IRRJ

INFORMATION RETRIEVAL
RESEARCH JOURNAL

Volume 1 · Number 2 · December 2025

Information Retrieval Research Journal
Published by RADBOUD UNIVERSITY PRESS
P.O. Box 9100, 6500 HA Nijmegen, The Netherlands
www.radbouduniversitypress.nl | radbouduniversitypress@ru.nl

ISSN: 3050-9106
E-ISSN: 3050-9114

Website: https://irrj.org

Cover design: Textcetera, The Hague
Print and distribution: Pumbo.nl

RADBOUD
UNIVERSITY
PRESS

Contents

Articles

A survey of Inclusive Information Access 167
Yue Zheng, Haiming Liu, Mike Wald

Evaluating Dense Model-based Approaches for Multimodal Medical Case Retrieval 197
Catarina Pires, Sérgio Nunes, Luís F. Teixeira

Effectiveness of In-Context Learning for Due Diligence: A Reproducibility Study of Identifying Passages for Due Diligence 221
Madhukar Dwivedi, Jaap Kamps

An Eigensystem for Topic Term Weighting yields Fair and Effective Document Rankings 247
Massimo Melucci

Exploring Embedding Interpretability by Correspondences Between Topic Models and Text Embeddings 281
Meng Yuan, Lida Rashidi, Justin Zobel

Emancipatory Information Retrieval 312
Bhaskar Mitra

Editorial

AI Resilience and Information Retrieval 339
Paul B. Kantor

Editorial board

Advisory board

Production editor

Webmaster

Mission & Scope

The Information Retrieval Research Journal (IRRJ) is a “diamond” open access journal that provides an international forum for the electronic and paper publication of high-quality scholarly articles in all areas of Information Retrieval. IRRJ commits to rigorous yet rapid reviewing. All published papers will be freely available online. Final versions are published electronically immediately upon receipt with a DOI (ISSN 3050-9114). Paper volumes are published and sold by Radboud University Press (ISSN 3050-9106). IRRJ does not charge article processing costs and aims to support researchers from low-income countries that currently have a hard time engaging with the field. IRRJ seeks unpublished papers on information retrieval research grounded in statistics, machine learning, linguistics, the cognitive sciences, and perhaps other related research fields, such as recommender systems. Papers may contain:

- new principled algorithms with sound empirical validation, and with justification of mathematical, theoretical, or psychological nature;
- experimental and/or theoretical studies yielding new insight into the design and behavior of information retrieval systems, including user-centric studies;
- reproducibility studies and applications of existing techniques that highlight the strengths and weaknesses of current methods;
- formalization of new information retrieval tasks (e.g., in the context of new applications) and methods for assessing the performance of those tasks;
- new evaluation approaches, including responsible information retrieval, fairness and non-discrimination in search;
- review and survey papers that contribute to the understanding of the state of the art in information retrieval.

Information Retrieval Research 1 (2025) 167-196 Submitted 12/24; Revised 05/25; Published 06/25

A survey of Inclusive Information Access

Yue Zheng YZ15U22@SOTON.AC.UK
University of Southampton
Southampton, United Kingdom

Haiming Liu H.LIU@SOTON.AC.UK
University of Southampton
Southampton, United Kingdom

Mike Wald M.WALD@SOTON.AC.UK
University of Southampton
Southampton, United Kingdom

Editor: Debarshi Kumar Sanyal

Abstract

Inclusive Information Access (IIA) is an essential field within information science and technology aimed at providing equitable information access for individuals from diverse backgrounds. Despite its significance, research on IIA remains dispersed across various domains, making it difficult to gain a comprehensive understanding of the field. This survey systematically examines the existing literature to map key themes, identify research trends, and highlight influential contributions within IIA. Using Latent Dirichlet Allocation (LDA) as a topic modeling method, this survey provides insights into the evolution of major topics, including publication trends, institutional engagement, and geographical reach. By analyzing authorship networks and citation patterns, this survey identifies key contributors, highly cited works, and influential research institutions, providing a structured overview of scholarly impact within IIA. This survey aims to support researchers in navigating the complex landscape of IIA and identifying avenues for future research.

Keywords: Inclusive Information Access, Information Retrieval, Digital divide

1 Introduction

In the digital era, access to information has evolved beyond mere convenience, becoming a fundamental human right (Bovens et al., 2000). Despite widespread access, systemic barriers related to age, gender, economic status, and disabilities continue to impede equitable information access, underscoring the importance of Inclusive Information Access (IIA) (Olphert et al., 2005; Hayes and Bulat, 2019). IIA seeks to ensure that all individuals have equitable access to the digital information landscape, irrespective of their challenges. This endeavor is not only about providing access but also about ensuring digital literacy, enabling individuals to effectively utilize and benefit from digital resources.

The scholarly exploration of IIA, while crucial, is still in its early stages. There is a growing consensus on the need for inclusive digital platforms and tools (Van Dijk and Hacker, 2003; Jaeger and Xie, 2009; Kiruki and Mutula, 2023), but comprehensive research

WWW: https://irrj.org DOI: 10.54195/irrj.21092

that clearly outlines the scope of IIA and proposes actionable solutions is limited. As the world becomes increasingly interconnected digitally, a nuanced understanding of IIA is vital.

Understanding the landscape of IIA requires a comprehensive approach that captures its complexity and breadth. A survey is essential for systematically mapping the diverse facets of IIA, providing a structured overview that highlights foundational contributions, major research trends, and emerging areas (Munn et al., 2018). By consolidating fragmented knowledge across topics like digital inclusion, accessibility technology, and information retrieval, this survey offers a detailed synthesis to guide researchers, policymakers, and practitioners striving to address the information divide.

Additionally, this survey integrates machine learning techniques to enhance the traditional survey process. By employing algorithms such as Latent Dirichlet Allocation (LDA) for topic modeling (Jelodar et al., 2019), the survey uncovers thematic patterns and provides insights into the field's evolution over time. This combination of scoping methodology with advanced computational tools ensures a more robust and efficient analysis, delivering a nuanced understanding of IIA's historical development, current trends, and future directions.

Despite the breadth of research on related areas like "digital inclusion" and the "digital divide" there remains a lack of comprehensive surveys specifically addressing the unique scope of IIA. Prior studies primarily focus on policy implications or sociocultural factors, often overlooking the technological dimensions that shape IIA.

In this study, we define Inclusive Information Access (IIA) as the design, implementation, and evaluation of information systems that enable equitable and effective access to digital content for individuals across the spectrum of physical, sensory, cognitive, socioeconomic, and linguistic diversity.

By combining bibliometric analysis and topic modeling, this review provides a structured synthesis of the IIA research landscape. It examines scholarly influence, thematic evolution, and cross-disciplinary engagement within the field. The findings reveal key trends, research gaps, and emerging directions that can inform future development in IIA. To achieve this, the following research questions are addressed:

RQ1: What are the dominant research patterns including influential authors, institutions, and major publication sources in IIA?

RQ2: How have research topic in IIA evolved over time?

RQ3: What are the existing gaps in IIA research, and what directions should future studies explore?

By addressing these questions, this study provides a comprehensive overview of IIA research, offering insights into its development, key challenges, and potential future directions.

2 Related work

The digital age has transformed the framework of information rights, shifting focus from mere acquisition to equitable access for all. This change has brought Inclusive Information Access (IIA) to the center of scholarly and policy discussions, emphasizing its critical role in the digital world. The 2003 World Summit on the Information Society highlighted this by advocating for an inclusive Information Society (Bryne, 2005).

Despite growing interest, a dedicated survey on IIA is lacking. This study aims to fill this gap by examining current reviews on the "digital divide" and "digital inclusion" drawing on methodologies and findings from these studies.

The concept of the digital divide, initially about access to technology, has evolved to focus on skills and usage. Studies have shifted from physical access to digital inclusion, considering social, cultural, and economic dimensions (Van Dijk, 2006; Parsons and Hick, 2008; Nemer, 2015). However, these primarily qualitative reviews lack empirical validation and overlook potential biases, highlighting the need for more rigorous research.

The discourse on "digital inclusion" has explored various barriers, particularly for older adults, and emphasized the need for holistic approaches that address root causes of digital exclusion (Olphert et al., 2005; Weerakkody et al., 2012). Pinder (2004) and Jaeger (2006) had linked digital inclusion to broader social inclusion, noting the gap between policy and practice in ICT accessibility for people with disabilities.

Methodologically, studies have employed various approaches to surveying quantitative literature, ranging from bibliometric analysis to topic modeling, each contributing to a structured understanding of the field (Kelly and Sugimoto, 2013; Istenic Starcic and Bagon, 2014; Chen et al., 2020; Vassilakopoulou and Hustad, 2023; Perez-Escolar and Canet, 2023). However, these approaches have limitations, highlighting the need for more comprehensive, interdisciplinary, and data-driven surveys to capture the evolving landscape of digital inclusion research.

The roots of IIA lie in special education, where initial efforts focused on individualized education for students with disabilities. However, these efforts fell short in bridging systemic gaps in equitable access to educational resources. Libraries, key to public knowledge, have faced criticism for inadequate infrastructure for universal accessibility (Subramaniam et al., 2013). Additionally, the engagement of students with disabilities in modern technology remains low (Ezeani et al., 2017).

The lack of a universally accepted methodology for information sharing among disadvantaged groups further complicates this issue (Hayes and Bulat, 2019). Addressing this requires increased funding for assistive technology infrastructure (Kiruki and Mutula, 2023). Information access challenges extend beyond physical access, encompassing comprehension and interpretation difficulties (Publishing, 2013). Despite efforts to simplify content for disadvantaged groups, there's no consensus on evaluating these interventions (Jaeger and Xie, 2009). A proposed multi-dimensional approach for IIA evaluation includes usability, accessibility, findability, comprehensibility, and reusability (Grenon et al., 2023).

Accessible design benefits not just those with impairments but society as a whole (LaCheen, 2000). Considering that most people will experience disability-like conditions at some point, creating an electronic society based on IIA principles is both an ethical and practical imperative.

Bibliometric analysis is a powerful tool for examining research trends across various academic fields. It involves analyzing bibliographic data like publication volume, key sources, prominent authors, citation rates, and relevant keywords (Snyder, 2019). This method reveals the development of research themes, collaboration patterns, and the overall intellectual landscape of a field. It identifies emerging topics and trends, guiding future research and encouraging interdisciplinary work. However, for rigor, a bibliometric analysis typically requires a dataset of at least 500 publications (Donthu et al., 2021). In newer research areas, meta-analysis (Aguinis et al., 2011) and citation analysis (De Groote, 2015) might be more appropriate.

Citation analysis, a key aspect of bibliometrics, examines the intertextual connections between scholarly works (Cronin, 2001). It identifies influential publications by analyzing citation patterns, thus assessing the impact of specific research contributions (Alzahrani et al., 2011). This analysis is useful for evaluating the citation metrics of articles, authors, institutions, and other indicators of academic productivity, aiding in reflective assessment and strategic planning (Nightingale and Marshall, 2012). However, citation analysis has limitations and should be applied carefully (Chikate and Patil, 2008). It quantitatively assesses citations from databases like Web of Science, Scopus, or Google Scholar, often visualized through citation networks or graphs to study publication trends and co-authorship dynamics.

3 Methods

For this survey, Web of Science (WoS) was selected as the principal data source. Although Web of Science (WoS) was chosen as the primary database due to its high-quality, peer-reviewed academic sources, we acknowledge its limitations (S. Adriaanse and Rensleigh, 2013). Future studies could supplement this dataset with Scopus, IEEE Xplore, or Google Scholar to enhance coverage and mitigate potential biases (Harzing and Alakangas, 2016; van Eck and Waltman, 2019). However, for this study, WoS was preferred for its structured citation data, facilitating bibliometric and topic modeling analyses.

Query statements were conducted in the Web of Science (WoS) Core Collection on June 11, 2023, using a Topic search(TS), which inherently includes Title, Abstract, Author Keywords, and Keywords Plus. The query applied the terms “Inclusive Information Access” OR “Inclusive Access Retrieval”, ensuring a focused retrieval of studies addressing IIA. The search was restricted to academic journal articles, conference proceedings, and review papers. The initial search retrieved 1,349 articles, which were refined by applying subject category filters to ensure relevance to Information Science, Computer Science, and Information and Communication Technology (ICT) topics, resulting in a final dataset of 299 articles for bibliometric and topic modeling analysis.

To ensure full reproducibility, the exact search query executed in WoS was as follows: `TS=("Inclusive Information Access" OR "Inclusive Access Retrieval")` AND `DT=(Article OR Proceedings Paper OR Review)` AND `LA=English` AND `PY=<=2023`

Also a Bibliometric analysis software has been introduced to improve graphical representation for more organized and insightful visualizations. The goal of VOSviewer (Van Eck and Waltman, 2010) is to produce maps that display bibliometric data. It can map phrases, authors, or journals, offering a simple yet effective way to comprehend a scientific field’s structure or the relationships between distinct concepts or writers. For data input, the software uses a user-friendly drag-and-drop interface. In the form of network maps, VOSviewer creates representations that are easy to see and comprehend. In contrast to CiteSpace (Chen, 2014), where many basic features require a paid upgrade, VOSviewer offers these functionalities for free, making it a more cost-effective choice

To identify journals that are most actively contributing to the field of IIA, we adopted the **h-index** and **g-index** as bibliometric indicators (Hirsch, 2005; Egghe, 2006). The **h-**

index of a journal indicates that it has published h articles that have each been cited at least h times, reflecting both productivity and consistent citation impact. The **g-index**, on the other hand, gives more weight to highly cited articles by identifying the largest number g such that the top g articles received together at least g^2 citations. These measures help surface journals that are not only frequently publishing in IIA-related topics but also whose publications are being consistently cited within this field.

To accomplish the domain topic modeling and evolution task, this study introduces the Latent Dirichlet Allocation (LDA) topic model, supplements word vectors using TF-IDF, calculates similarity using cosine distance, and visualizes topic evolution through Sankey diagrams (De Felice and Polimeni, 2020; Liu et al., 2022; Raza et al., 2019; Amjad and Ihsan, 2020; Zimmerman et al., 2021; Dhrangadhariya et al., 2020; Xiong et al., 2018; Owa et al., 2021). The workflow is presented as Figure 1, and the specific processes will be elaborated in detail in this section.

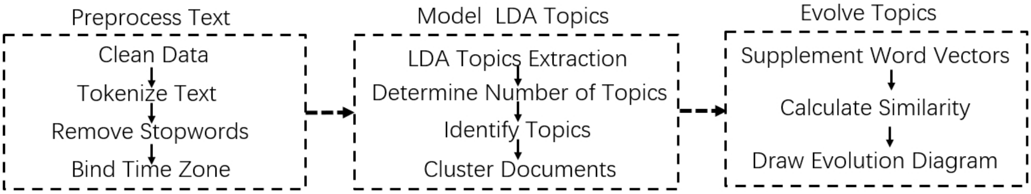


Figure 1: Workflow Diagram for Topic Modeling and Evolution

LDA is a generative probabilistic model crucial in topic modeling and Natural Language Processing (NLP). It assumes that each text in a corpus is composed of various topics, and each word in the text is associated with one of these topics. LDA uses a three-level hierarchical Bayesian model with Dirichlet distributions to uncover these latent topics (Blei et al., 2003).

In this study, LDA was employed to model the distribution of topics across documents and the distribution of words within topics. The articles' titles, contents, and publication years were retained for analysis. Text tokenization was performed using the Natural Language Toolkit (NLTK)'s word_tokenize function (Bird et al., 2009), and stopwords were filtered out using a unique list.

The evolution of the IIA field was divided into three periods: 2000–2010, 2011–2016, and 2017–2023. The rationale for selecting these periods as the basis for sorting publications is further explained in Section 4.1. LDA's perplexity metric, which measures how well the model's predicted probability distribution matches the actual distribution of words in texts, was used to determine the ideal number of topics. Lower perplexity scores indicate better model generalization (Wallach et al., 2009). The inflection point in the perplexity curve was chosen as the criterion for the optimal number of topics, adhering to the principle of parsimony.

After setting the topic parameters, the LDA model was used to identify topic terms for each period, followed by manual labeling to enhance interpretability by analyzing representative keywords and verifying them through a review of relevant articles. In this context, topic parameters refer to the predefined values that influence the model's structure and output, including the number of topics (K), which determines the granularity of thematic categorization, and the Dirichlet priors (α and β), which regulate the distribution of topics across documents and words across topics, respectively (Blei et al., 2003). Despite the effectiveness of LDA in identifying optimal topics and word distributions, manual intervention was necessary for summarizing and generalizing topic words. To ensure the reliability of topic interpretation and minimize potential bias, a second reviewer independently examined the LDA-generated keywords, validated the manually assigned topic labels, and cross-checked the phase boundaries used in the topic evolution analysis. Any discrepancies were discussed and resolved collaboratively.

The study also faced challenges in calculating cosine distances for topic evolution analysis due to low probability distribution weights of topic keywords. To overcome this, the Term Frequency-Inverse Document Frequency (TF-IDF) technique was used to identify high-frequency words within each topic (Zimmerman et al., 2021; Sparck Jones, 1972). Documents under each topic for each year were combined into a single document for TF-IDF vector computation.

Cosine similarity, measuring the similarity between two texts based on the cosine of the angle between their vectors in vector space, was used to establish connections between similar text sets (Wang et al., 2010; Ahad et al., 2016; Park et al., 2020). After normalizing the modeling to obtain vector representations, cosine similarity was calculated between texts. The study used topic vectors from TF-IDF computations to calculate cosine similarity between topics across different phases. An evolutionary relationship was established between topics when their similarity exceeded a predefined threshold. The final theme evolution diagram was visualized using the Sankey method in the pyecharts library (Chaudhuri, 2019).

To determine the optimal number of latent topics for LDA modeling, we trained a series of models with topic numbers K ranging from 2 to 10. The selection was guided by the perplexity score, a standard metric assessing how well the model predicts a sample. The inflection point in the perplexity curve indicated that $K = 5$ offered a balance between model complexity and generalizability, aligning with the principle of parsimony.

The LDA model was implemented using the Gensim library with symmetric Dirichlet priors. We adopted the default hyperparameters, setting $\alpha = 1/K$ and $\beta = 0.01$, which are widely recognized for providing robust results across diverse corpora. Given the exploratory nature of this study and the corpus size, these defaults were considered sufficient for capturing meaningful topics.

Text preprocessing was conducted using the Natural Language Toolkit (NLTK). This process involved lowercasing all texts, tokenization, stopword removal using an extended English stopword list, and filtering out punctuation and non-alphabetic characters. Stemming or lemmatization was intentionally omitted to preserve domain-specific semantics and enhance interpretability.

Topic quality was assessed both quantitatively and qualitatively. Perplexity scores were used for model comparison, and manual topic labeling was performed based on the top keywords per topic. Two independent reviewers collaboratively validated the labels to mitigate bias and enhance semantic coherence. This semi-automated evaluation approach is common in topic modeling applications where full automation of thematic interpretation remains limited.

4 Analysis and results

4.1 Publication Trend

Per the bibliometric framework delineated by Donthu et al. (2021), the temporal distribution of publications functions as a multifaceted indicator, invaluable for assessing the maturity and dynamism of a given research domain. Additionally, this temporal metric is an essential gauge for evaluating the field's evolutionary trajectory over a designated period. Key features such as the publication trend curve's growth rate and inflexion points yield incisive insights into the field's dynamic fluctuations. Such data prove indispensable for macro-level analyses to appraise the field's current scholarly prominence and for predictive models forecasting its developmental trajectory and emergent trends.

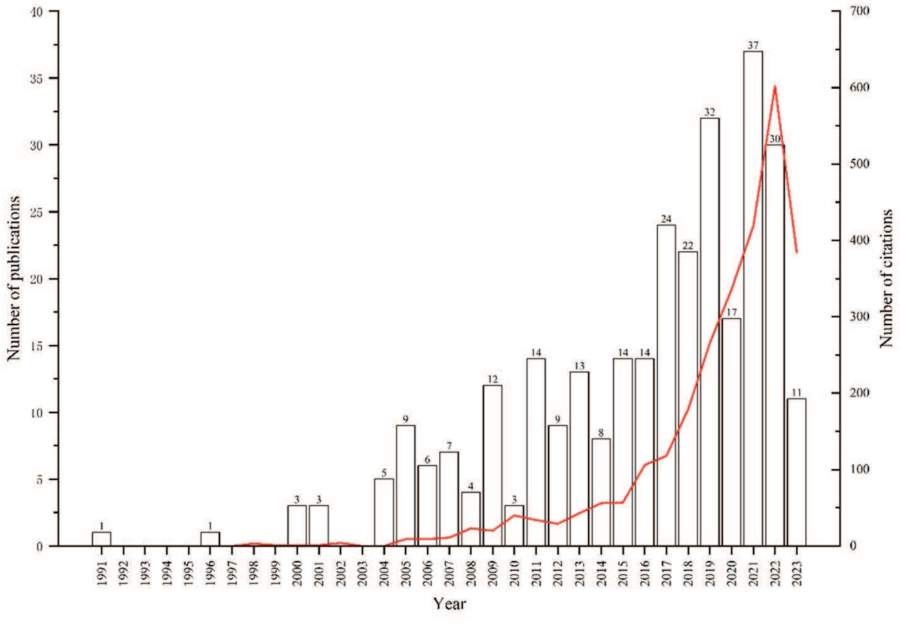


Figure 2: The number of annual research publications and citations

Upon statistical analysis of the temporal distribution of 299 publications, it became evident that IIA research originated as early as 1991. In this seminal work, Gerich (1991) explored the role of the National Science Foundation Network (NSFNET) in providing expansive network access to academic and research communities. The paper accentuated the dramatic escalation in network traffic and underscored the imperative for international collaboration and governance as the Internet began transcending national and global boundaries. This work marked the inaugural introduction of the term "Internet Inclusive" which, although not identical to the current focus on IIA, resonated with the burgeoning influx of information in the digital age.

It is worth noting that, following the convening of the 9th International Conference on Computers Helping People with Special Needs, the concept of Design for All (DfA) gained recognition as a significant paradigm within the realm of Information Society Technologies (IST) (Bühler and Stephanidis, 2004). This concept aimed to guarantee universal access to technology products and services by eliminating existing barriers. After 2004, the volume of publications pertinent to IIA consistently escalated, signalling an intensifying scholarly engagement with the subject matter.

Specific data reveals that from 1991 to 2023, an average of 9.34 articles were published yearly. These articles have been cited 2751 times, with an average annual citation rate of 85.97. Analyzing Figure 2, it can be observed that the field has undergone three distinct development phases.

From 1991 to 2010, the first stage served as an exploration phase. The field was still young, with only three papers published a year on average during this time. Due to the limited availability of documented outputs prior to the year 2000, with only two instances, the initial phase of the study was adjusted to focus on the period from 2000 to 2010. This modification was implemented to better align the study's timeframe with the broader temporal context, thereby enhancing the validity of subsequent analyses. The second phase, which lasted from 2011 to 2016, included fast expansion. At this time, the discipline produced 12 publications annually on average, and papers were getting many more citations.

The third and current phase started in 2017 and is still in progress. Academic output during this period was solid and consistent. The culmination was a record-breaking 37 publications in 2021. These papers have rec lasted from 2011 to 2016 andations as well. It is essential to keep in mind that the data for 2023 is insufficient because it only contains items released up until August. However, according to the most recent data, the number of articles published has reached the Phase II average, and the number of citations is higher than the previous year.

In conclusion, there has been a steady increase in IIA research in recent years. Three main developmental phases can be identified for this expansion: the original stage, which lasted from 2000 to 2010, the middle phase, which lasted from 2011 to 2016, and the most recent stage, which lasted from 2017 to 2023. Given the course it is taking, it is logical to assume that the field will continue to grow in the future.

4.2 Publication Source

In the dataset under investigation, 231 academic journals are represented. Among these, a minority subset of 13 journals is remarkably prolific, each contributing more than three articles to the corpus. In addition, 25 journals have contributed exactly two papers, while the vast majority, with 193 journals, have each contributed a single piece to the field. This distribution suggests a highly skewed landscape where a few journals are the primary conduits for scholarly output in this area. The top 10 prolific publication Sources in the field of IIA are shown in Table 4.2. "Universal Access in The Information Society" despite having the highest number of articles, has an h-index of 5 and a g-index of 8, which are not significantly higher than those of other journals. This suggests that while it may

Rank	Sources	h-index	g-index	TC	Articles
1	Universal Access in The Information Society	5	8	72	10
2	Journal of Documentation	3	8	80	8
3	Telecommunications Policy	5	6	109	6
4	Telematics and Informatics	4	4	182	4
5	IEEE Access	3	4	50	4
6	Ethics and Information Technology	3	3	26	3
7	Information Society	3	3	91	3
8	Journal Of The Australian Library and Information Association	3	3	16	3
9	Interacting With Computers	2	3	14	3
10	International Journal Of Electronic Government Research	2	3	11	3

Table 1: Top 10 Publication Sources

be a cornerstone in terms of volume, its overall impact, as measured by these indices, is comparable to other journals like "Telecommunications Policy" which has an h-index of 5 but a higher total citation count of 109.

This concentration of publications within a top-tier subset of journals suggests that the field is beginning to coalesce around a central academic nexus. Yet, it's worth noting that the area also enjoys contributions from a broad spectrum of journals, indicating its inherently multidisciplinary nature. Journals like "Ethics and Information Technology" and "IEEE Access" exemplify this trend, highlighting the intersection of ethical, technological, and policy dimensions within IIA research.

4.3 Publication Institution

For the purpose of locating potential partners in the field, it is essential to comprehend the institutional landscape of IIA research. In order to do this, the linkages between various research institutes based on their contributions to IIA were mapped using a network visualization Figure 3.

The network visualization offers a number of significant insights. First, each node in the network has a size that reflects the number of publications the institution has produced, acting as a measure of research output. As an illustration, the "University System of Maryland" node is clearly larger, demonstrating its considerable contribution to IIA research. Second, the strength and temporal characteristics of institutional collaborations are indicated by the thickness and colour of the lines joining the nodes. However, the lines connecting the nodes are neither numerous nor particularly thick or even almost obscured by the institutional nodes, suggesting that institutional cooperation is not common overall. This suggests the industry is still in an exploratory phase where institutions independently look into different aspects of IIA.

The absence of significant inter-institutional collaborations in the IIA research landscape is evident. However, Table 4.3 offers a different perspective. Specifically, the "University System of Ohio" stands out with a high total citation count (TC) of 161, indicating its substantial impact on the field.

According to Table 4.3, it is clear that the institutions with the highest publication output regularly work with other institutions, indicating that inter-institutional collabora-

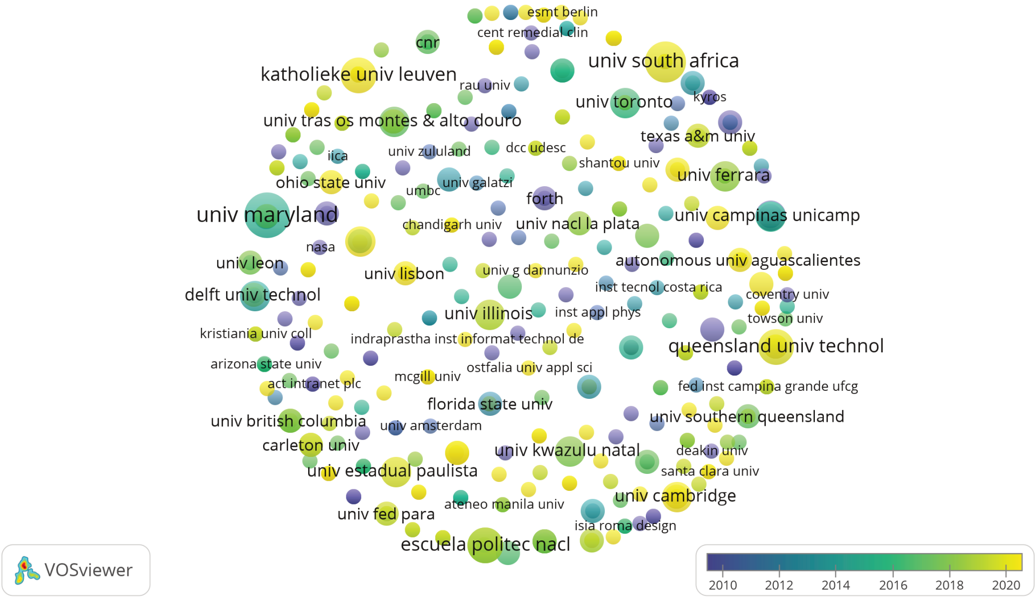


Figure 3: Institutional Cooperations

Institution	Country	NP	First Institution	Corresponding Institution	TC	PC
University System of Maryland	America	8	7	4	62	7.75
Escuela Politecnica Nacional Ecuador	Spain	6	3	4	71	11.83
Universidade Estadual De Campinas	Brazil	6	3	3	10	1.67
University of South Africa	South Africa	5	5	4	64	12.8
Consiglio Nazionale Delle Ricerche	Italy	4	3	1	7	1.75
Ku Leuven	Belgium	4	4	3	12	3
Pennsylvania Commonwealth System of Higher Education Pcshe	America	4	4	0	15	3.75
Queensland University of Technology	Australia	4	4	4	41	10.25
Universitat D Alacant	Spain	4	2	0	60	15
University of Maryland College Park	America	4	4	0	23	5.75
University System of Ohio	America	4	4	0	161	40.25

Table 2: Institutional Contributions

tion greatly boosts field developments. Notably, the Queensland University of Technology stands out for having a high preference for external cooperation; each of its four papers was co-authored with other institutions. This highlights the university's proactive strategy for creating relationships, which may work as a catalyst for encouraging innovation and research excellence in the field.

In conclusion, the state of collaboration is generally lacklustre. Nevertheless, institutions with the highest publication output exhibit a stronger propensity for cooperation. So, the

Country	Articles	SCP	MCP	Percent	MCP Ratio	TC	Avg. Article Citations
USA	57	51	6	0.191	0.105	578	10.1
UK	22	15	7	0.074	0.318	180	8.2
Brazil	21	20	1	0.07	0.048	101	4.8
Australia	19	14	5	0.064	0.263	154	8.1
Canada	16	14	2	0.054	0.125	122	7.6
South Africa	12	9	3	0.04	0.25	129	10.8
India	12	10	2	0.04	0.167	78	6.5
Italy	9	8	1	0.03	0.111	15	1.7
Portugal	8	8	0	0.027	0	9	1.1
China	7	5	2	0.023	0.286	167	23.9

Table 3: Top 10 Publication Countries

lack of connections between these academic institutions raises the possibility of collaborative research and knowledge sharing being underutilized.

Therefore, it is essential that future studies concentrate on establishing collaborative frameworks while appreciating individual institutions' contributions. Such a strategy could enhance the IIA research environment by ensuring a more comprehensive and broad spectrum of perspectives while accelerating field progress.

4.4 Publication Country

Understanding the worldwide geography of this study topic requires understanding the contributing nations in the field of IIA. With the use of citation rates, this analysis will be able to pinpoint the nations that produce the most research, contribute to it in the most meaningful ways, and have the most overall influence. The technique comprises a thorough analysis of the literature with a focus on the number of papers published in each nation, the average number of citations per article, and the proportion of multi-country to single-country publications.

According to the data, 64 nations or areas have contributed to the worldwide IIA research effort. Table 4.4 further categorizes the contributions by Single Country Publications (SCP), Multi-Country Publications (MCP), Percent of MCP, MCP Ratio, Total Citations, and Average Article Citations. For instance, the United States leads with 57 articles, of which 51 are SCPs, and has an average citation rate of 10.1. This indicates high research output and significantly impacts the academic community. With just seven articles, China has an average article citation rate of 239, highlighting the impact and quality of its research in this field.

In the realm of scientific collaboration, the network visualization Figure 4 offers a compelling snapshot of the intricate relationships between various countries in terms of their research output. The graph reveals several key trends and patterns that merit further scrutiny.

For the purpose of this analysis, data for England and Scotland are aggregated under the United Kingdom (UK). The United States and the United Kingdom prominently lead in both the volume of publications and the extent of their academic collaboration in the field

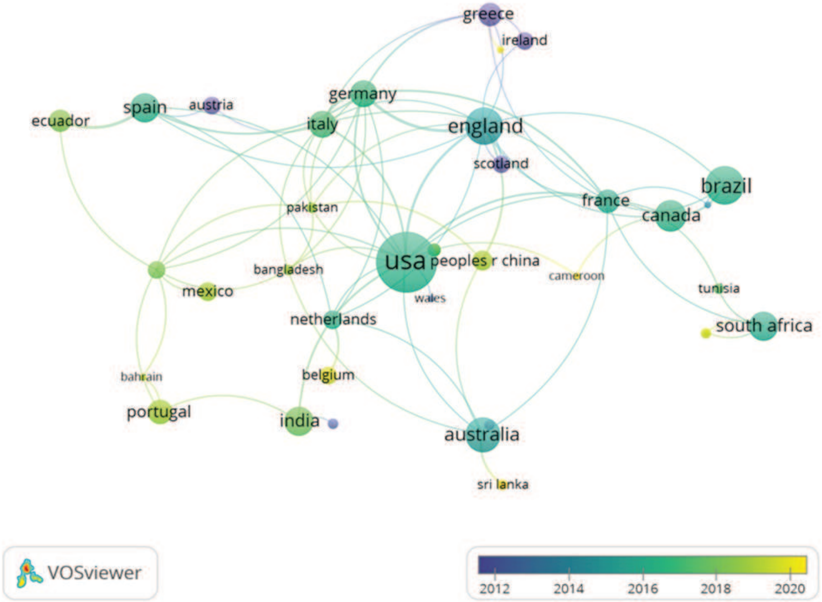


Figure 4: Collaborations of Countries

of IIA research. This suggests a significant strategic partnership, reflecting a high degree of synergy and shared research interests.

Additionally, other strong international collaborations are evident, such as the partnership between the United Kingdom and Germany, and the cooperation between Ecuador and Spain. These relationships, likely driven by shared research objectives or complementary expertise, highlight the global nature of scientific research.

Factors like common language and educational systems might facilitate the U.S.-U.K. collaboration, while cultural and historical connections could influence the Ecuador-Spain partnership.

Overall, the data underscores the global and interconnected nature of IIA research, emphasizing the importance of international collaboration in advancing knowledge. This suggests that the future of IIA research relies on a collective, globally-inclusive effort.

4.5 Authorship

Prolific authors serve as the cornerstone of any research field (Brito et al., 2023), and this is no less true for the domain of IIA. Between 1991 and 2023, a total of 840 authors have contributed to the body of literature on IIA. This study employs Price's Law, a well-established bibliometric principle, to identify core contributors within the expansive authorship network (Price, 1963). According to this law, authors with more than two publications can be considered as core authors.

However, the application of this criterion revealed that as many as 52 authors had published more than two articles. This introduced complexity into the analysis due to a high number of tied rankings among authors with fewer publications. This is most likely due to the small total number of articles studied, as well as the fact that the field is still in a developmental stage. To provide a more focused view, the criterion was adjusted to spotlight authors with three or more publications, as detailed in Table 4.5.

Author	h-index	g-index	TC	NP	PY start
Baranauskas MCC	2	3	10	6	2009
Lujan-mora S	3	5	68	5	2015
Sitbon L	3	4	40	4	2017
Stephanidis C	3	4	37	4	2004
Bonacin R	1	1	3	4	2010
Acosta T	3	3	57	3	2018
Brereton M	2	3	33	3	2018
Neris VPD	2	2	7	3	2009
Maietti F	1	2	5	3	2018
Dos reis JC	1	1	3	3	2010

Table 4: Top 10 core authors

By adopting this more stringent criterion, the study aims to offer a more focused view of the most impactful authors, thereby elucidating the intellectual landscape of IIA research

A review of articles published by core authors in the field reveals different research focus. Baranauskas MCC, Bonacin R, Dos reis JC, and Neris VPD engaged in the design and construction of inclusive social networks tailored to the specific characteristics of Brazilian society. Lujan-mora S and Acosta T primarily evaluated the accessibility features of various web pages. Sitbon L and Brereton M concentrated on extending learning opportunities in information technology for individuals with intellectual disabilities, while also encouraging their participation in information retrieval tasks. Stephanidis C proposed semantic-based user modeling to facilitate the adaptation of web-based user interfaces. Maietti F, on the other hand, focused on tasks related to the preservation of inclusive cultural heritage.

This diversity in research topics underscored the multidisciplinary nature of IIA and highlighted the range of approaches that were employed to address its complex challenges.

Expanding upon the foundational research previously discussed, a visual representation of author collaboration networks, as shown in Figure 5, was generated using VOSviewer. In this graph, it is noteworthy that among authors with more than two publications, there are 12 who have published independently, while the rest have engaged in collaborations to varying degrees. However, the majority of these collaborations occur within the same institution.

The field may still be in its early stages because the frequency and density of partnerships are often low. The small number of partnerships also raises the probability that many writers are working alone, which may reduce the variety of viewpoints and research methods in the subject.

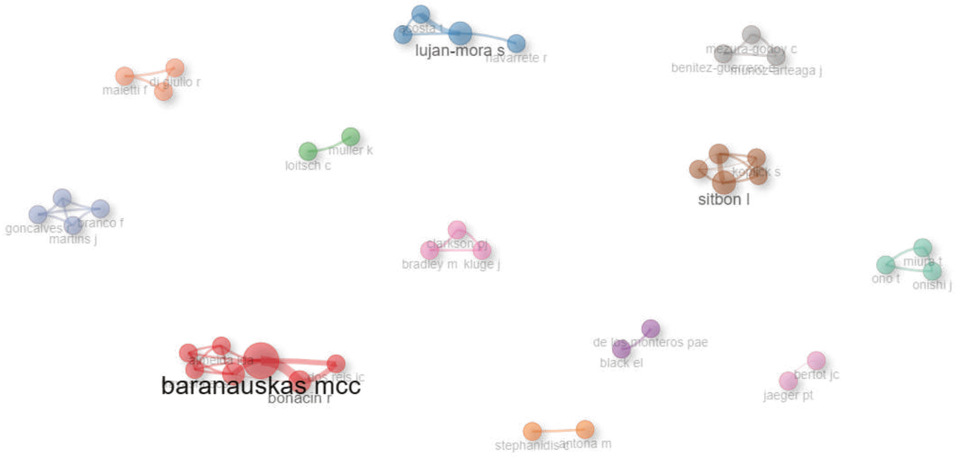


Figure 5: Collaborations of Authors

Given this context, it stands to reason that more varied and frequent partnerships ought to be promoted. The research environment in IIA may be considerably enhanced by encouraging interdisciplinary and interinstitutional cooperation, which would also provide researchers with a deeper knowledge of the complexities of the field.

4.6 Citation

To gain a deeper understanding of the influential works shaping the field of Information Inequality and Accessibility (IIA), it's useful to examine the most cited publications. Table 4.6 lists these key articles, which serve as important benchmarks in the existing body of knowledge and guide future research directions. The subsequent sections will delve into the details of each article, highlighting their contributions and significance in the IIA landscape.

Title	Global Citations	Local Citations
Web of Science (WoS) and Scopus: The Titans of Bibliographic Information in Today's Academic WorldPranckutė (2021)	269	0
Digital inclusiveness - Longitudinal study of Internet adoption by older adultsLam and Lee (2006)	138	2
Consumers' acceptance of information and communications technology in tourism: A reviewUkpabi and Karjaluoto (2017)	137	0
5G mobile technology: A surveyMitra and Agrawal (2015)	122	0
Mobile health technology adoption across generations: Narrowing the digital divideFox and Connolly (2018)	109	0

Table 5: Top 5 Cited Articles

A comprehensive overview of two major bibliographic databases, Web of Science (WoS) and Scopus, is provided by integrating information from the database owners with the latest research findings (Pranckutė, 2021). The investigation's primary area of interest is information accessibility. Similar to another study which used the review method to examine consumer adoption of web-based services, social media, and mobile information systems in the tourist industry, did not specifically address how inaccessible e-tourism is for populations with limited access to information (Ukpabi and Karjaluoto, 2017).

Lam's study explored the use of the Internet by the elderly, showing that training improves their online confidence and willingness to use the Internet (Lam and Lee, 2006). Fox's 2018 research (Fox and Connolly, 2018) found that older adults are hesitant to use mobile health technologies due to mistrust and privacy concerns, suggesting the need for designs and education that enhance their confidence and understanding of privacy.

Mitra (Mitra and Agrawal, 2015) offered a thorough review of 5G development activities. However the terms "inclusive" and "access" pertained to the hardware's compatibility with the technology. It is a different topic than the one explored in this study. This also reflects the confusing nature of nomenclature in the IIA field.

While the above publications receive the most global citations, only one paper receives two local citations, indicating a potential gap between local and global academic significance. This could imply that although these publications are acknowledged and mentioned in the larger academic world, the IIA research environment may not find them to be as significant or pertinent.

The significance and methodological rigor of the research they embody are both reflected in highly cited articles, which act as significant landmarks in the academic landscape. Table 4.6 displays the top ten cited articles in research in the field of IIA, reflecting the multifaceted nature of the field.

Title	Local Citations
Transforming our world: the 2030 Agenda for Sustainable DevelopmentResolution et al. (2015)	7
Development as freedomSen (2000)	7
Digital divide: Civic engagement, information poverty, and the Internet worldwideNorris (2001)	6
Cache template attacks: Automating attacks on inclusive {Last-Level} cachesGruss et al. (2015)	5
User-Sensitive Inclusive Design in Universal Access in the Information Society August 2011Newell et al. (2011)	5
Social network sites: Definition, history, and scholarshipBoyd and Ellison (2007)	5
Digital divide research, achievements and shortcomingsVan Dijk and Hacker (2003)	5
User acceptance of information technology: Toward a unified viewVenkatesh et al. (2003)	5
The impoverished life-world of outsidersChatman (1996)	5
Basic formal education quality, information technology, and inclusive human development in sub-Saharan AfricaAsongu and Odhiambo (2019)	4

Table 6: Top 10 Cited Reference

The most highly cited reference in the dataset is "The 2030 Agenda for Sustainable Development (Resolution et al., 2015)". Within this seminal work, the objective to "Make urban areas and habitations inclusive, secure, robust, and sustainable" emerges as the most frequently cited content, followed by discussions on inclusivity in learning and work environments. Collectively, citations to this article underscore the foundational premise in the field of IIA that access to the Internet, emergency services, and information and communication technologies constitutes a fundamental human right. Sen (2000) emphasized the importance of IIA, advocating for equipping those without information access with the necessary tools for self-reliance. Van Dijk and Hacker (2003) identified four types of digital divide access: motivation, physical, skills, and usage, underscoring the need for comprehensive theoretical frameworks, clear definitions, multidisciplinary approaches, qualitative analysis, and long-term studies in this area. This has influenced later research to fill these gaps.

A survey of the literature revealed a large worldwide digital divide between high-income and low-income countries. This finding forms the basis for future study into "e-government"-related issues in the area of IIA (Norris, 2001). It is important to remember, though, that the growth of "e-government" should not exclude people who are already informationally underprivileged, as doing so risked widening the digital divide already present. The research on sub-Saharan Africa has garnered considerable internal citations (Asongu and Odhiambo, 2019). The study posits that the low prevalence of mobile phone usage in the region is constrained by inadequate primary education, thereby highlighting a significant digital divide in comparison to global information development levels.

To strengthen ties between designers and users, Newell et al. (2011) encouraged the adoption of a moderately inclusive and user-sensitive design philosophy. According to the research, the use of inclusive design principles successfully reduced a variety of obstacles that people with impairments encountered when using technology. In the field of "inclusive design" this work served as a foundational study, and subsequent research used this methodology as a point of reference (de Almeida Neris et al., 2021). Additionally, Newell conceded that it was impractical to create goods that were accessible to all potential users. As a result, the article substituted the terms "inclusivity" and "sensitivity" for "universality" and "centrality" respectively, defining a more achievable and, in many cases, more appropriate goal. Venkatesh et al. (2003) introduced the Unified Theory of Acceptance and Use of Technology (UTAUT) model, which synthesizes eight prominent models from the user acceptance literature. Performance expectancy, effort expectancy, social influence, and enabling conditions are the four main factors identified by the UTAUT model. This study offers important insights for future research on universal information access since it focuses largely on how information is accepted in traditional cultures.

Many studies drawing on the characteristics of social networks, propose the use of user-specific data during retrieval to offer customized responses, a strategy that holds promise in the realm of inclusive design (Boyd and Ellison, 2007). However, the practical implementation of this approach can encounter ethical challenges, particularly concerning data privacy. Additionally, the re-collection of specific data could compromise the method's broad applicability.

To summarize, the highly referenced literature is divided into three categories, including: 1. The importance of study in the field of IIA 2. The global digital divide in its context 3. The underlying theories and procedures for inclusive design.

4.7 Topic Evolution

Figure 6 shows a thorough theme evolution map for the field of IIA spanning the years 2000 to 2023, building on the LDA topic modeling methodology described before. Each phase's individual color blocks correspond to different topics. These blocks' sizes show how many articles fall under each topic. The connections between blocks in neighboring phases represent the relationship between them and the course of progression. A relationship between two themes from different phases implies that a theme has evolved or changed through time.

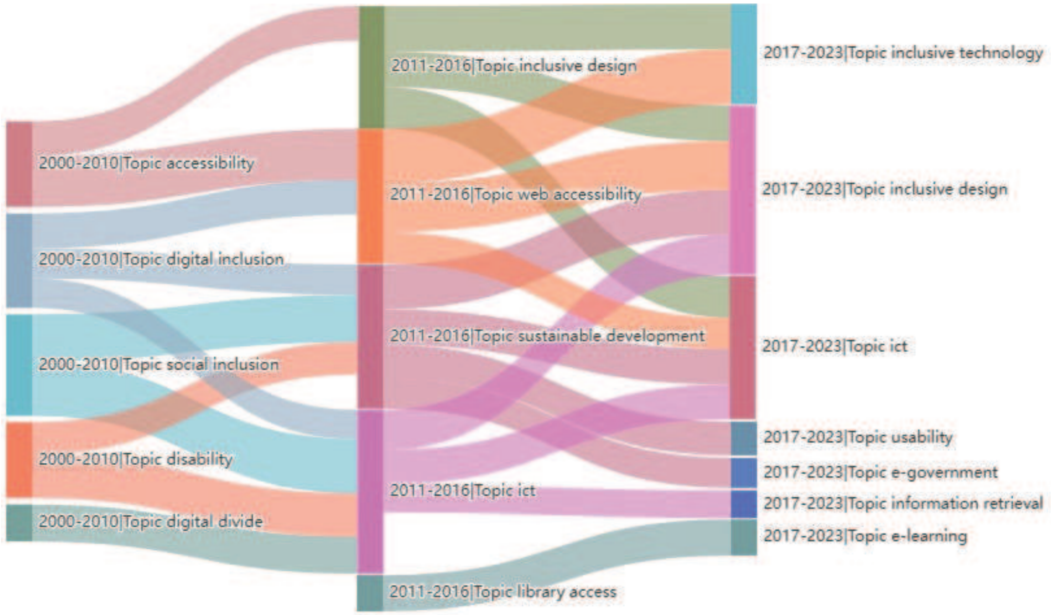

2000-2010|Topic accessibility
2000-2010|Topic digital inclusion
2000-2010|Topic social inclusion
2000-2010|Topic disability
2000-2010|Topic digital divide
2011-2016|Topic inclusive design
2011-2016|Topic web accessibility
2011-2016|Topic sustainable development
2011-2016|Topic ict
2011-2016|Topic library access
2017-2023|Topic inclusive technology
2017-2023|Topic inclusive design
2017-2023|Topic ict
2017-2023|Topic usability
2017-2023|Topic e-government
2017-2023|Topic information retrieval
2017-2023|Topic e-learning

Figure 6: Theme evolution

Based on the thematic evolution diagram, the field of IIA has evolved from its initial focus on broad concepts such as accessibility, digital inclusion, social inclusion, disability, and the digital divide. It has shifted towards more specialized topics like inclusive design, web accessibility, sustainable development, ICT, and library access. Currently, the research landscape is increasingly concentrating on practical scientific and technological methods, particularly in areas such as usability, e-government, information retrieval, and e-learning, which exhibit significant potential for further development.

5 Discussion

In the evolving landscape of IIA, this section focuses on inclusive technology and information retrieval, laying the groundwork for future advancements in inclusive information retrieval study. A comprehensive literature review on the topic of inclusive design has already been conducted by Li et al. (2023), serving as a valuable resource in this domain. The goal of inclusive technology is to make information and communication technologies available to everyone, regardless of age, physical condition, or socio-economic status. This field has gained a lot of attention in recent years as researchers focus on different target populations, such as the elderly and those with visual or hearing impairments. They use

Key Contributions	Target Population	Methodology
TV-based service to enhance well-being and self-esteem Epelde et al. (2013)	Elderly	User-centered design
iLiBra platform to streamline communication using sign language icons da Costa et al. (2019)	Hearing impairments	Assistive technology platform with sign language
Drawxi, an audio-haptic tool for collaborative diagram creation Chiplunkar et al. (2019)	Visual impairments and limited sight	Sensor-based solutions
Real-time tracking model for social interactions Grayson et al. (2020)	Visual impairments	Head-mounted HoloLens device
Force-FeedbackTablet (F2T) for interacting with 2D data Gay et al. (2021)	Visual impairments	Haptic architecture

Table 7: Summary of Inclusive Technology

various methods and technologies to tackle the specific challenges faced by these groups, ultimately contributing to the larger objective of providing universal access to information.

Table 5 provides a comprehensive summary of research articles with a clear focus on technology in the field of inclusive technology, outlining their contributions, target populations, and methodologies.

Current technology in this field primarily aids those with visual impairments, using audio conversion to overcome visual challenges. However, incorporating tactile feedback could enhance understanding of complex information. For individuals with hearing loss, the focus is on converting content into sign language. Technologies for the elderly prioritize usability and clarity.

Yet, there's a notable lack of comprehensive research on IIA for people with cognitive impairments, including learning disabilities, attention deficit disorders, and dementia. Current research inadequately addresses the unique challenges faced by these individuals in understanding and retaining information.

In a sequence of investigations focused on the development of inclusive social network systems in Brazil, a novel methodology for personalized system design is introduced(Neris et al., 2009; Almeida et al., 2009; Reis et al., 2010; dos Reis et al., 2010; Dos Reis et al., 2014; de Almeida Neris et al., 2021). This methodology diverges from traditional search solutions that rely on lexical-syntactic information processing. Instead, it advances an innovative approach to web ontology design by amalgamating semantic modeling techniques with a content-based strategy for ontology creation. This inclusive search mechanism yields semantic search results that are universally applicable, while simultaneously honoring the unique preferences and needs of individual users.

Complementing this, the inclusive design literature (summarized in Table 5) provides a methodological foundation for creating digital environments that are not only accessible, but also participatory and personalized. These studies emphasize the integration of user-centered and participatory design frameworks, ergonomic data visualization, and interface design patterns such as progressive disclosure, ensuring that digital systems are flexible and responsive to diverse user needs. Importantly, they also demonstrate how co-design workshops and iterative evaluation processes help translate user insights into usable interface prototypes, reinforcing inclusivity at the design level.

Title	Author (Year)	Methodology
Attuning Speech-Enabled Interfaces To User And Context For Inclusive Design: Technology, Methodology And Practice	Neerincx et al. (2008)	Accessibility analysis and iterative refinement of speech-based multimodal interface.
Accessing User Information For Use In Design	McGinley and Dong (2009)	Tool concept development and co-design workshop to improve ergonomic user data communication.
Providing Universally Accessible Interactive Services Through Tv Sets: Implementation And Validation With Elderly Users	Epelde et al. (2011)	User-centered design using interviews and focus groups to model older adults' needs.
User-Centered Design Journey for Pattern Development	Joshi et al. (2017)	User-centered design with iterative evaluation of UI solutions; developed the progressive disclosure pattern via case study.
Addressing Brazilian diversity in personal computing systems with a tailoring-based approach	Almeida Neris et al. (2020)	Participatory tailoring approach integrating PLuRaL and FAN frameworks; applied in inclusive system design.

Table 8: Summary of Inclusive Design

Title	Author (Year)	Methodology
Addressing Universal Access In Social Networks: An Inclusive Search Mechanism	Reis et al. (2013)	Semantic search design using Web ontology and Organizational Semiotics; evaluated via case study in a social network system.
A Reference Ontology For Digital Scientific Journals Applied To Systematic Literature Review Processes	Ghisi et al. (2012)	Used reference ontology with batch/on-the-fly processing to enhance interoperability and inclusiveness in scientific literature retrieval.

Table 9: Summary of Information Retrieval

Building upon these foundations, research on inclusive information retrieval (IIR) has begun to emerge as a critical subdomain within IIA. As summarized in Table 5, current efforts in this space have explored the use of semantic web technologies (Reis et al., 2013), organizational semiotics, and reference ontologies to develop more accessible and adaptive search mechanisms. These approaches aim to reduce the cognitive and linguistic barriers often encountered by users with low digital literacy, enabling them to retrieve information using familiar, meaningful terms rather than system-generated query syntax. However, these early investigations remain limited in scope and scale, suggesting a pressing need for more innovation in retrieval algorithms, interface design, and user modeling to support a wider range of impairments and contexts.

As attention increasingly shifts toward inclusive approaches in Information Retrieval (IR), recent research has moved beyond traditional, text-dominant keyword-based systems toward more adaptive, context-aware, and user-centered IR frameworks. These systems are designed to accommodate a broader spectrum of user profiles. For instance, several studies (Guo et al., 2018; Zhang et al., 2020; Erbacher et al., 2022) have emphasized the integration of interactive retrieval frameworks that support cognitive offloading. Such systems allow users to refine queries through dialog-based or visually guided interactions, thereby alleviating the need for precise keyword formulation. This paradigm shift aligns with the broader goal of reducing cognitive and linguistic load, particularly for users with learning difficulties, neurodiverse conditions, or limited language proficiency.

In parallel, other research (Hsieh et al., 2022; Zheng et al., 2024) has investigated the role of multimodal information retrieval, enabling users to access content through combinations of text, speech, and images. These systems are particularly beneficial for individuals with visual impairments, low literacy, or temporary accessibility challenges—such as users navigating complex physical environments. Multimodal interfaces expand the expressive capacity of users and diversify input modalities, further contributing to the development of inclusive IR systems.

Furthermore, the integration of user modeling and personalization into IR systems has become increasingly sophisticated. Recent studies (Kladouchou et al., 2025) have explored the use of real-time behavioral data, user preferences, and even physiological signals to dynamically adjust retrieval results. Such advancements enable IR systems to more accurately anticipate users' needs and deliver content in formats aligned with their individual capabilities and preferences (Zheng et al., 2024; Ji et al., 2024).

To achieve an inclusive information retrieval system, a multidisciplinary approach is essential. This involves integrating inclusive design principles to create user interfaces for various impairments, employing inclusive technology to enhance sensory accessibility, and utilizing specialized information retrieval algorithms for relevant personalized results. The aim is to address the various challenges in providing equitable access to information.

Recognizing limitations is critical, including potential data oversight from other databases and issues with keyword selection that resulted in incomplete or extraneous data. The findings are preliminary and may lose relevance over time. Although machine learning can help with theme recognition, accurate topic labeling still requires human interpretation.

Future research areas are highlighted, focusing on the need for technology that caters to a wide range of disabilities, particularly cognitive impairments, as well as the development of inclusive information retrieval systems. Addressing these issues has the potential to significantly advance the field of Inclusive Information Access, improving efficiency and accessibility for a broader audience.

6 Limitations and Future Work

Recognizing the limitations of this study is essential. First, despite the fact that the Web of Science is a credible bibliographic source that is acknowledged around the world, it is still possible that pertinent data from other databases may have been ignored. Second, the study's keyword selection procedure failed to consider any potential ambiguities, which led to the inclusion of superfluous information in the dataset. An incomplete dataset was also produced due to several important keywords being accidentally left out due to an early lack of a thorough grasp of the area. Moreover, the study's dependability may decline with time, it should be viewed as offering preliminary advice. Lastly, it is crucial to recognize that even though machine learning approaches may extract themes with topic words from hundreds of papers, correct labelling of these topics still needs thorough literature reviews. Results from machine learning can lead to future studies, but they cannot completely replace human interpretation and comprehension of the literature.

While this study offers an exhaustive analysis of the IIA domain, it also delineates several uncharted territories warranting future exploration. A critical area of concern is the creation of technology designed to accommodate individuals with a range of disabilities. Subsequent research could particularly target populations with cognitive impairments, aiming to integrate and innovate adaptive technologies that elevate the user experience for these groups.

When data were gathered, little study was done on the use of LLM in IIA. Although research on AI and LLM has been expanding rapidly, with models such as ChatGPT and GPT-4 transforming natural language processing and information retrieval (IR), their particular uses in inclusive information retrieval and accessibility were still in their infancy. In recent years, however, advances in AI-enhanced assistive technologies, adaptive retrieval systems, and LLM-driven conversational agents demonstrate the growing role of intelligent systems in breaking traditional barriers to information access (Zhu et al., 2023). These developments align with broader efforts in inclusive technology and information retrieval (IR) to ensure that digital content is available, navigable, and comprehensible for all users, including those with disabilities and those from underrepresented linguistic or cognitive backgrounds.

In order to improve accessibility for a variety of user groups, recent research highlights the potential of LLM in augmenting assistive technologies by facilitating multimodal interactions through text, audio, and image-based inputs (Adnin and Das, 2024; Raji et al., 2025). By enabling users to engage in natural discourse instead of intricate keyword-based searches, AI-driven conversational agents are proving crucial in information retrieval and improving the accessibility and navigability of digital content (Martínez et al., 2024). For users with impairments and those from low-resource language backgrounds in particular, these technologies, when combined with real-time translation and content simplification models, present a potential path toward overcoming linguistic and cognitive limitations (Fu et al., 2025).

Even with these developments, there are still significant obstacles to overcome in order to guarantee fair, transparent, and bias-free AI-driven retrieval. Research emphasizes the necessity of fairness-aware ranking algorithms that enhance accessibility-aware content ranking and reduce biases in training data (Sitbon et al., 2023). Future research must concentrate on explainability and trustworthiness to make sure users can understand and evaluate AI-generated responses as AI-generated material becomes more and more integrated into search and retrieval systems. Furthermore, the creation of fully inclusive AI solutions that meet practical accessibility requirements will depend on the advancement of co-design approaches, in which impaired users actively participate in system development. In order to make sure that information retrieval systems empower various user groups rather than exclude them, it will be essential to incorporate universal design principles into every phase of AI model training, evaluation, and deployment going ahead.

Furthermore, there is a specific need to focus on the design of inclusive information retrieval. This would involve developing methods and technologies that offer a more inclusive approach to information retrieval for individuals with disabilities. By doing so, we can empower them to independently access information, allowing for a more seamless integration into the information-driven world.

Future academic contributions can significantly contribute to the advancement of the IIA area by addressing these highlighted gaps and obstacles. With a broad user base catered to, it would become more productive, efficient, and inclusive.

7 Conclusion

In conclusion, this study thoroughly analyzes the IIA area, shedding light on its evolving trends in terms of publication timelines, institutional contributions, and geographical distribution. Even while the area is expanding year over year and there is a definite tendency toward globalization, it has not yet developed strong academic partnerships. The study offers an early knowledge of the present status of research in the IIA sector through a citation analysis of significant researchers and notable papers. The study demonstrates the thematic evolution within the IIA field using LDA for topic modeling in conjunction with the earlier findings. It highlights the present state of research in inclusive technology and information retrieval, pointing out certain areas that need more investigation. The paper also outlines the difficulties and knowledge gaps in the IIA area, offering a guide for the next research projects. By providing a comprehensive overview of the IIA area, this study, unlike prior studies, bridges a significant gap in the literature and acts as an invaluable resource for researchers. The highlighted possible study directions also provide insightful information that can guide future scholarly activity.

8 Disclosure of Funding

None of the authors have received any funding or possess any competing interest that could have influenced the research reported in this paper.

References

Rudaiba Adnin and Maitraye Das. I look at it as the king of knowledge: How blind people use and understand generative ai tools. In *Proceedings of the 26th International ACM SIGACCESS Conference on Computers and Accessibility*, pages 1–14, 2024.

Herman Aguinis, Ryan K Gottfredson, and Thomas A Wright. Best-practice recommendations for estimating interaction effects using meta-analysis. *Journal of Organizational Behavior*, 32(8):1033–1043, 2011.

Abdul Ahad, Muhammad Fayaz, and Abdul Salam Shah. Navigation through citation network based on content similarity using cosine similarity algorithm. *International Journal of Database Theory and Application*, 9(5):9–20, 2016.

Leonelo Dell Anhol Almeida, Vânia Paula de Almeida Neris, Leonardo Cunha de Miranda, Elaine Cristina Saito Hayashi, and Maria Cecília Calani Baranauskas. Designing inclusive social networks: a participatory approach. In *Online Communities and Social Computing: Third International Conference, OCSC 2009, Held as Part of HCI International 2009, Proceedings 3*, pages 653–662. Springer, 2009.

Vânia Paula Almeida Neris, Frederico Fortuna, Rodrigo Bonacin, Tatiana Silva de Alencar, Luciano de Oliveira Neris, and M. Cecília C. Baranauskas. Addressing brazilian diversity in personal computing systems with a tailoring-based approach. *Personal and Ubiquitous Computing*, 25(2):297–319, Aug 2020. doi: 10.1007/s00779-020-01444-w.

Salha Alzahrani, Vasile Palade, Naomie Salim, and Ajith Abraham. Using structural information and citation evidence to detect significant plagiarism cases in scientific publications. *Journal of the American Society for Information Science and Technology*, 63(2): 286–312, Oct 2011. doi: 10.1002/asi.21651.

Zainab Amjad and Imran Ihsan. Verbnet based citation sentiment class assignment using machine learning. *International Journal of Advanced Computer Science and Applications*, 11(9), 2020.

Simplice A Asongu and Nicholas M Odhiambo. Basic formal education quality, information technology, and inclusive human development in sub-saharan africa. *Sustainable Development*, 27(3):419–428, 2019.

Steven Bird, Ewan Klein, and Edward Loper. *Natural language processing with Python: analyzing text with the natural language toolkit.* " O'Reilly Media, Inc.", 2009.

David M Blei, Andrew Y Ng, and Michael I Jordan. Latent dirichlet allocation. *Journal of machine Learning research*, 3(Jan):993–1022, 2003.

MAP Bovens et al. Information right: Citizenship in the information society. In *International Political Science Association World Congress Quebec*, 2000.

Danah M Boyd and Nicole B Ellison. Social network sites: Definition, history, and scholarship. *Journal of computer-mediated Communication*, 13(1):210–230, 2007.

Ana CM Brito, Filipi N Silva, and Diego R Amancio. Analyzing the influence of prolific collaborations on authors productivity and visibility. *Scientometrics*, 128(4):2471–2487, 2023.

A Bryne. Promoting the global information commons: A commentary on the library and information implications of the wsis declaration of principles" building the information society: a global challenge in the new millennium"(document wsis/pc-3/dt/6). 2005.

Christian Bühler and Constantine Stephanidis. European co-operation activities promoting design for all in information society technologies: Introduction to the special thematic session. In *International Conference on Computers for Handicapped Persons*, pages 80–87. Springer, 2004.

Elfreda A Chatman. The impoverished life-world of outsiders. *Journal of the American Society for information science*, 47(3):193–206, 1996.

Abon Chaudhuri. A visual technique to analyze flow of information in a machine learning system. *arXiv preprint arXiv:1908.00754*, 2019.

Chaomei Chen. The citespace manual. *College of Computing and Informatics*, 1(1):1–84, 2014.

Xin Chen, Britt Östlund, and Susanne Frennert. Digital inclusion or digital divide for older immigrants? a scoping review. In *International conference on human-computer interaction*, pages 176–190. Springer, 2020.

RV Chikate and SK Patil. Citation analysis of theses in library and information science submitted to university of pune: A pilot study. *Library Philosophy and Practice*, 222: 31–55, 2008.

Suraj Chiplunkar, Anany Maini, Dinesh Ram, Zixuan Zheng, and Yaxin Zheng. Drawxi: an accessible drawing tool for collaboration. In *Extended Abstracts of the 2019 CHI Conference on Human Factors in Computing Systems*, pages 1–6, 2019.

Blaise Cronin. Bibliometrics and beyond: some thoughts on web-based citation analysis. *Journal of Information science*, 27(1):1–7, 2001.

Simone Erbs da Costa, Carla Diacui Medeiros Berkenbrock, Lucas Eduardo Rosa de Freitas, and Fabíola Ferreira Sucupira Sell. ilibras: using assistive and collaborative technology to support the communication of deaf people. *IEEE Revista Iberoamericana de Tecnologias del Aprendizaje*, 14(1):11–21, 2019.

Vânia Paula de Almeida Neris, Frederico Fortuna, Rodrigo Bonacin, Tatiana Silva de Alencar, Luciano de Oliveira Neris, and M Cecília C Baranauskas. Addressing brazilian diversity in personal computing systems with a tailoring-based approach. *Personal and Ubiquitous Computing*, 25:297–319, 2021.

Francesca De Felice and Antonella Polimeni. Coronavirus disease (covid-19): a machine learning bibliometric analysis. *in vivo*, 34(3 suppl):1613–1617, 2020.

S De Groote. Measuring your impact: Impact factor, citation analysis, and other metrics: Citation analysis. *UIC Libraries Research Guides*, 2015.

Anjani Dhrangadhariya, Roger Hilfiker, Roger Schaer, and Henning Müller. Machine learning assisted citation screening for systematic reviews. *Studies in health technology and informatics*, 270:302–306, Winter 2020. doi: 10.3233/SHTI200171.

Naveen Donthu, Satish Kumar, Debmalya Mukherjee, Nitesh Pandey, and Weng Marc Lim. How to conduct a bibliometric analysis: An overview and guidelines. *Journal of business research*, 133:285–296, 2021.

Júlio Cesar dos Reis, Rodrigo Bonacin, and M Cecilia C Baranauskas. New perspectives for search in social networks-a challenge for inclusion. In *International Conference on Enterprise Information Systems*, volume 2, pages 53–62. SCITEPRESS, 2010.

Julio Cesar Dos Reis, Rodrigo Bonacin, and M Cecília C Baranauskas. Addressing universal access in social networks: an inclusive search mechanism. *Universal access in the information society*, 13:125–145, 2014.

Leo Egghe. Theory and practise of the g-index. *Scientometrics*, 69(1):131–152, 2006.

Gorka Epelde, Xabier Valencia, Eduardo Carrasco, Jorge Posada, Julio Abascal, Unai Diaz-Orueta, Ingo Zinnikus, and Christian Husodo-Schulz. Providing universally accessible interactive services through tv sets: implementation and validation with elderly users. *Multimedia Tools and Applications*, 67(2):497–528, Dec 2011. doi: 10.1007/s11042-011-0949-0.

Gorka Epelde, Xabier Valencia, Eduardo Carrasco, Jorge Posada, Julio Abascal, Unai Diaz-Orueta, Ingo Zinnikus, and Christian Husodo-Schulz. Providing universally accessible interactive services through tv sets: implementation and validation with elderly users. *Multimedia tools and applications*, 67:497–528, 2013.

Pierre Erbacher, Ludovic Denoyer, and Laure Soulier. Interactive query clarification and refinement via user simulation, 2022. URL https://arxiv.org/abs/2205.15918.

Chinwe N Ezeani, Scholastica C Ukwoma, Esther Gani, Prince J Igwe, and Chidimma G Agunwamba. Towards sustainable development goals: What role for academic libraries in nigeria in assuring inclusive access to information for learners with special needs? 2017.

Grace Fox and Regina Connolly. Mobile health technology adoption across generations: Narrowing the digital divide. *Information Systems Journal*, 28(6):995–1019, 2018.

Biying Fu, Abdenour Hadid, and Naser Damer. Generative ai in the context of assistive technologies: Trends, limitations and future directions. *Image and Vision Computing*, 154:105347, 2025.

Simon L Gay, Edwige Pissaloux, Katerine Romeo, and Ngoc-Tan Truong. F2t: a novel force-feedback haptic architecture delivering 2d data to visually impaired people. *IEEE Access*, 9:94901–94911, 2021.

Elise Gerich. Expanding the internet to a global environment but... how to get connected? *Computer Networks and ISDN Systems*, 23(1-3):43–46, 1991.

Fernando Benedet Ghisi, Regina Bóries, Santos Marcos, Denilson Sell, and Jean Varvakis. A reference ontology for digital scientific journals applied to systematic literature review processes. *Transinformação*, 24:91–101, Aug 2012. URL `https://www.scielo.br/j/tinf/a/z7NkZ7Lwz4HbHFQLTMFhhCp/?lang=en`.

Martin Grayson, Anja Thieme, Rita Marques, Daniela Massiceti, Ed Cutrell, and Cecily Morrison. A dynamic ai system for extending the capabilities of blind people. In *Extended Abstracts of the 2020 CHI Conference on Human Factors in Computing Systems*, pages 1–4, 2020.

Marie Michèle Grenon, Julie Ruel, Patrick Fougeyrollas, Claude L Normand, André C Moreau, Alejandro Romero-Torres, and Sylvie Gravel. Conceptualizing access to and understanding of information. *Universal Access in the Information Society*, 22(1):83–94, 2023.

Daniel Gruss, Raphael Spreitzer, and Stefan Mangard. Cache template attacks: Automating attacks on inclusive {Last-Level} caches. In *24th USENIX Security Symposium (USENIX Security 15)*, pages 897–912, 2015.

Xiaoxiao Guo, Hui Wu, Yu Cheng, Steven Rennie, Gerald Tesauro, and Rogerio Schmidt Feris. Dialog-based interactive image retrieval, 2018. URL `https://arxiv.org/abs/1805.00145`.

Anne-Wil Harzing and Satu Alakangas. Google scholar, scopus and the web of science: a longitudinal and cross-disciplinary comparison. *Scientometrics*, 106:787–804, 2016.

Anne M. Hayes and Jennae Bulat. *Disabilities Inclusive Education Systems and Policies Guide for Low- and Middle-Income Countries*. RTI Press, Research Triangle Park (NC), 2019. URL `https://www.ncbi.nlm.nih.gov/books/NBK554622/`.

Jorge E Hirsch. An index to quantify an individual's scientific research output. *Proceedings of the National academy of Sciences*, 102(46):16569–16572, 2005.

Cheng-An Hsieh, Cheng-Ping Hsieh, and Pu-Jen Cheng. Mr. right: Multimodal retrieval on representation of image with text, 2022. URL `https://arxiv.org/abs/2209.13764`.

Andreja Istenic Starcic and Spela Bagon. Ict-supported learning for inclusion of people with special needs: Review of seven educational technology journals, 1970–2011. *British Journal of Educational Technology*, 45(2):202–230, 2014.

Paul T Jaeger. Telecommunications policy and individuals with disabilities: Issues of accessibility and social inclusion in the policy and research agenda. *Telecommunications Policy*, 30(2):112–124, 2006.

Paul T Jaeger and Bo Xie. Developing online community accessibility guidelines for persons with disabilities and older adults. *Journal of Disability Policy Studies*, 20(1):55–63, 2009.

Hamed Jelodar, Yongli Wang, Chi Yuan, Xia Feng, Xiahui Jiang, Yanchao Li, and Liang Zhao. Latent dirichlet allocation (lda) and topic modeling: models, applications, a survey. *Multimedia tools and applications*, 78:15169–15211, 2019.

Kaixin Ji, Danula Hettiachchi, Flora D Salim, Falk Scholer, and Damiano Spina. Characterizing information seeking processes with multiple physiological signals. *arXiv (Cornell University)*, Jul 2024. doi: 10.1145/3626772.3657793.

Sonali Joshi, Padmalata V. Nistala, Hetal Jani, Prachi Sakhardande, and Trevor Dsouza. User-centered design journey for pattern development. *Proceedings of the 22nd European Conference on Pattern Languages of Programs*, Jul 2017. doi: 10.1145/3147704.3147730.

Diane Kelly and Cassidy R Sugimoto. A systematic review of interactive information retrieval evaluation studies, 1967–2006. *Journal of the American Society for Information Science and Technology*, 64(4):745–770, 2013.

Beatrice Wamaitha Kiruki and Stephen Mudogo Mutula. Information communication technology (ict) use for information access by visually and physically impaired persons in public university libraries in kenya. *International Journal of Knowledge Content Development & Technology*, 13(1), 2023.

Vasiliki Kladouchou, Stephann Makri, Sylwia Frankowska-Takhari, Timothy Neate, Andrew MacFarlane, Stephanie Wilson, and Abi Roper. "the internet is hard. is words": Investigating information search difficulties experienced by people with aphasia and strategies for combatting them. In *Proceedings of the 2025 CHI Conference on Human Factors in Computing Systems*, CHI '25. Association for Computing Machinery, 2025. doi: 10.1145/3706598.3713808.

Cary LaCheen. Achy breaky pelvis, lumber lung and juggler's despair: The portrayal of the americans with disabilities act on television and radio. *Berkeley J. Emp. & Lab. L.*, 21:223, 2000.

Jolie CY Lam and Matthew KO Lee. Digital inclusiveness–longitudinal study of internet adoption by older adults. *Journal of Management Information Systems*, 22(4):177–206, 2006.

Guanyu Li, Dian Li, and Tang Tang. Bibliometric review of design for digital inclusion. *Sustainability*, 15(14):10962, 2023.

Chuan Liu, Rong Yu, Jixiang Zhang, Shuchun Wei, Fumin Xue, Yingyun Guo, Pengzhan He, Lining Shang, and Weiguo Dong. Research hotspot and trend analysis in the diagnosis of inflammatory bowel disease: A machine learning bibliometric analysis from 2012 to 2021. *Frontiers in Immunology*, 13:972079, 2022.

Paloma Martínez, Alberto Ramos, and Lourdes Moreno. Exploring large language models to generate easy to read content. *Frontiers in Computer Science*, 6:1394705, 2024.

Chris McGinley and Hua Dong. Accessing user information for use in design. *Lecture notes in computer science*, page 116–125, Jan 2009. doi: 10.1007/978-3-642-02707-9_13.

Rupendra Nath Mitra and Dharma P Agrawal. 5g mobile technology: A survey. *ICT express*, 1(3):132–137, 2015.

Zachary Munn, Micah DJ Peters, Cindy Stern, Catalin Tufanaru, Alexa McArthur, and Edoardo Aromataris. Systematic review or scoping review? guidance for authors when choosing between a systematic or scoping review approach. *BMC medical research methodology*, 18:1–7, 2018.

Mark A. Neerincx, Anita H. M. Cremers, Judith M. Kessens, David A. van Leeuwen, and Khiet P. Truong. Attuning speech-enabled interfaces to user and context for inclusive design: technology, methodology and practice. *Universal Access in the Information Society*, 8(2):109–122, Aug 2008. doi: 10.1007/s10209-008-0136-x.

David Nemer. From digital divide to digital inclusion and beyond. *The Journal of Community Informatics*, 11(1), 2015.

VP de A Neris, LD Almeida, LC Miranda, E Hayashi, and MCC Baranauskas. Towards a socially-constructed meaning for inclusive social network systems. In *International Conference on Informatics and Semiotics in Organisations. Beijing*, pages 247–254, 2009.

AF Newell, G Gregor, M Morgan, G Pullin, and C Macaulay. User-sensitive inclusive design in universal access in the information society august 2011. *Volume*, 10:235–24, 2011.

Julie M. Nightingale and Gill Marshall. Citation analysis as a measure of article quality, journal influence and individual researcher performance. *Radiography*, 18(2):60–67, 2012. ISSN 1078-8174. doi: 10.1016/j.radi.2011.10.044.

Pippa Norris. *Digital divide: Civic engagement, information poverty, and the Internet worldwide.* Cambridge university press, 2001.

C Wendy Olphert, Leela Damodaran, and AJ May. Towards digital inclusion–engaging older people in the 'digital world'. In *Accessible Design in the Digital World Conference 2005*, pages 1–7, 2005.

Denis Luiz Marcello Owa et al. Identification of topics from scientific papers through topic modeling. *Open Journal of Applied Sciences*, 10(04):541, 2021.

Kwangil Park, June Seok Hong, and Wooju Kim. A methodology combining cosine similarity with classifier for text classification. *Applied Artificial Intelligence*, 34(5):396–411, 2020.

Cheryl Parsons and Steven F Hick. Moving from the digital divide to digital inclusion. *Currents: Scholarship in the Human Services*, 7(2), 2008.

Marta Perez-Escolar and Fernando Canet. Research on vulnerable people and digital inclusion: toward a consolidated taxonomical framework. *Universal Access in the Information Society*, 22(3):1059–1072, 2023.

Andrew Pinder. Report of the digital inclusion panel. *The Stationery Office*, 2004.

Raminta Pranckutė. Web of science (wos) and scopus: The titans of bibliographic information in today's academic world. *Publications*, 9(1):12, 2021.

Derek J De Solla Price. *Little science, big science*. Columbia University Press, 1963.

OECD. Publishing. *The survey of adult skills: Reader's companion*. OECD Publishing, 2013.

NR Raji, CL Biji, and V Vineetha. Multi-modal generative ai for people with disabilities. In *Multimodal Generative AI*, pages 271–296. Springer, 2025.

Hassan Raza, M Faizan, Ahsan Hamza, Mushtaq Ahmed, and Naeem Akhtar. Scientific text sentiment analysis using machine learning techniques. *International Journal of Advanced Computer Science and Applications*, 10(12), 2019.

Júlio Cesar Reis, Rodrigo Bonacin, and Maria Cecília Calani Baranauskas. Prospecting an inclusive search mechanism for social network services. In *Enterprise Information Systems*. 2010.

Julio Cesar Reis, Rodrigo Bonacin, and M. Cecília C. Baranauskas. Addressing universal access in social networks: an inclusive search mechanism. *Universal Access in the Information Society*, Feb 2013. doi: 10.1007/s10209-013-0290-7.

General Assembly Resolution et al. Transforming our world: the 2030 agenda for sustainable development. *UN Doc. A/RES/70/1 (September 25, 2015)*, 2015.

Leslie S. Adriaanse and Chris Rensleigh. Web of science, scopus and google scholar: A content comprehensiveness comparison. *The Electronic Library*, 31(6):727–744, 2013.

Amartya Sen. Development as freedom. *Development in Practice-Oxford-*, 10(2):258–258, 2000.

Laurianne Sitbon, Gerd Berget, Margot Brereton, et al. Perspectives of neurodiverse participants in interactive information retrieval. *Foundations and Trends® in Information Retrieval*, 17(2):124–243, 2023.

Hannah Snyder. Literature review as a research methodology: An overview and guidelines. *Journal of business research*, 104:333–339, 2019.

Karen Sparck Jones. A statistical interpretation of term specificity and its application in retrieval. *Journal of documentation*, 28(1):11–21, 1972.

Mega Subramaniam, Rebecca Oxley, and Christie Kodama. School librarians as ambassadors of inclusive information access for students with disabilities. *School library research*, 16, 2013.

Dandison C Ukpabi and Heikki Karjaluoto. Consumers' acceptance of information and communications technology in tourism: A review. *Telematics and Informatics*, 34(5): 618–644, 2017.

Jan Van Dijk and Kenneth Hacker. The digital divide as a complex and dynamic phenomenon. *The information society*, 19(4):315–326, 2003.

Jan AGM Van Dijk. Digital divide research, achievements and shortcomings. *Poetics*, 34 (4-5):221–235, 2006.

Nees Van Eck and Ludo Waltman. Software survey: Vosviewer, a computer program for bibliometric mapping. *scientometrics*, 84(2):523–538, 2010.

Nees Jan van Eck and Ludo Waltman. Accuracy of citation data in web of science and scopus. *arXiv preprint arXiv:1906.07011*, 2019.

Polyxeni Vassilakopoulou and Eli Hustad. Bridging digital divides: A literature review and research agenda for information systems research. *Information Systems Frontiers*, 25(3): 955–969, 2023.

Viswanath Venkatesh, Michael G Morris, Gordon B Davis, and Fred D Davis. User acceptance of information technology: Toward a unified view. *MIS quarterly*, pages 425–478, 2003.

Hanna M Wallach, Iain Murray, Ruslan Salakhutdinov, and David Mimno. Evaluation methods for topic models. In *Proceedings of the 26th annual international conference on machine learning*, pages 1105–1112, 2009.

Jinlong Wang, Can Wen, Shunyao Wu, and Huy Quan Vu. A visual mining system for theme development evolution analysis of scientific literature. *International Journal of Digital Content Technology and its Applications*, 4(3):215–223, 2010.

Vishanth Weerakkody, Yogesh K Dwivedi, Ramzi El-Haddadeh, Ahlam Almuwil, and Ahmad Ghoneim. Conceptualizing e-inclusion in europe: An explanatory study. *Information Systems Management*, 29(4):305–320, 2012.

Zhaohan Xiong, Tong Liu, Gary Tse, Mengqi Gong, Patrick A Gladding, Bruce H Smaill, Martin K Stiles, Anne M Gillis, and Jichao Zhao. A machine learning aided systematic review and meta-analysis of the relative risk of atrial fibrillation in patients with diabetes mellitus. *Frontiers in physiology*, 9:835, 2018.

Neng Zhang, Qiao Huang, Xin Xia, Ying Zou, David Lo, and Zhenchang Xing. Chatbot4qr: Interactive query refinement for technical question retrieval. *IEEE Transactions on Software Engineering*, page 1–1, 2020. doi: 10.1109/tse.2020.3016006.

Yue Zheng, Lei Yu, Junmian Chen, Tianyu Xia, Yuanyuan Yin, Shan Wang, and Haiming Liu. Inclusive design insights from a preliminary image-based conversational search systems evaluation, 2024. URL `https://arxiv.org/abs/2403.19899`.

Yutao Zhu, Huaying Yuan, Shuting Wang, Jiongnan Liu, Wenhan Liu, Chenlong Deng, Haonan Chen, Zheng Liu, Zhicheng Dou, and Ji-Rong Wen. Large language models for information retrieval: A survey. *arXiv preprint arXiv:2308.07107*, 2023.

John Zimmerman, Robin E Soler, James Lavinder, Sarah Murphy, Charisma Atkins, LaShonda Hulbert, Richard Lusk, and Boon Peng Ng. Iterative guided machine learning-assisted systematic literature reviews: a diabetes case study. *Systematic Reviews*, 10(1): 1–8, 2021.

Information Retrieval Research 1 (2025) 197-220 Submitted 10/24; Revised 03/25; Published 09/25

Evaluating Dense Model-based Approaches for Multimodal Medical Case Retrieval

Catarina Pires UP201907925@FE.UP.PT
INESC TEC, Faculty of Engineering of the University of Porto,
Rua Dr. Roberto Frias, s/n 4200-465 Porto, Portugal

Sérgio Nunes SSN@FE.UP.PT
INESC TEC, Faculty of Engineering of the University of Porto,
Rua Dr. Roberto Frias, s/n 4200-465 Porto, Portugal

Luís Filipe Teixeira LUISFT@FE.UP.PT
INESC TEC, Faculty of Engineering of the University of Porto,
Rua Dr. Roberto Frias, s/n 4200-465 Porto, Portugal

Editor: Ben He

Abstract

Medical case retrieval plays a crucial role in clinical decision-making by enabling healthcare professionals to find relevant cases based on patient records, diagnostic images, and textual descriptions. Given the inherently multimodal nature of medical data, effective retrieval requires models that can bridge the gap between different modalities. Traditional retrieval approaches often rely on unimodal representations, limiting their ability to capture cross-modal relationships. Recent advances in dense model-based techniques have shown promise in overcoming these limitations by encoding multimodal information into a shared latent space, facilitating retrieval based on semantic similarity. This paper investigates the potential of dense models to enhance multimodal search systems. We evaluate various dense model-based approaches to assess which model characteristics have the greatest impact on retrieval effectiveness, using the medical case-based retrieval task from ImageCLEFmed 2013 as a benchmark. Our findings indicate that different dense model approaches substantially impact retrieval effectiveness, and that applying the CombMAX fusion method to combine their output results further improves effectiveness. Extending context length, however, yielded mixed results depending on the input data. Additionally, domain-specific models—those trained on medical data—outperformed general models trained on broad, non-specialized datasets within their respective fields. Furthermore, when text is the dominant information source, text-only models surpassed multimodal models.

Keywords: Medical Search, Multimodal Retrieval, Dense Retrieval

1 Introduction

The increasing volume of digital medical records and imaging data has made medical case retrieval an important tool for clinical decision-making (Sivarajkumar et al., 2024). Physicians and researchers often need to retrieve relevant cases that share similar characteristics with a given medical query, which may include text and images. This retrieval supports comparing diagnostic outcomes, exploring treatment options, or gaining insights from historical cases. However, the multimodal nature of medical case data—often consisting of

WWW: https://irrj.org DOI: 10.54195/irrj.19769

both textual descriptions (e.g., reports, diagnoses) and visual content (e.g., radiographs, MRIs)—poses significant challenges for retrieval systems. Traditional information retrieval methods frequently rely on text-based searches, which may fail to capture the intricate relationships between textual and visual data fully. The challenge lies in effectively integrating these modalities to improve retrieval accuracy and relevance, i.e., to retrieve results that are not only topically relevant but also visually and semantically aligned with the query.

Recent advances in deep learning, particularly in dense model-based approaches, offer new opportunities for multimodal retrieval. Multimodal models encode multiple modalities into a shared latent space, allowing for retrieval based on cross-modal semantic similarity.

In this paper, we explore whether dense multimodal models can outperform traditional retrieval methods by addressing the challenge of integrating text and images through dense representations within a multimodal framework. To evaluate the effectiveness of various dense model-based approaches, we focus on the case-based retrieval task from ImageCLEFmed 2013. The ImageCLEFmed 2013 dataset is distinguished by its multimodal collection of text and images, along with relevance judgments that are key for effectiveness evaluation. To guide our investigation, we pose the following research questions:

RQ1 Which characteristics of dense models have the greatest impact on retrieval effectiveness in multimodal search systems?

RQ2 How does the effectiveness of dense multimodal models compare to traditional search systems in medical case retrieval, and what factors influence their relative effectiveness?

To answer **RQ1**, we conduct a series of experiments, exploring different result fusion methods and dense models. These experiments analyze how factors like context length and domain specificity, particularly within the medical domain, influence retrieval effectiveness. Our findings show that dense model approaches significantly influence results, with the CombMAX fusion method yielding the best effectiveness, specialized domain-specific models surpassing general ones, context length extensions producing mixed effects, and text-based models outperforming multimodal models when text is the primary information source. To answer **RQ2**, we perform a comparative analysis against the leading submissions from the ImageCLEFmed 2013 case-based retrieval task, which predominantly employed traditional sparse approaches. The results suggest that dense retrieval holds great potential, particularly for improving semantic similarity searches across different modalities.

2 Related work

Our work explores dense retrieval models for multimodal medical ad hoc search, a field that builds upon two key areas: medical case retrieval and multimodal fusion. Medical case retrieval has evolved from traditional text-based methods, such as keyword search, query expansion, and relevance feedback, to multimodal systems that integrate various types of medical data. While systems like PubMed,[1] which use keywords, Boolean operators, and MeSH terms to refine search results, remain foundational in biomedical information

[1] https://pubmed.ncbi.nlm.nih.gov/

retrieval (Jin et al., 2024; Lu et al., 2009), they are primarily text-centric. In contrast, real-world clinical scenarios often demand the integration of heterogeneous evidence, including both textual and visual data, to support accurate diagnosis and treatment.

Multimodal retrieval systems integrate multiple evidence types by combining information from different modalities. Early examples include MedGIFT (Group, 2009), which allows independent text or image-based searches, and img(Anaktisi) (Zagoris et al., 2009), focused on image-based retrieval across medical datasets. Later approaches, like NovaMed-Search (Mourão and Martins, 2013), fuse textual and visual data to improve case retrieval performance. The ImageCLEF medical retrieval task (Müller and Kalpathy-Cramer, 2010) introduced fusion techniques specifically for case retrieval, emphasizing the integration of textual and visual information. Comparisons of different fusion methods in medical case retrieval tasks were further explored by Garcia Seco de Herrera et al. (2015).

Multimodal fusion systems combine information from diverse sources to support decision-making by creating a context-aware representation. Fusion techniques include early fusion, late fusion, and hybrid fusion, which blend aspects of both approaches (Bayoudh et al., 2022; Zhang et al., 2021; Bruni et al., 2014).

Early fusion combines raw data or features from multiple modalities at the beginning of the data processing pipeline, allowing a more thorough cross-modal correlation analysis. However, it often requires data standardization, such as dimensionality reduction, to ensure compatibility across modalities. In contrast, late fusion integrates information from different modalities at the decision stage, where each modality is processed independently before merging results through operations like concatenation or averaging (Feng et al., 2021). This approach offers more flexibility and can better handle individual modality errors but often fails to capture complex cross-modal dependencies and interactions (Zhang et al., 2021).

Recent studies have explored deep learning-based multimodal fusion for medical tasks. Li et al. (2024) reviewed deep learning-based information fusion techniques for medical image classification, highlighting their role in enhancing medical decision-making. Similarly, Cui et al. (2023) provided a comprehensive review of deep multimodal fusion of image and non-image data for disease diagnosis and prognosis, demonstrating the growing importance of multimodal techniques in biomedical applications. These works underscore the need for optimized fusion techniques tailored to medical retrieval tasks.

In this work, we focus on late fusion strategies to aggregate individual results (e.g., derived from different modalities) into a final score. This choice is motivated by the desire to leverage the strengths of individual modality-specific models while maintaining flexibility in handling different data types and potential errors within each modality. Within late fusion, there are score-based methods (e.g., CombSUM, CombMNZ) that merge normalized scores and rank-based methods (e.g., CombMAX) that prioritize document order (Hsu and Taksa, 2005). Rank-based methods are further categorized into positional (e.g., Borda Fuse (Aslam and Montague, 2001), and Reciprocal Rank Fusion (Cormack et al., 2009)) and majoritarian (e.g., Condorcet (Montague and Aslam, 2002)), with positional approaches assigning votes by rank and majoritarian methods using pairwise comparisons between documents.

Despite extensive exploration of image–text multimodal models, their applications in the biomedical field remain under-researched, particularly in areas such as clinical case retrieval. Guo et al. (2024) provided a detailed survey on advancements in these models, emphasizing their impact on biomedical multimodal technologies. A major breakthrough in multimodal retrieval was the introduction of CLIP (Radford et al., 2021), which learns joint representations of images and text through contrastive learning. CLIP has shown strong zero-shot and few-shot learning capabilities across diverse image-text tasks, making it a central component in many recent multimodal retrieval systems, including our work. Although more recent approaches such as ColPali (Faysse et al., 2025) have emerged, CLIP remains a foundational model that influenced the development of our system, particularly given its relevance at the time of our experiments.

Building on these insights, our study addresses the specific challenges in medical case-based retrieval tasks by evaluating various dense model-based approaches. The findings contribute to a deeper understanding of which dense multimodal approaches are most effective for biomedical applications, contributing to the broader development of image-text retrieval technologies in this domain.

3 Methodology

This section presents the methodology for achieving the research goals, including defining the retrieval task, presenting the experimental pipeline, and detailing the experimental variables considered and how they will be varied to assess their impact on the final results.

3.1 ImageCLEFmed 2013 case-based retrieval task

We selected the case-based medical retrieval task from ImageCLEFmed 2013 (Garcia Seco de Herrera et al., 2013) due to its multimodal collection of text and images and the availability of relevance judgments, which are crucial for evaluating a search system. To our knowledge, this was the only available dataset with these characteristics, making it our best choice. The task simulates a clinician's diagnostic workflow by finding articles from a vast collection of biomedical literature (PubMed Central[2]) that could aid in differential diagnosis, based on a given case description and images of a patient's case. The dataset includes 75,000 articles and 35 query topics (i.e., cases), all following a well-defined structure, and a total of 300,000 images. Each article is structured into sections such as the title, author names, abstract, full text, figures, and captions, while the query topics, also divided into sections, contain a detailed case description and several relevant images.

Our analysis of the dataset revealed significant heterogeneity in the textual component, with text sections ranging from concise titles (average of 21 tokens) to fulltexts (up to 90,605 tokens). For the visual component, it was found that images in relevant articles differ notably from query images. Article images are mostly graphs and charts, while query images are medical exam images, creating a disadvantage for visual retrieval. Compound images, which contain multiple sub-images within a single frame, also add complexity.

In ImageCLEF, the articles were evaluated for relevance based on their contribution to differential diagnosis, using a three-point scale: relevant, partly relevant, or non-relevant (Kalpathy-Cramer et al., 2015). Analysis showed that physicians prioritized textual information over visual content in their decision-making. The images in relevant articles differed significantly from the query images, suggesting that visual content played a secondary role in their relevance assessments. Consequently, it is unlikely that a system relying solely on

[2] https://www.ncbi.nlm.nih.gov/pmc/

visual information could effectively retrieve relevant articles. Furthermore, Garcia Seco de Herrera et al. (2017) observed that incorporating visual data into a multimodal approach did not enhance retrieval effectiveness for the specific topics of the ImageCLEF task.

A total of 15,028 relevance judgments were made across all query topics, with only 0.57% of the collection judged per topic, as detailed in Section 5.7. The limited judgments, especially the lack of relevant documents, posed challenges, as also noted by Garcia Seco de Herrera et al. (2015). This impacted top results in the case-based task with relatively low scores, where MAP scores ranged from 0.0281 to 0.2429, depending on retrieval type (visual, mixed, or textual), as discussed in Section 5.6. Mitigation efforts are explored in Section 6.

3.2 Experimental multimodal retrieval pipeline

To systematically experiment with and evaluate dense model-based approaches, we built a functional prototype of the retrieval system. The source code is openly available on GitHub.[3] The workflow is organized into five key steps:

1. **Dataset collection and article encoding**: In this step, raw data is processed and encoded into dense embeddings.
2. **Storage and indexing of embeddings**: The encoded articles are represented in an embedding space and indexed for efficient retrieval. We used Faiss[4] (Johnson et al., 2021) version 1.8.0 with the GPU implementation, employing its HNSW index with squared Euclidean (L2) distance, which preserves ranking while improving efficiency by avoiding square root calculations.
3. **Query encoding**: Query documents are transformed into embeddings followed by a similarity search, using Faiss, against the pre-computed indexed embeddings of the articles for retrieval.
4. **Results fusion**: Results from multiple retrieval approaches are combined, matching each section of the query with corresponding sections in the article, which may contain textual or visual information.
5. **Retrieval**: A ranked list of documents based on the fused results is produced.

3.3 Experimental variables

With the functional pipeline established as the foundation for experimentation, we have the flexibility to adjust and refine various elements of the system. Through this process, we systematically assess how different changes impact the effectiveness of the retrieval task.

3.3.1 Results fusion

Since the documents consist of multiple distinct sections, each section of the article document must be compared with each section of the topic document, resulting in multiple

[3] https://github.com/catarinaopires/eval-multimodal-medical-case-retrieval

[4] https://faiss.ai

ranked lists that require results fusion. For instance, when comparing a topic description with article images, we need to compute the similarity between the textual description and each article image, generating multiple ranked lists. To obtain a single final ranking, we then apply one of the result fusion methods.

We experiment with CombSUM, CombMAX, and CombMNZ, given their demonstrated effectiveness within the Comb family, introduced by Shaw and Fox (1994). For each document i, the score after fusion can be computed as:

$$CombSUM(i) = \sum_{k=1}^{N(i)} S_k(i), \tag{1}$$

$$CombMAX(i) = max(S), \forall S \subset D_i, \tag{2}$$

$$CombMNZ(i) = N(i) * CombSUM(i), \tag{3}$$

where $S_k(i)$ is the score of the i-th document in the k-th result list, $N(i)$ refers to the number of times a document appears in the result lists, and D_i denotes the set of scores (S) assigned to document i across all result lists in which it appears.

3.3.2 Models

Our study explores the use of dense models in multimodal search, focusing on how various model architectures and capabilities impact retrieval effectiveness, particularly when handling both text and image data. The models employed, primarily using HuggingFace[5] implementations in Python, include CLIP's ViT-B/16 variant (Radford et al., 2021), Long-Clip B/16 (Zhang et al., 2024), a fine-tuned version of CLIP that extends the token capacity from 77 to 248 for longer text–image pairs, PubMedCLIP (Eslami et al., 2021), a fine-tuned CLIP model for medical image-caption tasks, as well as Llama 3 (Grattafiori et al., 2024) with 8 billion parameters and LLaVA-1.5 (Liu et al., 2023) with 7 billion parameters.

One of the challenges in multimodal retrieval is managing different modalities. A key decision is whether to use a multimodal model that processes both text and images together or to handle each modality separately using unimodal models. While CLIP effectively manages visual data, it struggles with long texts due to its limited token capacity. Aggravating this problem, Zhang et al.'s experimental findings suggest that the effective length of text that CLIP can handle optimally is no more than 20 tokens, beyond which it struggles to utilize the additional information effectively. This limitation can result in the loss of crucial information necessary for accurate retrieval. Although models like LongCLIP attempt to address CLIP's token limit by increasing token capacity, alternatives like large language models (e.g., Llama 3) may be better suited for handling lengthy text inputs.

The main strength of multimodal models lies in their capacity to encapsulate data across various modalities within a shared latent space, facilitating comparison and relationship establishment. In contrast, employing unimodal models would compromise this key capability, impeding cross-modality comparisons. Nevertheless, it is feasible to mitigate the limitation of unimodal models in cross-modality comparisons by homogenizing data modalities into a

[5] `https://huggingface.co/`

singular format, like text. A potential solution is to convert visual data into text through image descriptions, allowing unimodal models to handle the transformed data exclusively in text format. Using a multimodal generative model such as LLaVA, images are translated into textual descriptions, which can then be compared with existing text-based content. Transforming visual data into text-based representations facilitates the comparison of images and text within a unified latent space, even when using text-only models.

Finally, our study evaluates the potential benefits of domain-specific models, such as PubMedCLIP, over general-purpose models like CLIP for improving multimodal search systems in medical information retrieval.

4 Experimental setup

This section presents the experimental setup, outlining the planned experiments along with their objectives and focus. It also presents the evaluation metrics used to assess the impact of each approach and details the computational resources employed for execution.

4.1 Experiments

To address the outlined research questions, we conducted five experiments, focusing on model variations to investigate how different factors influence retrieval effectiveness.

Exp. 1 **Results fusion effectiveness:** Firstly, we will evaluate the impact of using different results fusion approaches, such as CombSUM, CombMAX, and CombMNZ, by applying them to retrieval outputs from CLIP and comparing their effectiveness.

Exp. 2 **Effect of context length:** Our second experiment will examine whether the context length of a model influences the obtained results. To explore this, we will compare the effectiveness of the CLIP model with the LongCLIP model.

Exp. 3 **Domain-specific model effectiveness:** Our third experiment will investigate whether a domain-specific model can outperform a general-purpose model. This experiment will involve comparing the effectiveness of PubMedCLIP against CLIP.

Exp. 4 **Unimodal vs. Multimodal effectiveness:** Our fourth experiment examines whether a unimodal approach can outperform a multimodal baseline with the dataset at hand. This will involve comparing the outcomes of Llama against CLIP.

Exp. 5 **Dominant data type approach:** The fifth experiment assesses whether selecting a dominant data type (text) and converting visual content to a textual representation can outperform the baseline, which uses the original multimodal data. To investigate this, we will compare the effectiveness of searches using the existing topic sections with those using LLaVa's generated topic image descriptions. This comparison will involve all utilized models, not just the text model Llama.

4.2 Measure

Following established methodologies and metrics from ImageCLEFmed 2013, we report MAP as the primary metric, along with GM-MAP (Geometric Mean, or GMAP), bpref, and P@10/@30 as complementary metrics. The highest scores in each column are bolded, and statistically significant results are marked in the tables, based on a two-tailed paired permutation test with 100,000 permutations. A Holm-Bonferroni correction was applied at the 0.05 significance level (95% confidence interval) to account for multiple comparisons when evaluating the effectiveness of different dense model-based approaches on the selected dataset. In addition, we report effect sizes (Cohen's d_z), standard errors (SE), and 95% confidence intervals (CI), computed for each comparison relative to its respective baseline.

4.3 Computational resources

Initial experiments were conducted on a server equipped with two NVIDIA GeForce RTX 2080 Ti GPUs, each with 11GB of VRAM, which provided sufficient resources for running smaller models like CLIP. As the complexity of the experiments increased, the computational tasks were migrated to a more advanced computing environment managed by SLURM (Jette and Wickberg, 2023). This setup featured multiple GPUs, including NVIDIA Tesla V100 and NVIDIA A100 models, with 32GB and 80GB of VRAM, respectively. Some steps were run without GPUs and on the less advanced setup to minimize resource usage when possible.

5 Results

The overall results for all experiments are summarized in Table 1, with the scores presented as averages to provide an overview of effectiveness across all topics. Additionally, effect sizes, standard errors, and 95% confidence intervals are reported for all statistically significant comparisons ($p < 0.05$, Holm-Bonferroni corrected) in Table 2. To evaluate each proposed approach, we conducted individual searches for each section of the topic documents against each section of the article documents, testing all possible combinations. Thus, the results tables are organized by topic section (Description, Images) and article section (Title, Abstract, Fulltext, Images, Captions). The scores represent the outcomes of comparing each section from the topic (left-most label) to each section from the article (row-label).

5.1 Results fusion effectiveness

The first experiment examines the impact of different result fusion methods—CombSUM, CombMAX, and CombMNZ—without altering the underlying CLIP model. CombSUM serves as the baseline for comparison, and the effectiveness results for each method are presented in the top section of Table 1. Since topic descriptions and article titles, abstracts, and fulltexts are directly compared, results fusion is unnecessary. Therefore, CombMAX and CombMNZ scores remain unchanged from the baseline and are omitted from the table.

CombMAX consistently outperforms the baseline, particularly when comparing topic images to article images and captions. The medium effect size indicates that the observed improvement is likely to be meaningful in practice, even if not large, indicating better prioritization of relevant documents at the top results. In contrast, CombMNZ produces results nearly identical to CombSUM, with no meaningful improvements observed.

Table 1: Summary of experimental results. Statistically significant scores are marked with an asterisk (*) based on a two-tailed paired permutation test with 100,000 permutations, using Holm-Bonferroni correction at the 0.05 significance level. Significance is computed relative to the baseline (CLIP CombSUM in Experiment 1, and CLIP CombMAX for the remaining). In Experiment 5 (in “Gen. I. Desc.” rows), statistically significant differences relative to the description and image baselines are marked with ‘d’ and ‘i’, respectively. Highest scores per column are bolded.

		CLIP CombSUM$^{\text{Exp. 1}}$					**CLIP CombMNZ**$^{\text{Exp. 1}}$					**CLIP CombMAX**$^{\text{Exp. 1-5}}$				
		MAP	GM-MAP	bpref	P10	P30	MAP	GM-MAP	bpref	P10	P30	MAP	GM-MAP	bpref	P10	P30
Description	Title	**0.0396**	**0.0074**	**0.1028**	**0.0714**	**0.0543**	-	-	-	-	-	-	-	-	-	-
	Abstract	0.0177	0.0020	0.0799	0.0371	0.0305	-	-	-	-	-	-	-	-	-	-
	Fulltext	0.0280	0.0022	0.1013	0.0343	0.0324	-	-	-	-	-	-	-	-	-	-
	Images	0.0096	0.0022	0.0554	0.0143	0.0143	0.0096	**0.0022**	0.0554	0.0143	0.0143	0.0142	0.0026	0.0593	0.0143	0.0229
	Captions	0.0204	0.0018	0.0664	0.0371	0.0295	**0.0204**	0.0018	**0.0664**	**0.0371**	**0.0295**	0.0370*	**0.0031**	0.0941	**0.0629**	**0.0467***
Images	Title	0.0074	0.0015	0.0470	0.0200	0.0124	0.0074	0.0015	0.0470	0.0200	0.0124	0.0164*	0.0027	0.0597	0.0286	0.0210
	Abstract	0.0048	0.0006	0.0420	0.0171	0.0095	0.0048	0.0006	0.0420	0.0171	0.0095	0.0181	0.0011	0.0725	0.0343	0.0276
	Fulltext	0.0042	0.0007	0.0465	0.0086	0.0067	0.0042	0.0007	0.0465	0.0086	0.0067	0.0140	0.0012	0.0674	0.0171	0.0171
	Images	0.0170	0.0018	0.0546	0.0143	0.0105	0.0100*	0.0015	0.0510	0.0086	0.0124	**0.0397***	0.0034	**0.0963**	0.0486*	0.0305*
	Captions	0.0076	0.0015	0.0480	0.0200	0.0143	0.0076	0.0015	0.0480	0.0200	0.0143	0.0212*	0.0028	0.0718	0.0257	0.0267
Gen. I. Desc.	Title	-	-	-	-	-	-	-	-	-	-	0.0048^{di}	0.0002	0.0341^{d}	0.0086^{d}	0.0086^{d}
	Abstract	-	-	-	-	-	-	-	-	-	-	0.0128	0.0003	0.0537	0.0171	0.0095^{d}
	Fulltext	-	-	-	-	-	-	-	-	-	-	0.0073^{di}	0.0002	0.0435^{d}	0.0200	0.0095^{d}
	Images	-	-	-	-	-	-	-	-	-	-	0.0085^{i}	0.0016	0.0599	0.0086^{i}	0.0143^{i}
	Captions	-	-	-	-	-	-	-	-	-	-	0.0050^{di}	0.0007	0.0482^{d}	0.0086^{d}	0.0067^{d}

		LongCLIP CombMAX$^{\text{Exp. 2}}$					**PubMedCLIP CombMAX**$^{\text{Exp. 3}}$					**Llama CombMAX**$^{\text{Exp. 4}}$				
Description	Title	0.0122*	0.0010	0.0671	0.0314	0.0238*	0.0437	0.0047	0.1122	**0.0686**	**0.0562**	0.0304	0.0067	0.0771	0.0714	0.0419
	Abstract	0.0304	**0.0035**	0.0944	**0.0514**	0.0314	0.0504	**0.0069**	**0.1180**	0.0629	0.0457	0.0457*	0.0119	0.0947	0.0771	0.0629*
	Fulltext	**0.0370**	0.0019	**0.1046**	0.0457	**0.0352**	0.0184	0.0016	0.0866	0.0457	0.0314	0.0682*	**0.0141**	0.1152	0.0971*	**0.0771***
	Images	0.0192	0.0035	0.0748	0.0343	0.0210	0.0220	0.0014	0.0651	0.0429	0.0371	-	-	-	-	-
	Captions	0.0209	0.0033	0.0687	0.0486	0.0333	0.0344	0.0064	0.0869	0.0600	0.0524	**0.0686**	0.0123	**0.1171**	**0.1057**	0.0667
Images	Title	0.0184	0.0023	0.0644	0.0371	0.0295	0.0102	0.0007	0.0705	0.0200	0.0143	-	-	-	-	-
	Abstract	0.0189	0.0026	0.0618	0.0200	0.0267	0.0176	0.0016	0.0832	0.0143	0.0143	-	-	-	-	-
	Fulltext	0.0106	0.0008	0.0548	0.0171	0.0133	0.0101	0.0006	0.0655	0.0171	0.0133	-	-	-	-	-
	Images	0.0305	0.0030	0.0913	0.0371	0.0324	**0.0530**	0.0057	0.1011	0.0714	0.0429	-	-	-	-	-
	Captions	0.0114	0.0020	0.0560	0.0057	0.0152	0.0224	0.0041	0.0760	0.0457	0.0314	-	-	-	-	-
Gen. I. Desc.	Title	0.0002^{di}	0.0000	0.0245^{d}	0.0000^{di}	0.0000^{di}	0.0087^{d}	0.0005	0.0416^{d}	0.0114^{d}	0.0124^{d}	0.0016^{d}	0.0002	0.0354^{d}	0.0029^{d}	0.0029^{d}
	Abstract	0.0035^{di}	0.0003	0.0432^{d}	0.0086^{d}	0.0086^{d}	0.0150^{d}	0.0012	0.0669^{d}	0.0257	0.0181^{d}	0.0027^{d}	0.0003	0.0337^{d}	0.0029^{d}	0.0057^{d}
	Fulltext	0.0021^{di}	0.0001	0.0417^{d}	0.0029^{d}	0.0057^{d}	0.0045^{d}	0.0004	0.0492^{d}	0.0086	0.0114	0.0012^{d}	0.0001	0.0185^{d}	0.0000^{d}	0.0000^{d}
	Images	0.0100^{di}	0.0015	0.0549	0.0086	0.0124^{i}	0.0030^{di}	0.0004	0.0250^{di}	0.0029^{i}	0.0029^{di}	-	-	-	-	-
	Captions	0.0026^{di}	0.0002	0.0283^{d}	0.0057^{d}	0.0048^{d}	0.0047^{di}	0.0008	0.0359^{di}	0.0057^{di}	0.0057^{di}	0.0012^{d}	0.0002	0.0352^{d}	0.0000^{d}	0.0010^{d}

Table 2: Effect sizes (Cohen's d_z), standard errors (SE), and 95% confidence intervals (CI) for significant comparisons in Tables 1, 4, and 5. Interpretations of effect sizes follow standard thresholds: small (0.2), medium (0.5), and large (0.8).

		Comparison	Metric	d_z	SE	95% CI	Interpretation
Original Qrels	CLIP CombMNZ	Topic Images vs. Article Images	MAP	-0.18	0.17	[-0.53, 0.16]	Small
	CLIP CombMAX	Topic Description vs. Article Captions	MAP	0.40	0.17	[0.05, 0.76]	Small
			P30	0.60	0.18	[0.23, 0.98]	Medium
		Topic Images vs. Article Title	MAP	0.48	0.18	[0.12, 0.84]	Small
		Topic Images vs. Article Images	MAP	0.38	0.17	[0.03, 0.74]	Small
			P10	0.58	0.18	[0.21, 0.95]	Medium
			P30	0.66	0.19	[0.28, 1.03]	Medium
		Topic Images vs. Article Captions	MAP	0.38	0.17	[0.02, 0.74]	Small
	LongCLIP CombMAX	Topic Description vs. Article Title	MAP	-0.51	0.18	[-0.88, -0.14]	Medium
			P30	-0.51	0.18	[-0.88, -0.15]	Medium
	Llama CombMAX	Topic Description vs. Article Abstract	MAP	0.58	0.18	[0.21, 0.95]	Medium
			P30	0.61	0.18	[0.24, 0.99]	Medium
		Topic Description vs. Article Fulltext	MAP	0.44	0.18	[0.08, 0.80]	Small
			P10	0.56	0.18	[0.19, 0.93]	Medium
			P30	0.52	0.18	[0.16, 0.89]	Medium
Subset	LongCLIP CombMAX	Topic Description vs. Article Title	MAP	-0.46	0.18	[-0.82, -0.10]	Small
Expanded Qrels	CLIP CombMNZ	Topic Images vs. Article Images	MAP	-0.43	0.18	[-0.79, -0.07]	Small
			bpref	-0.35	0.17	[-0.71, -0.00]	Small
	CLIP CombMAX	Topic Description vs. Article Captions	bpref	0.48	0.18	[0.12, 0.84]	Small
		Topic Images vs. Article Title	MAP	0.59	0.18	[0.22, 0.96]	Medium
			bpref	0.53	0.18	[0.16, 0.90]	Medium
		Topic Images vs. Article Abstract	MAP	0.43	0.18	[0.08, 0.79]	Small
			bpref	0.49	0.18	[0.13, 0.86]	Small
		Topic Images vs. Article Captions	MAP	0.50	0.18	[0.14, 0.87]	Medium
			bpref	0.55	0.18	[0.18, 0.92]	Medium
	LongCLIP CombMAX	Topic Images vs. Article Captions	MAP	-0.46	0.18	[-0.82, -0.10]	Small
	PubMedCLIP CombMAX	Topic Description vs. Article Abstract	MAP	0.54	0.18	[0.17, 0.90]	Medium
			bpref	0.50	0.18	[0.14, 0.87]	Medium
	Llama CombMAX	Topic Description vs. Article Abstract	MAP	0.44	0.18	[0.08, 0.80]	Small
	PubMedCLIP CombMAX	Topic Description vs. Article Images	bpref	0.36	0.17	[0.00, 0.71]	Small
			P30	0.30	0.17	[-0.05, 0.65]	Small
		Topic Description vs. Article Captions	P30	0.46	0.18	[0.10, 0.82]	Small

An example of CombMAX's advantage is seen in topic 29, where the query includes two head CT scans, and a relevant article (per qrels) contains one matching CT and one unrelated MRI. CombMAX ranked the article 16th by emphasizing the strongest match, while CombSUM ranked it 124th due to the MRI diluting the overall score. This highlights CombMAX's strength when a single strong match determines relevance.

Overall, **CombMAX emerges as the most effective result fusion method in this context**, consistently outperforming CombSUM and CombMNZ across most metrics.

5.2 Effect of context length

The second experiment investigates the impact of context length by comparing CLIP (77 tokens) with LongCLIP (248 tokens), using CombMAX across all runs. The baseline involves CLIP, and the experiment assesses whether longer context lengths improve results by maintaining the same fusion method for comparison. Results are at the bottom of Table 1.

Results indicate that CLIP generally performs better with short texts like titles and captions, aligning with its original training setup. LongCLIP shows mixed results: it slightly improves effectiveness on longer inputs such as abstracts, but often underperforms on shorter texts, with some statistically significant drops. Visual comparisons show varied results across models, likely due to differences in the text–image alignment learned during training.

In summary, while **context length influences retrieval effectiveness, a longer context does not consistently yield better results**. The benefits of longer context length depend on the type and structure of the input data.

5.3 Domain-specific model effectiveness

The third experiment evaluates whether a domain-specific model, PubMedCLIP, outperforms the general-purpose CLIP model in biomedical retrieval tasks, using CombMAX. As shown in the bottom portion of Table 1, the goal is to determine if fine-tuning a model for a specific domain yields better results than a model pre-trained on diverse data.

While statistical significance is limited, PubMedCLIP generally performs better, particularly in retrieving abstracts and article images. Its improvements suggest stronger domain alignment, especially in extracting abstract-level semantics and visual information. However, inconsistent effectiveness across certain queries, reflected in lower GM-MAP, indicates challenges in handling difficult or ambiguous topics.

In summary, the results show that the **domain-specific model (PubMedCLIP) outperforms the general-purpose model (CLIP) for biomedical searches**, achieving higher effectiveness in retrieving relevant cases.

5.4 Unimodal vs. Multimodal effectiveness

The fourth experiment evaluates whether a unimodal text model, Llama, can outperform a multimodal model, CLIP, using the CombMAX fusion method. Since Llama handles only textual input, it is evaluated solely on text-based searches, while CLIP also includes mixed and visual comparisons (see bottom of Table 1) to assess the effectiveness of unimodal compared to multimodal models.

Llama generally outperforms CLIP across most textual inputs, particularly with longer texts like abstracts and fulltexts, reflecting its stronger language modeling capabilities and training on longer contexts. The observed medium effect sizes suggest that these improvements are not only statistically significant but also practically meaningful. A clear example is topic 29, where Llama ranked a relevant article at the top, while CLIP placed it at 309. The key factor is context length: CLIP's 77-token limit only covered the introduction, missing critical content, whereas Llama, with its 8,192-token capacity, processed the entire article, including two case studies and conclusions. This enabled a deeper semantic understanding, crucial to determining the relevance of the article to this case description.

While CLIP performs slightly better on short texts like titles, Llama surprisingly surpasses it on image captions as well, possibly due to CLIP's limited token capacity when handling complex topic descriptions.

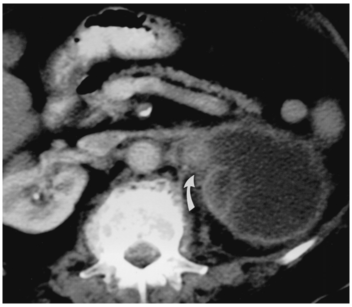
Overall, **Llama consistently outperforms CLIP in nearly all textual sections, as well as in mixed and visual searches**. Llama's superior effectiveness highlights its strength with purely textual content, while CLIP's multimodal capabilities offer no significant advantage in this context. This supports the hypothesis from Section 3.1 that **textual information is prioritized over visual content for this task under evaluation**.

5.5 Dominant data type approach

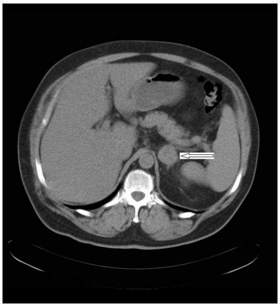
The fifth experiment assesses whether converting visual content into text and prioritizing text can outperform a multimodal approach. This is tested by comparing searches using existing topic sections with those using LLaVa-generated image descriptions (labeled "Gen. I. Desc."). While the focus is on the effectiveness of these generated descriptions, mixed searches that combine them with article images are also analyzed to assess multimodality against unimodality. Instead of relying solely on the text-only Llama model, all models are tested using the CombMAX fusion method. To ensure fair comparisons, each setup is assessed against the corresponding model with statistical relevance tests conducted against the "Topic Description" and "Images" sections. Results show that nearly all metrics are statistically significant for at least one baseline, supporting multiple significant conclusions.

The generation of image descriptions imposed a token limit of 1024 per image to ensure detail and prevent truncation, but actual token counts ranged from 47 to 133, averaging 79 tokens. Descriptions typically began with high-level features, such as color or subfigure count, and then moved on to more specific details within the image. However, the evaluation revealed mixed accuracy due to the model's lack of medical domain fine-tuning, resulting in some errors and inconsistencies. Additionally, a recurring issue was the inconsistency between the topic-generated image descriptions and the image captions, differing in detail, wording, length, and intent. This mismatch, shown by an example in Figure 1, contributed to the observed effectiveness differences.

For CLIP and LongCLIP, visual searches consistently outperformed those using generated descriptions, suggesting that these text surrogates fail to capture the necessary detail or alignment with article content, potentially due to errors in the description generation process and varying levels of detail. PubMedCLIP showed some improvement when comparing generated descriptions to article abstracts, but effectiveness dropped sharply in visual com-



(a) Query image example with LLaVa-generated description: “The image is a black and white medical image of a person’s abdomen, likely an X-ray or CT scan. The abdomen is filled with various organs, including the liver, spleen, and pancreas. The liver is located on the left side of the image, while the spleen is situated in the middle, and the pancreas is on the right side. There is a small arrow pointing towards the right side of the image, possibly indicating a specific area of interest or a point of reference. The overall image provides a detailed view of the internal organs within the abdomen.”.



(b) Article image example with caption: “CT scan showing an adrenal metastasis to the contralateral gland, 2 years after a right nephrectomy for primary RCC.”.

Figure 1: Example of similar abdominal CT scan images, both showing an arrow pointing to a region but with completely different descriptions in terms of wording and medical detail.

parisons. Llama, relying solely on text, also performed worse with generated descriptions than with original topic descriptions, likely due to differences in wording mentioned above.

The experiment assessed whether using text as the dominant data type could outperform a multimodal baseline. While a comparison between the unimodal Llama model and the multimodal CLIP model was obvious, intended to test text-only versus multimodal effectiveness, it was not entirely fair due to differences in models and augmented data. Results showed that Llama with generated image descriptions did not surpass CLIP, though the comparison is inequitable. The experiment suggests that comparing article images with generated captions and topic images could further explore whether visual searches have an advantage over mixed searches, as in the previous experiment the multimodal approach showed no advantage over the unimodal. However, in this experiment, **all textual searches using generated descriptions performed worse than their multimodal baselines, likely due to limitations in the generated data**. Additionally, the misalignment between topic and article captions hindered effectiveness, supporting prior conclusions that topic and article images differ significantly. Despite this effectiveness disparity, CLIP and LongCLIP performed best in visual searches, while PubMedCLIP excelled in comparisons involving topic images and article captions, benefiting from medical domain fine-tuning.

5.6 Comparative analysis against top submissions

To contextualize the effectiveness of our retrieval system, we compare it to the top submissions from the ImageCLEFmed 2013 case-based retrieval task. However, this comparison is not entirely fair due to differences in evaluation methodology. In a typical ImageCLEF

Table 3: Top submissions from ImageCLEFmed 2013 Case-based retrieval task (2013) alongside our highest scores for the same task. Asterisk (*) mark results discussed in Section 6.

		Runid	Retrieval type	MAP	GM-MAP	bpref	P10	P30
Top sub.		SNUMedinfo9 (Choi et al., 2013)	Textual	**0.2429**	**0.1163**	**0.2417**	**0.2657**	**0.1981**
		FCT_CB_MM_rComb (Mourão et al., 2013)	Mixed	**0.1608**	**0.0779**	**0.1426**	**0.1800**	**0.1257**
		FCT_SEGHIST_6x6_LBP (Mourão et al., 2013)	Visual	0.0281	0.0009	0.0335	0.0429	0.0238
Ours	original	Llama CombMAX (desc. & cap.)	Textual	0.0686	0.0123	0.1171	0.1057	0.0667
		PubMedCLIP CombMAX (img. & cap.)	Mixed	0.0224	0.0041	0.0760	0.0457	0.0314
		PubMedCLIP CombMAX (img. & img.)	Visual	**0.0530**	**0.0057**	**0.1011**	**0.0714**	**0.0429**
Ours*	subset	Llama CombMAX (desc. & cap.)	Textual	0.1210	0.0301	0.1171	0.1629	0.1267
		PubMedCLIP CombMAX (img. & abs.)	Mixed	0.0767	0.0144	0.0832	0.1514	0.1229
		PubMedCLIP CombMAX (img. & img.)	Visual	**0.0967**	**0.0134**	**0.1011**	**0.1257**	**0.1105**
Ours*	expanded	Llama CombMAX (desc. & full.)	Textual	0.0419	0.0293	0.0733	0.1657	0.1448
		CLIP CombMAX (img. & title)	Mixed	0.0229	0.0130	0.0425	0.1143	0.0905
		PubMedCLIP CombMAX (img. & img.)	Visual	**0.0282**	**0.0146**	**0.0515**	**0.1257**	**0.0914**

evaluation campaign, the top 30–60 results from each submitted run are merged to create judgment pools of approximately 1,000 cases per topic, which are combined to create pools of approximately 1,000 cases per topic, which are then manually assessed (Kalpathy-Cramer et al., 2015). Since our retrieved documents were not included in this pooling process, only an average of 19.73% of our retrieved articles were judged across all topics, introducing a considerable margin of uncertainty in the effectiveness evaluation.

As shown in Table 3, even our best-performing setup falls significantly behind the top submissions, but we achieved better results in purely visual searches using all the studied multimodal models (CLIP, LongCLIP, PubMedCLIP) compared to top visual submission, which relied on sparse feature extraction methods. This suggests that dense approaches can capture more relevant information. However, in mixed and textual searches, our system significantly underperformed. The top submissions enhanced their text components using an external corpus (MEDLINE) to perform term expansion, contributing to their success. Although we used a somewhat similar approach using the medical-specific model (PubMedCLIP), its inherent token capacity limitations prevented it from handling the large textual sections effectively, likely explaining the discrepancies in retrieval effectiveness.

5.7 Summary

From the five experiments, we can draw several conclusions. The majority of the effect sizes fall within the small and medium ranges. No large effects ($d_z \geq 0.8$) are observed, indicating that while meaningful differences exist, they are generally modest. We begin by analyzing the overall results shared across all experiments and summarizing the key findings.

In the ImageCLEFmed 2013 case-based retrieval task, a total of 709 documents were judged as relevant and 14,319 as non-relevant across 35 query topics, with between 372 and 480 documents judged per topic, covering only about 0.57% of the 75,000 article collection per query topic. This small sample size, due to the pooling technique, may limit the completeness of relevance assessments. Ideally, all documents should be judged to ensure a more accurate evaluation, as unjudged documents are often assumed non-relevant (Clough and Sanderson, 2013), potentially overlooking relevant ones and affecting retrieval effectiveness.

Our overall results are lower than the top submissions of the task. The bpref measure consistently yields better results than MAP across all experiments, indicating that more unjudged documents were retrieved, some of which could be relevant. On average, no more than 19.73% of our retrieved articles were judged across all topics, limiting the evaluation. While bpref accounts for incomplete relevance judgments, it only focuses on the ranking of relevant over non-relevant documents. The highest percentage of judged retrieved articles corresponds with the highest MAP achieved, suggesting that unjudged articles might be relevant. Conversely, the lowest percentage of judged articles resulted in our lowest MAP score (0.0002), yet it did not show the lowest bpref measure. This difference highlights the known property of MAP, which treats all unjudged articles as non-relevant, while bpref handles incomplete judgments by considering only on judged documents.

From all the experiments, we conclude that for the dataset used, CombMAX is the best fusion method out of the ones tested (Exp. 1). Context length affects effectiveness (Exp. 2), showing both advantages and disadvantages based on the input data. Domain-specific models are better suited for their respective domains (Exp. 3). Text-based models can outperform multimodal models when text is the primary information source (Exp. 4). Finally, text searches based on generated descriptions significantly underperform those using the original model on both text and visuals, probably due to limitations in the augmented topic section (Exp. 5).

6 Mitigating incomplete judgments

Many missing judgments in the ground truth may affect reliability of results and conclusions. This can be addressed by adapting the evaluation or expanding the ground truth.

Adapting the evaluation focuses on the subset of the dataset with existing ground truth. This approach, explored in Section 6.1, evaluates system effectiveness within this subset, providing a partial but informative picture. It is a simple, practical method using existing data, but it has limitations. The evaluation assumes that the subset represents the entire collection, which requires careful interpretation as it does not provide a complete evaluation.

Expanding the ground truth can be achieved through manual or semi-supervised techniques. While manual annotation is the most straightforward method, it is often infeasible due to the need for domain experts, as well as its time-intensive and costly nature. In contrast, semi-supervised learning involves using a small labeled dataset to train a model that predicts relevance for unlabeled documents. This method, explored in Section 6.2, can efficiently label large datasets with minimal manual effort, but its accuracy heavily depends on the quality of the initial labeled data and the model's generalization capabilities.

6.1 Retrieval on judged documents

Using trec_eval, we re-ranked our runs and excluded all unjudged documents from the retrieved set, enabling us to calculate metrics solely based on the judged documents, whether relevant or non-relevant. Table 4 presents the overall results for all experiments, considering only the judged documents in the retrieval set. Effect sizes, standard errors, and 95% confidence intervals are also reported for all statistically significant comparisons in Table 2 under "Subset". We analyze these results and compare them with those obtained without excluding unjudged documents, as discussed in Section 5.

Table 4: Summary of experimental results considering only judged documents. Statistically significant scores are marked with an asterisk (*) based on a two-tailed paired permutation test with 100,000 permutations, using Holm-Bonferroni correction at the 0.05 significance level. Significance is computed relative to the baseline (CLIP CombSUM in Experiment 1, and CLIP CombMAX for the remaining). In Experiment 5 (in “Gen. I. Desc.” rows), statistically significant differences relative to the description and image baselines are marked with ‘d’ and ‘i’, respectively. Highest scores per column are bolded.

		CLIP CombSUM$^{\text{Exp. 1}}$					**CLIP CombMNZ**$^{\text{Exp. 1}}$					**CLIP CombMAX**$^{\text{Exp. 1-5}}$				
		MAP	GM-MAP	bpref	P10	P30	MAP	GM-MAP	bpref	P10	P30	MAP	GM-MAP	bpref	P10	P30
Description	Title	**0.1009**	**0.0207**	**0.1028**	**0.1629**	0.1219	-	-	-	-	-	-	-	-	-	-
	Abstract	0.0672	0.0070	0.0799	0.1343	0.1181	-	-	-	-	-	-	-	-	-	-
	Fulltext	0.0968	0.0093	0.1013	0.1371	**0.1305**	-	-	-	-	-	-	-	-	-	-
	Images	0.0509	0.0096	0.0554	0.0943	0.0914	0.0509	**0.0096**	0.0554	0.0943	**0.0914**	0.0565	0.0094	0.0593	0.0886	0.0971
	Captions	0.0592	0.0063	0.0664	0.1029	0.0876	**0.0592**	0.0063	**0.0664**	**0.1029**	0.0876	0.0865	0.0083	0.0941	**0.1371**	**0.1086**
Images	Title	0.0489	0.0092	0.0470	0.0714	0.0733	0.0489	0.0092	0.0470	0.0714	0.0733	0.0662	**0.0128**	0.0597	0.1029	0.0933
	Abstract	0.0389	0.0036	0.0420	0.0686	0.0676	0.0389	0.0036	0.0420	0.0686	0.0676	0.0673	0.0055	0.0725	0.1029	0.0933
	Fulltext	0.0488	0.0053	0.0465	0.0914	0.0895	0.0488	0.0053	0.0465	0.0914	0.0895	0.0717	0.0082	0.0674	0.1057	0.1038
	Images	0.0539	0.0081	0.0546	0.0800	0.0838	0.0465	0.0072	0.0510	0.0743	0.0800	**0.0895**	0.0098	**0.0963**	0.1171	0.1019
	Captions	0.0523	0.0092	0.0480	0.0943	0.0829	0.0523	0.0092	0.0480	0.0943	0.0829	0.0710	0.0124	0.0718	0.0914	0.1029
Gen. I. Desc.	Title	-	-	-	-	-	-	-	-	-	-	0.0328^{d}	0.0011	0.0341^{d}	0.0743^{d}	0.0486^{di}
	Abstract	-	-	-	-	-	-	-	-	-	-	0.0465	0.0016	0.0537	0.0743^{d}	0.0619^{d}
	Fulltext	-	-	-	-	-	-	-	-	-	-	0.0351^{di}	0.0011	0.0435^{d}	0.0743	0.0476^{di}
	Images	-	-	-	-	-	-	-	-	-	-	0.0494^{i}	0.0090	0.0599	0.0857	0.0943
	Captions	-	-	-	-	-	-	-	-	-	-	0.0384^{di}	0.0033	0.0482^{d}	0.0829	0.0781^{d}

		LongCLIP CombMAX$^{\text{Exp. 2}}$					**PubMedCLIP CombMAX**$^{\text{Exp. 3}}$					**Llama CombMAX**$^{\text{Exp. 4}}$				
Description	Title	0.0545*	0.0043	0.0671	0.1400	0.1076	0.1102	0.0127	0.1122	**0.1657**	0.1467	0.0774	0.0225	0.0771	0.1343	0.1152
	Abstract	0.0807	0.0127	0.0944	0.1343	0.1200	**0.1136**	**0.0261**	**0.1180**	0.1571	**0.1562**	0.0971	0.0267	0.0947	0.1571	0.1162
	Fulltext	**0.0855**	0.0062	**0.1046**	**0.1486**	**0.1305**	0.0727	0.0076	0.0866	0.1486	0.1324	0.1200	0.0296	0.1152	**0.1714**	**0.1410**
	Images	0.0697	**0.0138**	0.0748	0.0829	0.0971	0.0578	0.0032	0.0651	0.1057	0.1095	-	-	-	-	-
	Captions	0.0604	0.0119	0.0687	0.1171	0.0981	0.0797	0.0163	0.0869	0.1371	0.1305	**0.1210**	**0.0301**	**0.1171**	0.1629	0.1267
Images	Title	0.0652	0.0120	0.0644	0.1171	0.1019	0.0710	0.0047	0.0705	0.1229	0.1133	-	-	-	-	-
	Abstract	0.0712	0.0135	0.0618	0.1057	0.0943	0.0767	0.0144	0.0832	0.1514	0.1229	-	-	-	-	-
	Fulltext	0.0478	0.0045	0.0548	0.0943	0.0857	0.0579	0.0037	0.0655	0.1229	0.1095	-	-	-	-	-
	Images	0.0853	0.0092	0.0913	0.1143	0.1010	0.0967	0.0134	0.1011	0.1257	0.1105	-	-	-	-	-
	Captions	0.0529	0.0101	0.0560	0.0943	0.0848	0.0755	0.0191	0.0760	0.1257	0.1152	-	-	-	-	-
Gen. I. Desc.	Title	0.0183^{di}	0.0002	0.0245^{d}	0.0486^{d}	0.0210^{di}	0.0447^{d}	0.0030	0.0416^{d}	0.0743^{d}	0.0705^{di}	0.0219^{d}	0.0011	0.0354^{d}	0.0571^{d}	0.0590^{d}
	Abstract	0.0369^{di}	0.0019	0.0432^{d}	0.0686^{d}	0.0667^{d}	0.0665^{d}	0.0080	0.0669^{d}	0.1000	0.0838^{di}	0.0285^{d}	0.0014	0.0337^{d}	0.0514^{d}	0.0533^{d}
	Fulltext	0.0346^{d}	0.0005	0.0417^{d}	0.0714^{d}	0.0476^{d}	0.0370^{d}	0.0025	0.0492^{d}	0.0886	0.0733^{di}	0.0138^{d}	0.0005	0.0185^{d}	0.0286^{d}	0.0438^{d}
	Images	0.0525	0.0067	0.0549	0.0886	0.0943	0.0233^{di}	0.0020	0.0250^{di}	0.0429^{di}	0.0590^{di}	-	-	-	-	-
	Captions	0.0234^{d}	0.0017	0.0283^{d}	0.0543^{d}	0.0543^{d}	0.0321^{di}	0.0053	0.0359^{di}	0.0543^{di}	0.0590^{di}	0.0185^{d}	0.0014	0.0352^{d}	0.0600^{d}	0.0533^{d}

As expected, the bpref values remain unchanged since this metric already excludes unjudged documents, aligning with this new approach. However, the other metrics (MAP, GM-MAP, P10, P30) consistently improve over the previous results. Notably, the MAP reaches 0.1210—nearly twice the highest MAP when unjudged documents are included. Precision also stands out, reaching 0.1714 for P10 and 0.1562 for P30.

Overall, the scores approach those of the best-submitted runs, particularly for mixed retrieval. While our approaches outperform visual retrieval, they remain below the top textual retrieval submission, as seen in the "Ours subset" section of Table 3.

Since we did not have access to the top submitted runs from the ImageCLEF 2013 competition, it was not possible to re-rank or directly evaluate our models against those exact submissions. Instead, by evaluating only on the official judged pool of ImageCLEF 2013, our leave out unjudged results scenario ensures a consistent and fair comparison, as it uses the same set of documents assessed in the original campaign. This approach mitigates biases from unjudged documents and aligns our evaluation methodology with the official judging process as closely as possible, reinforcing the validity of our comparisons.

Importantly, the analysis of retrieval effectiveness considering only judged documents confirms that the conclusions presented in response to the research questions remain consistent and valid. The main trends observed in model effectiveness and fusion strategies hold true under this evaluation strategy, supporting the robustness of our findings.

6.2 Using expanded relevance judgments dataset

We expanded the ImageCLEFmed 2013 case-based retrieval task relevance judgments (qrels) dataset using an MLLM-as-a-Judge approach (Pires et al., 2025), which used Gemini 1.5 Pro to simulate human assessment, increasing the original qrels from 15,028 to 558,653 relevance judgments. Table 5 presents the overall results for all experiments using the expanded dataset, which we analyze and compare to those from the original qrels. We also report effect sizes, standard errors, and 95% confidence intervals for all statistically significant comparisons in Table 2 under "Expanded Qrels".

Overall, the expanded qrels yield 16 statistically significant results, one more than the original qrels. The findings align with previous observations, reinforcing key observations: CombMAX emerges as the most effective result fusion method in this medical context; different context lengths impact effectiveness, but LongCLIP's larger context does not consistently outperform shorter ones, as seen with CLIP; the domain-specific PubMedCLIP surpasses the general-purpose CLIP for biomedical searches; the Llama model consistently outperforms the CLIP model; and all textual searches using generated descriptions underperform compared to their multimodal baselines.

Compared to the results with the original qrels, some MAP values show an increase, but the improvement is not substantial. Notably, Llama continues to achieve the highest MAP when comparing topic descriptions with article captions. However, the top MAP value is still seen with the original qrels. In contrast, GM-MAP, P10, and P30 consistently show clear improvements, while bpref exhibits a substantial decline across all experiments.

Table 5: Summary of experimental results using expanded relevance judgments dataset. Statistically significant scores are marked with an asterisk (*) based on a two-tailed paired permutation test with 100,000 permutations, using Holm-Bonferroni correction at the 0.05 significance level. Significance is computed relative to the baseline (CLIP CombSUM in Experiment 1, and CLIP CombMAX for the remaining). In Experiment 5 (in "Gen. I. Desc." rows), statistically significant differences relative to the description and image baselines are marked with 'd' and 'i', respectively. Highest scores per column are bolded.

		CLIP CombSUM$^{\text{Exp. 1}}$					**CLIP CombMNZ**$^{\text{Exp. 1}}$					**CLIP CombMAX**$^{\text{Exp. 1-5}}$				
		MAP	GM-MAP	bpref	P10	P30	MAP	GM-MAP	bpref	P10	P30	MAP	GM-MAP	bpref	P10	P30
Description	Title	**0.0283**	**0.0188**	**0.0566**	**0.1371**	**0.1152**	-	-	-	-	-	-	-	-	-	-
	Abstract	0.0179	0.0104	0.0416	0.1114	0.0990	-	-	-	-	-	-	-	-	-	-
	Fulltext	0.0249	0.0123	0.0466	0.0943	0.0933	-	-	-	-	-	-	-	-	-	-
	Images	0.0137	0.0080	0.0349	0.0800	0.0733	0.0137	0.0080	0.0349	0.0800	0.0733	0.0165	0.0089	0.0375	0.0714	0.0781
	Captions	0.0185	0.0101	0.0372	0.0914	0.0762	**0.0185**	**0.0101**	**0.0372**	**0.0914**	**0.0762**	0.0257	0.0129	**0.0491***	**0.1171**	**0.0933**
Images	Title	0.0150	0.0077	0.0283	0.0886	0.0705	0.0150	0.0077	0.0283	0.0886	0.0705	**0.0229***	**0.0130**	0.0425*	0.1143	0.0905
	Abstract	0.0098	0.0043	0.0217	0.0571	0.0486	0.0098	0.0043	0.0217	0.0571	0.0486	0.0183*	0.0069	0.0336*	0.0800	0.0752
	Fulltext	0.0100	0.0040	0.0203	0.0457	0.0410	0.0100	0.0040	0.0203	0.0457	0.0410	0.0172	0.0053	0.0304	0.0600	0.0571
	Images	0.0125	0.0070	0.0284	0.0686	0.0610	0.0118*	0.0063	0.0271*	0.0686	0.0610	0.0218	0.0108	0.0417	0.0943	0.0829
	Captions	0.0110	0.0060	0.0233	0.0686	0.0552	0.0110	0.0060	0.0233	0.0686	0.0552	0.0212*	0.0099	0.0396*	0.0800	0.0743
Gen. I. Desc.	Title	-	-	-	-	-	-	-	-	-	-	0.0074^{di}	0.0021	0.0197^{di}	0.0400^{di}	0.0457^{di}
	Abstract	-	-	-	-	-	-	-	-	-	-	0.0074^{di}	0.0012	0.0165^{di}	0.0229^{d}	0.0238^{di}
	Fulltext	-	-	-	-	-	-	-	-	-	-	0.0054^{di}	0.0008	0.0124^{di}	0.0400^{d}	0.0219^{di}
	Images	-	-	-	-	-	-	-	-	-	-	0.0102^{di}	0.0060	0.0250^{di}	0.0400^{i}	0.0410^{di}
	Captions	-	-	-	-	-	-	-	-	-	-	0.0074^{di}	0.0047	0.0203^{di}	0.0314^{d}	0.0410^{di}
		LongCLIP CombMAX$^{\text{Exp. 2}}$					**PubMedCLIP CombMAX**$^{\text{Exp. 3}}$					**Llama CombMAX**$^{\text{Exp. 4}}$				
Description	Title	0.0169	0.0062	0.0382	0.0829	0.0743	**0.0416**	**0.0246**	**0.0687**	**0.1600**	**0.1371**	0.0185	0.0124	0.0423	0.1286	0.0895
	Abstract	0.0203	**0.0120**	0.0421	**0.1200**	0.0867	0.0391*	0.0215	0.0622*	0.1257	0.1181	0.0285*	0.0209	0.0557	0.1514	0.1276
	Fulltext	**0.0255**	0.0111	**0.0475**	0.1200	**0.0990**	0.0255	0.0110	0.0469	0.1200	0.0933	**0.0419**	**0.0293**	**0.0733***	**0.1657**	**0.1448***
	Images	0.0174	0.0082	0.0357	0.0914	0.0800	0.0204	0.0064	0.0467	0.1114	0.0981	-	-	-	-	-
	Captions	0.0189	0.0112	0.0416	0.0886	0.0810	0.0292	0.0161	0.0582	0.1314	0.1229	0.0381	0.0221	0.0694	0.1629	0.1343*
Images	Title	0.0219	0.0098	0.0421	0.1057	0.0838	0.0179	0.0051	0.0294	0.0543	0.0514	-	-	-	-	-
	Abstract	0.0189	0.0068	0.0342	0.0686	0.0695	0.0156	0.0057	0.0275	0.0486	0.0610	-	-	-	-	-
	Fulltext	0.0148	0.0046	0.0303	0.0486	0.0581	0.0129	0.0028	0.0255	0.0629	0.0438	-	-	-	-	-
	Images	0.0204	0.0100	0.0425	0.0714	0.0781	0.0282	0.0146	0.0515	0.1257	0.0914	-	-	-	-	-
	Captions	0.0141*	0.0081	0.0308	0.5710	0.0695	0.0220	0.0102	0.0407	0.1086	0.0819	-	-	-	-	-
Gen. I. Desc.	Title	0.0001^{di}	0.0002	0.0024^{di}	0.0086^{di}	0.0048^{di}	0.0137^{d}	0.0042	0.0296^{d}	0.0714^{d}	0.0638^{d}	0.0021^{d}	0.0007	0.0076^{d}	0.0143^{d}	0.0181^{d}
	Abstract	0.0068^{di}	0.0031	0.0165^{di}	0.0400^{d}	0.0333^{d}	0.0130^{d}	0.0042	0.0253^{d}	0.0714	0.0562^{d}	0.0029^{d}	0.0009	0.0094^{d}	0.0171^{d}	0.0190^{d}
	Fulltext	0.0036^{di}	0.0007	0.0110^{di}	0.0171^{d}	0.0257^{di}	0.0065^{d}	0.0026	0.0188^{d}	0.0400^{d}	0.0343^{d}	0.0025^{d}	0.0005	0.0071^{d}	0.0143^{d}	0.0114^{d}
	Images	0.0116^{di}	0.0058	0.0296^{i}	0.0429^{d}	0.0524^{d}	0.0039^{di}	0.0016	0.0104^{di}	0.0314^{di}	0.0248^{di}	-	-	-	-	-
	Captions	0.0029^{di}	0.0013	0.0116^{di}	0.0371^{d}	0.0286^{di}	0.0068^{di}	0.0037	0.0178^{di}	0.0486^{di}	0.0390^{d}	0.0017^{d}	0.0006	0.0068^{d}	0.0114^{d}	0.0133^{d}

The slight variation in MAP values can be attributed to the nature of the expanded judgments dataset. Despite a significant increase in the number of judgments, approximately 99% were labeled as not relevant. Among all the gathered metrics, MAP and GM-MAP are the most affected by dataset imbalance, as they depend on the ranking of relevant documents across the entire retrieved list. In contrast, precision focuses only on a fixed number of top-ranked results (e.g., top 10 and top 30), making it less sensitive to overall dataset distribution. While bpref is designed to mitigate the impact of missing relevance judgments, it is still influenced by dataset imbalance due to the limited number of relevant documents in the ranked list. Since most of the previously excluded unjudged documents are now considered non-relevant, a decline in metric values was expected.

Compared to the top submission of the ImageCLEFmed 2013 competition, we highlight the precision values listed under the "Ours expanded" section of Table 3, which closely approach the top mixed retrieval results. This indicates that the number of relevant documents retrieved in the top 10 and 30 results is similar to the top submissions.

7 Conclusions

To the best of our knowledge, our paper is the first to apply dense models to multimodal medical case retrieval. Our work investigates the effectiveness of different dense-model approaches in improving multimodal ad hoc search, focusing on retrieving articles relevant to medical differential diagnosis. The findings underscore the limitations of the dataset in a dual-modality search scenario, as incorporating visual data did not enhance retrieval effectiveness. The lack of improvement in retrieval effectiveness was likely due to the physician assessors' preference for textual information. Most documents in the judgment pool were retrieved through textual searches (33 submissions), while far fewer submissions focused on visual (5 submissions) or multimodal (4 submissions) tasks. This aligns with the experimental results, which revealed a clear emphasis on text-based retrieval in the retrieved article. Compounding the issue, no more than about 20% of our retrieved articles across all topics were initially judged, leaving a substantial margin of uncertainty in the evaluation process.

We addressed the challenge of highly incomplete relevance judgments by adapting the evaluation by excluding unjudged documents from our retrieval sets, and using an expanded relevance judgment set that covered all missing judgments across experiments. These strategies led to major improvements, with increases in nearly all metrics, especially precision, bringing scores closer to top submissions. Importantly, the overall findings remain consistent with those obtained using the original qrels, reinforcing their robustness.

To answer **RQ1**: *"Which characteristics of dense models have the greatest impact on retrieval effectiveness in multimodal search systems?"*, the results indicate that various dense model characteristics influence retrieval effectiveness. Context length plays a crucial role, with different lengths offering both advantages and disadvantages depending on the input data. Truncated versions of larger texts were used due to the models' token limits, potentially omitting important information. The Llama 3 model, which has the largest token capacity among the models tested, attained the highest MAP in Experiment 4 (Unimodal vs. Multimodal effectiveness), demonstrating the value of larger context lengths. Additionally, domain-specific models significantly improved retrieval effectiveness over general-purpose models.

To answer **RQ2**: *"How does the effectiveness of dense multimodal models compare to traditional search systems in medical case retrieval, and what factors influence their relative effectiveness?"*, the experiments suggest that dense retrieval holds great potential, particularly for semantic similarity searches across different modalities. However, limitations in context length hinder effectiveness, as multimodal models are often trained on shorter inputs, resulting in lower effectiveness than top submissions. Nonetheless, our approaches excelled in visual retrieval, suggesting that a multimodal large language model, especially if fine-tuned for the medical domain, could greatly enhance effectiveness, though at a high computational cost.

Future work could focus on improving both textual and visual data integration to enhance multimodal medical case retrieval. For textual data, handling large inputs more effectively through text-splitting techniques could enable a more thorough analysis, as Llama, despite its large context length, struggles with longer instances. Splitting fulltexts into logical sections and encoding them separately may improve retrieval effectiveness. Additionally, testing a unified description generator, such as LLaVA or a fine-tuned variant for the medical domain, could help resolve inconsistencies observed in the fifth experiment (Dominant data type approach) due to different generators. On the visual side, exploring medical image modality classification to filter out non-relevant images and compound figure separation to isolate relevant subfigures could reduce retrieval noise. As suggested by Garcia Seco de Herrera et al. (2015), these techniques have the potential to improve case-based retrieval but require further investigation and integration. Finally, integrating the top-performing textual and visual search methods, combining sparse and dense models, could further enhance the accuracy and effectiveness of multimodal search systems. Future studies could also extend experiments to additional datasets, reinforcing the findings and broadening applicability. Additionally, topic-level analysis of system behavior remains important future work to better understand model strengths and limitations across different query topics.

Acknowledgments and Disclosure of Funding

We thank the anonymous reviewers for their valuable comments and suggestions, which helped improve this paper. This work was financed by Component 5 - Capitalization and Business Innovation, integrated into the Resilience Dimension of the Recovery and Resilience Plan, within the scope of the Recovery and Resilience Mechanism (MRR) of the European Union (EU), framed under Next Generation EU, for the period 2021–2026, within the project HfPT (reference 41). Sérgio Nunes acknowledges the use of computational resources provided by the project "PTicola – Increasing Computationally Language Resources for Portuguese" (reference `https://doi.org/10.54499/CPCA-IAC/AV/594794/2023`).

References

Javed A. Aslam and Mark H. Montague. Models for metasearch. In W. Bruce Croft, David J. Harper, Donald H. Kraft, and Justin Zobel, editors, *SIGIR 2001: Proceedings of the 24th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, September 9-13, 2001, New Orleans, Louisiana, USA*, pages

275–284. ACM, 2001. doi: 10.1145/383952.384007.

Khaled Bayoudh, Raja Knani, Fayçal Hamdaoui, and Abdellatif Mtibaa. A survey on deep multimodal learning for computer vision: Advances, trends, applications, and datasets. *Vis. Comput.*, 38(8):2939–2970, 2022. doi: 10.1007/S00371-021-02166-7.

Elia Bruni, Nam-Khanh Tran, and Marco Baroni. Multimodal distributional semantics. *J. Artif. Intell. Res.*, 49:1–47, 2014. doi: 10.1613/JAIR.4135.

Sungbin Choi, Jeongeun Lee, and Jinwook Choi. SNUMedinfo at ImageCLEF 2013: Medical retrieval task. In Pamela Forner, Roberto Navigli, Dan Tufis, and Nicola Ferro, editors, *Working Notes for CLEF 2013 Conference , Valencia, Spain, September 23-26, 2013*, volume 1179 of *CEUR Workshop Proceedings*. CEUR-WS.org, 2013. URL `https://ceur-ws.org/Vol-1179/CLEF2013wn-ImageCLEF-ChoiEt2013.pdf`.

Paul D. Clough and Mark Sanderson. Evaluating the performance of information retrieval systems using test collections. *Inf. Res.*, 18(2), 2013. URL `http://www.informationr.net/ir/18-2/paper582.html`.

Gordon V. Cormack, Charles L. A. Clarke, and Stefan Büttcher. Reciprocal rank fusion outperforms condorcet and individual rank learning methods. In James Allan, Javed A. Aslam, Mark Sanderson, ChengXiang Zhai, and Justin Zobel, editors, *Proceedings of the 32nd Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, SIGIR 2009, Boston, MA, USA, July 19-23, 2009*, pages 758–759. ACM, 2009. doi: 10.1145/1571941.1572114.

Can Cui, Haichun Yang, Yaohong Wang, Shilin Zhao, Zuhayr Asad, Lori A. Coburn, Keith T. Wilson, Bennett A. Landman, and Yuankai Huo. Deep multimodal fusion of image and non-image data in disease diagnosis and prognosis: A review. *Progress in Biomedical Engineering*, 5(2):022001, apr 2023. doi: 10.1088/2516-1091/acc2fe.

Sedigheh Eslami, Gerard de Melo, and Christoph Meinel. Does CLIP benefit visual question answering in the medical domain as much as it does in the general domain? *CoRR*, abs/2112.13906, 2021. URL `https://arxiv.org/abs/2112.13906`.

Manuel Faysse, Hugues Sibille, Tony Wu, Bilel Omrani, Gautier Viaud, Céline Hudelot, and Pierre Colombo. ColPali: Efficient document retrieval with vision language models. In *The Thirteenth International Conference on Learning Representations, ICLR 2025, Singapore, April 24-28, 2025*. OpenReview.net, 2025. URL `https://openreview.net/forum?id=ogjBpZ8uSi`.

Di Feng, Christian Haase-Schütz, Lars Rosenbaum, Heinz Hertlein, Claudius Gläser, Fabian Timm, Werner Wiesbeck, and Klaus Dietmayer. Deep multi-modal object detection and semantic segmentation for autonomous driving: Datasets, methods, and challenges. *IEEE Trans. Intell. Transp. Syst.*, 22(3):1341–1360, 2021. doi: 10.1109/TITS.2020.2972974.

Alba Garcia Seco de Herrera, Jayashree Kalpathy-Cramer, Dina Demner-Fushman, Sameer K. Antani, and Henning Müller. Overview of the ImageCLEF 2013 medical tasks. In Pamela Forner, Roberto Navigli, Dan Tufis, and Nicola Ferro, editors, *Working Notes for CLEF 2013 Conference , Valencia, Spain, September 23-26, 2013*, volume 1179 of *CEUR Workshop Proceedings*. CEUR-WS.org, 2013. URL `https://ceur-ws.org/Vol-1179/CLEF2013wn-ImageCLEF-SecoDeHerreraEt2013b.pdf`.

Alba Garcia Seco de Herrera, Roger Schaer, Dimitrios Markonis, and Henning Müller. Comparing fusion techniques for the ImageCLEF 2013 medical case retrieval task. *Comput. Medical Imaging Graph.*, 39:46–54, 2015. doi: 10.1016/J.COMPMEDIMAG.2014.04.004.

Alba Garcia Seco de Herrera, Roger Schaer, and Henning Müller. Shangri-la: A medical case-based retrieval tool. *J. Assoc. Inf. Sci. Technol.*, 68(11):2587–2601, 2017. doi: 10.1002/ASI.23858.

Aaron Grattafiori, Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Alex Vaughan, et al. The Llama 3 herd of models. *CoRR*, abs/2407.21783, 2024. doi: 10.48550/ARXIV.2407.21783.

MedGIFT Group. medSearch – medical search engine by HES-SO Valais (medSearch 2009). `http://fast.hevs.ch:8080/MedSearch/faces/Search.jsp`, 2009.

Ruifeng Guo, Jingxuan Wei, Linzhuang Sun, Bihui Yu, Guiyong Chang, Dawei Liu, Sibo Zhang, Zhengbing Yao, Mingjun Xu, and Liping Bu. A survey on advancements in image-text multimodal models: From general techniques to biomedical implementations. *Comput. Biol. Medicine*, 178:108709, 2024. doi: 10.1016/J.COMPBIOMED.2024.108709.

D. Frank Hsu and Isak Taksa. Comparing rank and score combination methods for data fusion in information retrieval. *Inf. Retr.*, 8(3):449–480, 2005. doi: 10.1007/S10791-005-6994-4.

Morris A. Jette and Tim Wickberg. Architecture of the slurm workload manager. In Dalibor Klusáček, Julita Corbalán, and Gonzalo P. Rodrigo, editors, *Job Scheduling Strategies for Parallel Processing - 26th Workshop, JSSPP 2023, St. Petersburg, FL, USA, May 19, 2023, Revised Selected Papers*, volume 14283 of *Lecture Notes in Computer Science*, pages 3–23. Springer, 2023. doi: 10.1007/978-3-031-43943-8_1.

Qiao Jin, Robert Leaman, and Zhiyong Lu. PubMed and beyond: Biomedical literature search in the age of artificial intelligence. *eBioMedicine*, 100:104988, 2024. ISSN 2352-3964. doi: https://doi.org/10.1016/j.ebiom.2024.104988.

Jeff Johnson, Matthijs Douze, and Hervé Jégou. Billion-scale similarity search with GPUs. *IEEE Trans. Big Data*, 7(3):535–547, 2021. doi: 10.1109/TBDATA.2019.2921572.

Jayashree Kalpathy-Cramer, Alba García Seco de Herrera, Dina Demner-Fushman, Sameer Antani, Steven Bedrick, and Henning Müller. Evaluating performance of biomedical image retrieval systems—an overview of the medical image retrieval task at ImageCLEF 2004–2013. *Computerized Medical Imaging and Graphics*, 39:55–61, 2015. ISSN 0895-6111. doi: https://doi.org/10.1016/j.compmedimag.2014.03.004.

Yihao Li, Mostafa El Habib Daho, Pierre-Henri Conze, Rachid Zeghlache, Hugo Le Boité, Ramin Tadayoni, Béatrice Cochener, Mathieu Lamard, and Gwenolé Quellec. A review of deep learning-based information fusion techniques for multimodal medical image classification. *Comput. Biol. Medicine*, 177:108635, 2024. doi: 10.1016/J.COMPBIOMED.2024.108635.

Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. Visual instruction tuning. In Alice Oh, Tristan Naumann, Amir Globerson, Kate Saenko, Moritz Hardt, and Sergey Levine, editors, *Advances in Neural Information Processing Systems 36: Annual Conference on Neural Information Processing Systems 2023, NeurIPS 2023, New Orleans, LA, USA, December 10 - 16, 2023*, 2023. URL `http://papers.nips.cc/paper_files/paper/2023/hash/6dcf277ea32ce3288914faf369fe6de0-Abstract-Conference.html`.

Zhiyong Lu, Won Kim, and W. John Wilbur. Evaluation of query expansion using MeSH in PubMed. *Inf. Retr.*, 12(1):69–80, 2009. doi: 10.1007/S10791-008-9074-8.

Mark H. Montague and Javed A. Aslam. Condorcet fusion for improved retrieval. In *Proceedings of the 2002 ACM CIKM International Conference on Information and Knowledge Management, McLean, VA, USA, November 4-9, 2002*, pages 538–548. ACM, 2002. doi: 10.1145/584792.584881.

André Mourão and Flávio Martins. NovaMedSearch: A multimodal search engine for medical case-based retrieval. In João Ferreira, João Magalhães, and Pável Calado, editors, *Open research Areas in Information Retrieval, OAIR 2013, Lisbon, Portugal, May 15-17, 2013*, pages 223–224. ACM, 2013. URL `http://dl.acm.org/citation.cfm?id=2491798`.

André Mourão, Flávio Martins, and João Magalhães. NovaSearch on medical ImageCLEF 2013. In Pamela Forner, Roberto Navigli, Dan Tufis, and Nicola Ferro, editors, *Working Notes for CLEF 2013 Conference , Valencia, Spain, September 23-26, 2013*, volume 1179 of *CEUR Workshop Proceedings*. CEUR-WS.org, 2013. URL `https://ceur-ws.org/Vol-1179/CLEF2013wn-ImageCLEF-MouraoEt2013.pdf`.

Henning Müller and Jayashree Kalpathy-Cramer. The ImageCLEF medical retrieval task at ICPR 2010 - information fusion. In *20th International Conference on Pattern Recognition, ICPR 2010, Istanbul, Turkey, 23-26 August 2010*, pages 3284–3287. IEEE Computer Society, 2010. doi: 10.1109/ICPR.2010.803.

Catarina Pires, Sérgio Nunes, and Luís Filipe Teixeira. Expanding relevance judgments for medical case-based retrieval task with multimodal LLMs. *CoRR*, abs/2506.17782, 2025. doi: 10.48550/ARXIV.2506.17782. Presented at the Third Workshop on Large Language Models for Evaluation in Information Retrieval (LLM4Eval 2025), co-located with SIGIR 2025, Padua, Italy, July 17, 2025.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. Learning transferable visual models from natural language supervision. In Marina Meila and Tong Zhang, editors, *Proceedings of the 38th International Conference on Machine Learning, ICML 2021, 18-24 July 2021, Virtual Event*, volume

139 of *Proceedings of Machine Learning Research*, pages 8748–8763. PMLR, 2021. URL http://proceedings.mlr.press/v139/radford21a.html.

Joseph A. Shaw and Edward A. Fox. Combination of multiple searches. In Donna K. Harman, editor, *Proceedings of The Third Text REtrieval Conference, TREC 1994, Gaithersburg, Maryland, USA, November 2-4, 1994*, volume 500-225 of *NIST Special Publication*, pages 105–108. National Institute of Standards and Technology (NIST), 1994. URL http://trec.nist.gov/pubs/trec3/papers/vt.ps.gz.

Sonish Sivarajkumar, Haneef Ahamed Mohammad, David Oniani, Kirk Roberts, William R. Hersh, Hongfang Liu, Daqing He, Shyam Visweswaran, and Yanshan Wang. Clinical information retrieval: A literature review. *J. Heal. Informatics Res.*, 8(2):313–352, 2024. doi: 10.1007/S41666-024-00159-4.

Konstantinos Zagoris, Savvas A. Chatzichristofis, Nikos Papamarkos, and Yiannis S. Boutalis. img(anaktisi): A web content based image retrieval system. In Tomás Skopal and Pavel Zezula, editors, *Second International Workshop on Similarity Search and Applications, SISAP 2009, 29-30 August 2009, Prague, Czech Republic*, pages 154–155. IEEE Computer Society, 2009. doi: 10.1109/SISAP.2009.15.

Beichen Zhang, Pan Zhang, Xiaoyi Dong, Yuhang Zang, and Jiaqi Wang. Long-CLIP: Unlocking the long-text capability of CLIP. In Ales Leonardis, Elisa Ricci, Stefan Roth, Olga Russakovsky, Torsten Sattler, and Gül Varol, editors, *Computer Vision - ECCV 2024 - 18th European Conference, Milan, Italy, September 29-October 4, 2024, Proceedings, Part LI*, volume 15109 of *Lecture Notes in Computer Science*, pages 310–325. Springer, 2024. doi: 10.1007/978-3-031-72983-6_18.

Yifei Zhang, Désiré Sidibé, Olivier Morel, and Fabrice Mériaudeau. Deep multimodal fusion for semantic image segmentation: A survey. *Image Vis. Comput.*, 105:104042, 2021. doi: 10.1016/J.IMAVIS.2020.104042.

Submitted 02/25; Revised 09/25; Published 10/25

Effectiveness of In-Context Learning for Due Diligence
A Reproducibility Study of Identifying Passages for Due Diligence

Madhukar Dwivedi MDWIVEDIUVA@GMAIL.COM
Institute for Logic, Language and Computation
University of Amsterdam, The Netherlands

Jaap Kamps KAMPS@UVA.NL
Institute for Logic, Language and Computation
University of Amsterdam, The Netherlands

Editor: Daniela Godoy

Abstract

In recent years, Information Retrieval (IR) has evolved from ad hoc document retrieval to passage and answer retrieval, incorporating downstream Natural Language Processing (NLP). This led to remarkable progress in models when evaluated on early precision, yet at the same time, the potential to improve recall aspects has received less attention. This paper investigates an extremely high-recall task by a reproducibility study on a massive collection of merger and acquisition documents in due diligence passage retrieval. We have replicated previous work using Conditional Random Fields (CRF) and introduced a Python version of the effective CRFsuite approach. In addition, we explore the utility of open-source and closed-source Large Language Models (LLMs) with zero-shot and few-shot learning techniques on 50 different due diligence topics. Our findings reveal the potential for few-shot learning in due diligence, delivering acceptable levels of performance in terms of recall, marking an essential step towards developing advanced due diligence models that minimize the dependency on extensive training data typically required by domain-specific IR and NLP models. More generally, our results are an important first step toward developing advanced due diligence models for any legal information need.

Keywords: Information retrieval, Legal search, Due diligence passage retrieval

1 Introduction

The general public often thinks that "Search" is a solved problem, as modern Information Retrieval (IR) and Natural Language Processing (NLP) models exhibit high precision: the top retrieved results will likely be relevant to our user. While precision is of obvious importance, high recall is of equal or even greater relevance in many professional and domain-specific IR applications. These use cases necessitate high recall because the critical implications of overlooking key information present a significant challenge. They also present needle-in-a-haystack problems that are very challenging for NLP and text classification, as the labels are extremely unevenly distributed: collections are large, and a tiny fraction contains relevant information.

These aspects are particularly critical in specialized applications such as legal due diligence for mergers and acquisitions. Due diligence refers to a systematic legal process where

WWW: https://irrj.org DOI: 10.54195/irrj.22626

No.	Text	Pred.	Ref.
1	(a) the arithmetic mean of the rates (rounded upwards to four decimal places) as supplied to the Agent at its request by the Reference Banks # C #	B	1
2	5.15 cm	B	1
3	the Reference Bank is a contributor to the applicable Screen Rate	1	1
4	“Revolving Facility Loan“ means a loan made or to be made under the Revolving Facility or the principal amount outstanding for the time being of that loan	B	B
5	“ Rollover Loan “ means one or more Revolving Facility Loans # C #	B	B
6	and]	B	1

Table 1: Text classification using CRFsuite PA on sentence-level data: 1 denotes relevant, B denotes non-relevant, using the KIRA due diligence data (Roegiest, Hudek, and McNulty, 2018)

a potential buyer evaluates a company’s assets, liabilities, contractual obligations, and associated risks before completing a transaction. The goal is to uncover hidden risks and ensure informed decisions, making it essential to retrieve all potentially relevant information from complex legal documents. In this context, the manual extraction of key information from extensive legal documents is labor-intensive and prone to significant financial risks (Klaber, 2013; Sherer, Hoffman, and Ortiz, 2015; Sherer, Hoffman, Wallace, et al., 2016). This form of legal document retrieval extends beyond mere document retrieval, demanding a nuanced understanding of complex legal terminologies and stipulations to ensure that no potential liabilities are overlooked. This leads to relatively generic topics and a formidable challenge in sentence or passage retrieval, with extremely high recall requirements to find relevant passages. Table 1 shows an example where a retrieved sentence is compared with the annotated relevant sentence for the topic ‘*change of control definition credit agreement.*’ This example highlights both the difficulty of matching semantically relevant content and the noise introduced by text extraction processes (e.g., tokenization errors, OCR noise). As a case in point, the references include `5.15 cm` as a relevant sentence, as it was included in a larger passage annotated as relevant to the topic. Such challenges, common in large-scale document analysis, directly affect the reliability of automated due diligence systems.

There is great variability in contractual agreements, and the specific accuracy required for mergers and acquisitions due diligence underscores the need for models capable of adapting to a wide array of legal contexts and jurisdictions. There is great interest in developing advanced IR and NLP approaches for automating this intricate process, which led to an emerging literature on finding essential risk passages in large corpora of legal documents (Parikh, Poojary, and Gupta, 2023; Moriarty et al., 2019; Roegiest, Hudek, and McNulty, 2018). In addressing these challenges, the study by Roegiest, Hudek, and McNulty (2018) entitled “*A Dataset and an Examination of Identifying Passages for Due Diligence*” is of particular interest. This research highlights the importance of framing legal due diligence as an IR task, particularly in dealing with issues such as high recall and uneven data distribution, which are critical to preventing oversights in sensitive legal environments. The study's contributions include developing a specialized dataset and evaluation data of due diligence passage retrieval. This paper aims to perform a reproducibility study of Roegiest, Hudek, and McNulty (2018) and present an extended analysis that enables future research on automated high-recall information retrieval and passage extraction in legal document analysis.

Motivation and Objectives of our Study We structure our work in two phases. First, we rigorously reproduce and validate the original study by Roegiest, Hudek, and McNulty (2018), ensuring the robustness of their methods. Second, we extend the analysis by exploring whether recent advances in Large Language Models (LLMs) can offer practical alternatives in high-recall, low-annotation settings common in due diligence tasks.

Our study focuses on four main objectives:

Reproducibility We aim to reproduce the original experiments using the same data and analytical tools; in addition to this, we specifically adapt the CRFsuite (PA) algorithm implementation in Python to confirm cross-framework compatibility.

Robustness We compare standard text processing to the original paper's tailored preprocessing and feature engineering techniques like using a customized 'punkt' trained on 1M EDGAR documents, n-gram inclusion, case normalization, and POS tagging.

Large Language Models Analysis We explore the effectiveness of open-source LLMs for the due diligence problem, and compare to proprietary LLMs in zero-shot and few-shot scenarios using prompt-based methods.

Resources We offer an implementation of the CRFSuite and other models, including modern LLMs, within a Python environment. Moreover, we create a subset of the original data to analyse the effectiveness of of LLMs for the due diligence problem.

The rest of this paper is structured as follows. Next, §2 details the used due diligence dataset. §3 summarizes the original paper and experimental setup. §4 details our reproduction and replication of the main results. §5 details an LLM approach to due diligence. Finally, §6 discusses our findings and draws conclusions.

2 Kira Data and Evaluation Subset for LLMs

This section discusses the used due diligence dataset, framing due diligence as a high recall information retrieval problem.

Kira Dataset Roegiest, Hudek, and McNulty (2018) developed the Kira Systems collection to support academic research by identifying key information within legal documents. This dataset comprises 4,412 legal documents, primarily related to credit agreements pertinent to mergers and acquisitions, annotated across 50 topics that reflect due diligence needs. These documents contain over 15 million sentences, with a significant portion remaining unannotated, posing a challenge in pinpointing relevant information within such a voluminous dataset. Despite the high prevalence at the document level, with two-thirds of the documents containing relevant examples, the prevalence of specifically annotated sentences is notably low, ranging from 0.01% to 0.7% per topic. This indicates a much finer granularity in targeting sentence-level relevance over document-level, underscoring the complexity of accurately identifying crucial information.

Legal professionals, including law students and experienced lawyers, meticulously annotated each document to ensure the accuracy and legal validity of the annotations. This rigorous process ensures that the annotations not only reflect genuine legal analysis but are also precisely aligned with the intricacies and requirements of due diligence. For instance, an example from the dataset illustrates the dataset's utility in extracting key information for due diligence, highlighting conditions that affect transaction risk profiles:

> *In the event of a Change of Control, the Borrower must provide written notice to the Lender within 30 days, triggering a reassessment of the loan terms.*

The dataset comprises real-life legal contracts spanning 50 topics, each accompanied by expert-generated titles and descriptions. We used these titles and descriptions to create prompts for LLMs, guiding them to generate appropriate responses for the complex due diligence task. These clear and detailed descriptions provide the nuanced context required for accurate model predictions. For example, 'Evidence of Loans' - one of the 50 annotated topics- is described as:

> *To avoid any future debate as to how much the borrower owes, this topic captures provisions that typically set out that a lender's internal records or accounts are conclusive evidence of the amount owed to the lender by the borrower and may further be evidenced by a promissory note by the borrower to the lender.*

LLM Evaluation Subset To evaluate the performance of Large Language Models (LLMs), we created an evaluation subset from the Kira dataset, selecting all relevant sentences along with a random subset of 1,000 non-relevant sentences per topic due to computational and time constraints. These were not passages in the traditional sense, each selected sentence was required to be at least 240 characters long to reduce the risk of truncation during LLM inference. We chose this threshold to filter out fragmented or broken sentences, such as artifacts, such as ##C##, that often occur during text segmentation. Additionally, sentences shorter than 240 characters may lack sufficient context for LLMs, increasing the risk of unpredictable outputs in prompt-based inference. Running inference over the entire Kira dataset with LLMs would have been prohibitively expensive and time-consuming, which is why we opted for a smaller evaluation subset.

This smaller subset was designed to maintain enough challenge for a fair evaluation while making the LLM experiments computationally feasible. Although this setup reduces class imbalance compared to the original dataset, it still simulates the low-prevalence condition typically encountered in real-world due diligence tasks. Each topic includes a fixed number of 1,000 non-relevant sentences to allow uniform testing across topics, while the number of relevant ones varies considerably.

Our evaluation subset is significantly smaller than the original Kira dataset, as we retain only a few thousand of the sentences per topic from the original 15 million. The subset remains imbalanced, with a lower fraction of relevant sentences compared to non-relevant ones. This imbalance, though less severe than in the full data, continues to influence model behavior. Specifically, the non-relevant category is fixed at 1,000 sentences per topic, while the relevant category ranges from 15 to 1,307 (median 124, average 210).

This subset provides a manageable yet meaningful benchmark for exploring the utility of current LLMs in Due Diligence. While we don't claim equivalence to the full Kira dataset, this design enables fair comparative evaluation in a practical setting, without requiring large-scale inference runs.

Although the reduced class imbalance might affect absolute performance metrics (e.g. precision), our primary goal is to compare models under consistent conditions. Since all LLMs and baselines are evaluated on the same balanced subset, the relative rankings and comparative insights remain valid.

3 Overview of Original Study

This section summarizes the original paper and experimental setup, exploiting large-scale train data for traditional machine learning models.

The original study (Roegiest, Hudek, and McNulty, 2018) proposes the use of advanced information retrieval techniques to automate the due diligence process in mergers and acquisitions, aiming to replace the traditional, labor-intensive scrutiny of legal documents. The research focuses on developing a reliable tool capable of pinpointing key passages within extensive legal texts—a crucial task given the high recall needs and the rarity of relevant information. Utilizing machine learning, specifically Conditional Random Fields, this approach enables precise detection and extraction of critical data points that indicate potential risks in mergers and acquisitions transactions. This work not only enhances the precision and efficiency of legal evaluations but also significantly contributes to the field by advancing the application of machine learning in complex legal scenarios. Below is a detailed description of the methodology, evaluation measures, and key takeaways from the study.

Methodology The original study employs multiple models for sentence-level classification, each designed to capture different aspects of the due diligence task. Conditional Random Fields (CRF) are used via CRFsuite, treating each sentence as an independent instance. The term ”entity” in this context refers to relevant sentence-level segments, not named entities. Features used capture lexical, structural, and topic-related characteristics.

CRFsuite is trained with both Passive-Aggressive (PA) (Crammer et al., 2006) and LBFGS (Nocedal, 1980) optimizers. PA is particularly suited for high-recall tasks like legal due diligence, as it updates only on errors, promoting more inclusive classification. This mirrors the configuration used in the original study (Roegiest, Hudek, and McNulty, 2018). For PA, we used a moderate aggressiveness setting ($c = 0.1$), the second variant of the PA algorithm ($type = 2$), and capped the maximum number of training iterations at 100. For the LBFGS optimizer, we similarly limited training to 100 iterations to maintain consistency in convergence time.

In addition to CRFsuite, the original setup also included SVMhmm, which applies Support Vector Machines for sequence modeling of sentences, offering improved non-linear discrimination over traditional HMMs.

Separately, Vowpal Wabbit (Langford, Li, and Strehl, 2025) is used to train a logistic regression classifier on the same sentence-level features. Configurations include `--holdout_off`

`--loss_function logistic --passes 50`, and when bigram features are used: `--ngram 2 -b 24`.

We retain all three models—CRFsuite, SVMhmm, and logistic regression—directly from the original study to ensure reproducibility and comparative consistency across methods.

Evaluation Measures The original paper evaluated performance using two distinct metrics to measure the effectiveness of sentence and passage classification. We follow the same setup and compute precision, recall, and F1 scores for the relevant class only, as the task focuses on retrieving legally important content while ignoring non-relevant text. Macro-averaging across both classes is not meaningful in this context, where recall of relevant content is critical. First, in *sentence-level* evaluation, precision, recall, and F1 scores are calculated by treating each sentence as a separate data point. This directly reflects how well the model isolates relevant due diligence material at the sentence level. Second, *annotation-level* evaluation assesses a model's ability to label text sequences accurately by treating groups of sentences as unified entities (a paragraph), unlike sentence-level evaluation's focus on individual sentences. For example, imagine a legal document where sentences 5 through 10 pertain to a specific legal issue important for a due diligence task.

Suppose a model correctly classifies some of these six sentences as relevant. In that case, our user will still have discovered the relevant due diligence information, and we can regard this as a success. Hence, annotation-level evaluation is a more lenient measure that can be interpreted directly by our legal user, having located all the relevant information flagged for further inspection.

4 Reproduction and Replication

This section details our reproduction and replication of the main results. The first part discusses reproducing the original experiment results and implementing the CRFsuite algorithm in Python. The second part focuses on replicating the CRFsuite algorithm with simpler features and evaluating its performance.

4.1 Reproduction

We first reproduced the experiments from the original study using the same code, feature data, parameters, and algorithm versions on the same platform. According to the ACM Artifact Review and Badging Guidelines (v1.1), this qualifies as a **reproduction**: a different research team repeating the original experimental setup to verify published results.

The original study by Roegiest, Hudek, and McNulty (2018) used CRFsuite for sentence-level classification with feature vectors derived from a custom preprocessing pipeline. While the source code for this pipeline was not released, the authors shared the resulting feature representations, enabling us to replicate their CRF results exactly. For a detailed breakdown of this preprocessing pipeline and feature engineering, see the next section.

4.1.1 Feature Engineering from Original Paper

The original study applied custom preprocessing to address challenges specific to legal texts. A key step involved adapting the `punkt` sentence segmentation algorithm for legal documents. Legal filings from the EDGAR repository often include non-standard punctuation,

Source	Model	Precision	Recall	F1-Score
(a) Original	CRF-PA	0.92 [0,91,0.93]	**0.85** [0.83,0.88]	**0.88** [0.87,0.90]
	CRF-LBFGS	**0.94** [0.93,0.95]	0.80 [0.77,0.83]	0.86 [0.84,0.88]
	SVM-HMM	0.93 [0.89,0.96]	0.69 [0.64,0.74]	0.78 [0.74,0.82]
	VW-Tuned	0.92 [0.91,0.94]	0.62 [0.58,0.65]	0.74 [0.71,0.76]
	VW-Sent	0.90 [0.89,0.92]	0.65 [0.62,0.68]	0.75 [0.72,0.78]
(b) Replication	CRF-PA	0.9235 [0.91,0.93]	**0.8473** [0.83,0.88]	**0.8812** [0.87,0.90]
	CRF-LBFGS	**0.9440** [0.93, 0.95]	0.8091 [0.77,0.83]	0.8691 [0.84,0.88]
	SVM-HMM	0.9273 [0.89, 0.96]	0.6881 [0.64,0.74]	0.7755 [0.74,0.82]
	VW-Tuned	0.9240 [0.91, 0.94]	0.6225 [0.58,0.65]	0.7396 [0.71,0.76]
	VW-Sent	0.8993 [0.89, 0.92]	0.6460 [0.62,0.68]	0.7487 [0.72,0.78]
(c) Python	CRF-PA	0.9211 [0.91,0.93]	0.8509 [0.83,0.88]	0.8826 [0.87,0.90]
(d) Text Features	CRF-PA	0.9405 [0.92, 0.95]	0.7214 [0.68, 0.76]	0.8089 [0.78, 0.84]

Table 2: Sentence level evaluation: Comparison between (a) the original results; (b) replication results; (c) Python re-implementation; and (d) CRF replication with text features; Square brackets indicate the 95% confidence intervals.

enumerated clauses (e.g., "Section 5(b)"), and inconsistent formatting, which cause off-the-shelf sentence splitters to perform poorly. The customized punkt variant was developed to handle such structures, ensuring accurate sentence boundaries—a critical factor for reliable sentence-level classification.

Beyond segmentation, the authors engineered lexical and semantic features tailored to this domain. These included bigram and trigram token features derived from word2vec embeddings trained on over 1 million EDGAR documents. Token vectors were clustered using k-means, and cluster IDs were used as features to capture broader semantic patterns. This approach was particularly beneficial in the low-resource, high-recall setting of due diligence.

Feature generation and classification leveraged the Vowpal Wabbit toolkit, combining both in-house feature sets and VW's n-gram hashing capabilities. This hybrid setup enabled effective learning of both domain-specific and generalizable patterns, providing robust input to CRF, SVMhmm, and logistic regression models.

Our study successfully reproduced the original study's work, focusing on models like Conditional Random Fields (CRFs) via CRFsuite (PA, LBFGS), SVM-HMM, and two Vowpal Wabbit (VW) configurations: VW Tuned and VW Sent. These models were key for analyzing legal documents in mergers and acquisitions due diligence in the original study.

Our findings affirm the original study's reliability and impact, endorsing the methods proposed by Roegiest, Hudek, and McNulty (2018). The detailed metrics such as precision, recall, and F1 scores, along with their 95% confidence intervals shown in square brackets, are provided in Table 2(a,b) and Table 3(a,b) respectively. The close match of our results with the original study, consistently within two decimal points, confirms the success of our reproduction efforts.

Source	Model	Precision	Recall	F1-Score
(a) Original	CRF-PA	0.92 [0.90,0.93]	**0.94** [0.94,0.95]	**0.93** [0.92,0.94]
	CRF-LBFGS	**0.97** [0.96,0.98]	0.85 [0.83,0.88]	0.90 [0.89,0.92]
	SVM-HMM	0.92 [0.88,0.95]	0.84 [0.81,0.88]	0.88 [0.84,0.91]
	VW-Tuned	0.84 [0.80,0.87]	0.83 [0.81,0.86]	0.83 [0.80,0.85]
	VW-Sent	0.79 [0.75,0.83]	0.88 [0.86,0.90]	0.82 [0.79,0.85]
(b) Replication	CRF-PA	0.9189 [0.90,0.93]	**0.9436** [0.94,0.95]	**0.9300** [0.92,0.94]
	CRF-LBFGS	**0.9720** [0.96,0.98]	0.8540 [0.83,0.88]	0.9033 [0.89,0.92]
	SVM-HMM	0.9160 [0.88,0.95]	0.8426 [0.81,0.88]	0.8752 [0.84,0.91]
	VW-Tuned	0.8377 [0.80,0.87]	0.8326 [0.81,0.86]	0.8265 [0.80,0.85]
	VW-Sent	0.7885 [0.75,0.83]	0.8811 [0.86,0.90]	0.8224 [0.79,0.85]
(c) Python	CRF-PA	0.9190 [0.90,0.93]	0.9428 [0.94,0.95]	0.9298 [0.92,0.94]
(d) Text Features	CRF-PA	0.9272 [0.91, 0.94]	0.8624 [0.84, 0.88]	0.8893 [0.87, 0.90]

Table 3: Annotation level evaluation: Comparison between (a) the original results; (b) replication results; (c) Python re-implementation; and (d) CRF replication with text features; Square brackets indicate the 95% confidence intervals.

4.1.2 Python code for CRFsuite

We also reproduced the CRFsuite experiments within a Python environment using the `sklearn-crfsuite` package. This Python-based setup served two purposes: first, to confirm that CRFsuite's performance remains consistent when integrated with standard Python NLP workflows; second, to make the reproduction accessible to a broader community familiar with Python tools. Importantly, we did not modify or re-implement the CRFsuite algorithm, our work simply wraps the existing CRFsuite functionality using Python bindings.

Our findings, demonstrating consistent model performance, are detailed in Table 2(c) and Table 3(c) for sentence level and annotation level, respectively.

4.2 Replication

We extend our analysis beyond just reproduction. The original paper uses proprietary features, and we try to reproduce these from the source text. We carefully align our text features with those used in the original paper, focusing on the best-performing CRF model. For our experiments, we utilized the Python implementation of CRFsuite, which is available through sklearn-crfsuite. We specifically chose the Passive Aggressive (PA) version of the CRFsuite algorithm, favored for its slight bias toward recall. We adhered to the same CRF parameters as in the original study.

Feature Type	**Description**
Token Attributes	Token, lower, is first, is last, is capitalized, is all caps, is all lower
Morphological	prefix-1, prefix-2, prefix-3, suffix-1, suffix-2, suffix-3
Contextual	prev token, next token, is numeric
N-Gram	unigram, bigram, trigram

Table 4: Token-level features employed in the analysis

4.2.1 Feature Engineering

We directly used raw sentences with their corresponding labels from the original dataset, bypassing the featured data used for reproducibility. Our feature engineering approach closely follows the spirit of the original paper, adapted for our own setup.

We customized the Punkt tokenizer using 1 million legal sentences to improve sentence segmentation. Token-level features include lowercase form, capitalization, numeric flags, prefixes/suffixes, and part-of-speech tags. These are aggregated into sparse binary vectors to represent each sentence.

In addition, we added basic sentence-topic compatibility signals such as word overlap with the topic title and description, sentence length, and character count—useful for reducing noisy matches.

While binary n-grams and POS tags individually provided marginal improvements, we observed better performance when they were combined with customized tokenization and case normalization. This is consistent with the observations reported by Roegiest, Hudek, and McNulty (2018), who noted that binary token n-grams were most effective when used alongside carefully tuned preprocessing pipelines.

4.2.2 Results

Roegiest, Hudek, and McNulty (2018) released preprocessed feature vectors for all documents based on proprietary in-house trained and optimized preprocessing, allowing us to reproduce and replicate their experiments above. We study the effectiveness of going back to the source text with "normal" preprocessing choices as used in IR/NLP to understand the impact of their advanced proprietary preprocessing in the original paper.

The results are in Table 2(d), and Table 3(d) looking at the sentence-level and annotation level precision, recall, and F1-score respectively for our version of the CRFsuite (PA) model working on the source text. This model scores sentence-level recall (72.14%) and F1-score (80.89%). This is lower than the proprietary in-house preprocessing results in Table 2(a,b) before, scoring sentence-level recall (85.09%) and F1-score (88.26%). This considerable difference both highlights the value of the proprietary preprocessing, as well as that traditional classification models like CRF require tailored feature engineering in order to excel. Similarly, our standard pre-processing model scores annotation-level recall (86.24%) and F1-score (88.93%). This is lower than the proprietary preprocessing results in original study before, scoring annotation-level recall (94.28%) and F1-score (92.98%). These scores show the value of the proprietary preprocessing and that the model, on relatively standard preprocessing, obtains high levels of effectiveness.

Our feature set did not incorporate some of the advanced n-gram features described in the original study, as they were proprietary. Although the EDGAR documents are publicly available, the Kira Systems features are based on training a word2vec model on proprietary labels and then clustering to create enriched bigram and trigram features. Our main aim here is to assess the effectiveness of standard NLP preprocessing techniques, similar to those used in other IR tasks. The development of optimized features tailored for this specific task, while promising, is left for future work.

To summarize, we can reproduce and replicate the experiments framing the high-recall due diligence task as an information retrieval experiment, both using the original code and with a Python version using the precomputed proprietary features. Additionally, we can replicate the results by returning to the original text, which avoids the proprietary preprocessing of the original paper and enables the model to be applied to new data outside the Kira dataset.

5 An LLM Approach to Due Diligence

This section explores an LLM-based approach to due diligence.

The models discussed in earlier sections are based on extensive labeled training data and hand-crafted feature engineering. Although effective, these traditional approaches assume task-specific supervision at scale, an assumption that is often impractical in legal domains, where annotating data requires costly expert input. This limitation becomes more pressing when applying these models to new or evolving legal topics, where annotated data may be sparse or unavailable.

An alternative class of models, including semantic similarity methods such as Sentence-Transformers, could address some of these challenges by encoding sentences into embedding spaces and retrieving semantically similar passages. However, such approaches typically require additional infrastructure for retrieval (e.g., dense indexing), and may still fall short in capturing task-specific instructions or legal nuances not present in the training data. These methods also struggle with long, multi-sentence contexts or when fine-grained classification decisions are required.

Recent Large Language Models (LLMs) offer a compelling alternative. Not only do they support zero-shot and few-shot learning, but they also allow us to inject detailed task guidance directly into the prompt, enabling inference without additional training or indexing. The Kira dataset, originally designed for human assessors, includes rich topic titles and descriptions that align well with LLM prompt inputs. This motivates us to evaluate whether LLMs, when guided by these descriptions and a small number of examples, can perform due diligence classification effectively, even without large-scale labeled training data.

To this end, we investigate multiple open-source and closed-source LLMs across zero-shot and few-shot prompting conditions, comparing their performance against supervised models on a curated subset of the original dataset.

5.1 Related Work

Generative AI models like GPT-4 (Achiam et al., 2023), Llama (Touvron et al., 2023), and Gemini (Mesnard et al., 2024), have significantly advanced Information Retrieval (IR) and NLP tasks (Ma et al., 2024; Shi et al., 2024). These models can perform complex tasks such as document analysis, summarization, and contract review (Pradeep and Lin, 2024; Jang and Stikkel, 2024; Roegiest, Chitta, et al., 2023). Their performance can improve further with zero-shot and few-shot learning methods (Yu, Quartey, and Schilder, 2022; Sanh et al., 2022; Wang et al., 2024), or in-context learning, which requires little to no fine-tuning. This is especially valuable in legal due diligence, where fast and accurate document interpretation is crucial.

Zero-shot and few-shot learning, combined with prompt engineering, have shown success in legal applications. Roegiest, Chitta, et al. (2023) reported that prompt engineering with GPT-3.5-Turbo outperformed traditional contract analysis methods in accuracy and consistency. Similarly, Jang and Stikkel (2024) demonstrated GPT-4's superior recall in mergers and acquisitions tasks, though models like BERT and Legal-BERT lagged. However, these studies also have limitations. Roegiest, Chitta, et al. (2023) focus solely on question-answering tasks and do not explore full-scale due diligence, which involves retrieving information from diverse legal documents. Jang and Stikkel (2024) demonstrated GPT-4's high recall for due diligence tasks on the Kira dataset but limited their study to a single topic (1243) and 100 samples rather than the complete set of 50 different due diligence topics in the Kira dataset.

5.2 Experimental Setup

Models In this analysis, we have explored the application of open-source LLMs, including dolphin-2.9-llama3-8B, Meta-Llama-3.1-8B, and gemma2-9B (Hartford, Atkins, and Fernandes, 2024; Dubey et al., 2024; Mesnard et al., 2024), alongside OpenAI's latest proprietary GPT-4o-mini model (OpenAI, 2024), which is optimized for fast and less complex tasks. We tested these models for all 50 topics in our dataset, facilitating a direct comparison between proprietary and open-source models. These models were selected based on their performance within the Ollama framework as of July 2024 (Ollama, 2024), known for their local execution capabilities, ensuring data privacy and adaptability to projects with limited computational resources.

We also assessed the earlier CRF-suite model trained on the entire dataset, using the same evaluation set as the LLMs. This directly compares the performance of LLMs against a traditional model trained on a large domain-specific dataset. We evaluate the trained model over the exact same test data, ensuring we select the particular instance of the model depending on the test split the passage was in initially.

We utilized Ollama platform to run open-source models like Dolphin - Llama 3, Llama3.1, and Gemma2, which were tasked with classifying input data as 'Relevant' or 'Not Relevant.' Additionally, we used the OpenAI API to run the GPT-4o-mini model. The performance of these models was assessed using metrics such as Precision, Recall, and F1-score. To assess the comparative performance of the open-source and closed-source models against the CRF-Baseline across 50 topics, we conducted a series of paired sample t-tests.

Topic	Title	Description
1086	Evidence of Loans	To avoid any future debate as to how much the borrower owes, this topic captures provisions which typically set out that a lender's internal records or accounts are conclusive evidence of the amount owed to the lender by the borrower and may further be evidenced by a promissory note by the lender.
1244	Collateral Documents / Security Documents	Where lenders take security/collateral, the rights and obligations of all parties involved are typically contained in separate documents called 'Collateral Documents' or 'Security Documents'. This topic assists in identifying such documents to ensure that there are obligations within the credit agreement for the borrower and other obligors to enter into such documents and to comply with the obligations therein.
1247	Coverage Ratio / Interest Cover	Coverage covenants are negative covenants requiring borrowers to maintain enough income to service interest payments and/or principal repayments under the loan. These are often calculated as ratios between (i) EBITDA (i.e., earnings) and (ii) interest, principal repayments, and/or other regular charges under the credit agreement. This topic captures which, if any, of the various coverage covenants apply.

Table 5: Descriptions of selected legal due diligence topics

Prompts We exploit the detailed descriptions provided in the KIRA data set. Table 5 shows descriptions of selected legal due diligence topics (1086, 1244, 1247). These topic titles and descriptions, originally crafted to guide human domain experts, also serve as detailed task statements that may help direct the LLMs toward the correct labels. *Note that while the supervised models from the original study use task-specific features based on the topic descriptions (such as word overlap and similarity), they do not leverage the full descriptive text or example-based prompts, as we explore with LLMs in this part of the study.*

In addition, we used zero-shot and few-shot example prompts, where in the few-shot case, we show three relevant and three non-relevant examples. Detailed examples are shown in Appendix B. These components were used to build task-specific prompts. Specifically, we provide an increasing amount of information in the prompts, varying *Title Only*, *Title + Description*, and *Title + Description + Examples* prompts using the templates shown in Appendix A.

5.3 Results

Table 6 shows the performance of Dolphin-Llama3, Gemma2, Llama3.1, and GPT-4o-mini was evaluated across 50 topics under three different prompt configurations: *Title Only*, *Title + Description*, and *Title + Description + Examples.* For reference, we also report the

Prompt	**Precision**	**Recall**	**F1-Score**
Supervised (Inference on Evaluation Data)			
CRF-Baseline	0.9991 [0.99, 1.00]	0.8863 [0.87, 0.90]	0.9379 [0.93, 0.95]
Dolphin-llama3			
Title Only	0.4391 [0.37, 0.50]	0.7449 [0.70, 0.79]	0.5029 [0.45, 0.55]
T. + Description	**0.6512**‡ [0.59, 0.71]	0.7677^{-} [0.72, 0.82]	**0.6549**‡ [0.61, 0.70]
T. + D. + Examples	0.3969$^{-,**}$ [0.34, 0.45]	**0.9261**‡,** [0.90, 0.94]	0.5241$^{-,**}$ [0.47, 0.58]
Gemma2			
Title Only	0.7202 [0.65, 0.78]	0.7118 [0.66, 0.77]	0.6781 [0.63, 0.72]
T. + Description	**0.8198**† [0.77, 0.87]	0.7929† [0.75, 0.84]	0.7779‡ [0.74, 0.81]
T. + D. + Examples	0.7974†,* [0.75, 0.85]	**0.8734**‡,** [0.84, 0.91]	**0.8134**‡,** [0.78, 0.85]
Llama3.1			
Title Only	0.6657 [0.59, 0.74]	0.3984 [0.33, 0.46]	0.4542 [0.39, 0.52]
T. + Description	**0.8943**‡ [0.86, 0.93]	0.6168‡ [0.55, 0.68]	0.6916^{-} [0.64, 0.74]
T. + D. + Examples	0.8243‡,** [0.78, 0.87]	**0.8176**‡,** [0.78, 0.85]	**0.8016**‡,** [0.77, 0.83]
Gpt4o-mini			
Title Only	0.8354 [0.78, 0.89]	**0.7044** [0.65, 0.75]	0.7347 [0.69, 0.77]
T. + Description	0.8853‡ [0.85, 0.93]	0.7007^{-}[0.65, 0.75]	**0.7560**$^{-}$ [0.72, 0.80]
T. + D. + Examples	**0.9360**‡,** [0.91, 0.97]	0.6633†,* [0.61, 0.72]	0.7537$^{-,-}$ [0.72, 0.79]

Table 6: Performance Metrics Across Different Prompt Configurations for Dolphin-llama3, Gemma2, Llama3.1, and Gpt4o-mini. Significant differences: † for $p < 0.001$ and ‡ for $p < 0.0001$ when compared to *Title Only*, * for $p < 0.05$ and ** for $p < 0.001$ when compared to *Title + Description*, $^{-}$ for non-significant differences

CRF-Baseline, the supervised CRF model trained on the full Kira dataset but evaluated on the same LLM evaluation subset, serving as a direct benchmark against the LLMs on identical data. We are particularly interested in comparing the performances of open-source models to the closed-source model. We make a number of observations. First, we observe that more in-context learning, adding more context, is generally beneficial. The Title Only setting yields inconsistent results, especially for Dolphin-llama3 and Llama3.1. Including Title + Description refines focus, boosting F1-score, particularly in Gemma2 and GPT-4o-mini. Few-shot learning (Title + Description + Examples) enhances recall (Dolphin-llama3: 93%, Gemma2: 87%) but reduces precision, indicating a trade-off between generalization and false positives. Second, we observe that open-source models perform competitively with the closed-source model. Although GPT-4o-mini outperforms on title-only, the open-source models Llama3.1 and Gemma2 benefit more from the additional context information, and their performance is highest on the few-shot learning prompts. Third, the LLM's performance is approaching the extensively trained CRF baseline closely. Although the CRF-baseline achieves near-perfect precision and F1-score, this is only possible after exhaustive training on labeled data, whereas the LLMs exhibit promising performance without further training or fine-tuning. In particular, some models exhibit competitive recall, which is of key importance in legal due diligence. Obviating the need for extensive training is a key strength as labeled train data is usually not available in real-world due diligence use case.

Our results show the potential of LLMs with few-shot learning for due diligence, delivering acceptable levels of performance in terms of recall under simplified, but class-imbalanced conditions. Although these results cannot be directly compared to the models on the entire Kira dataset as used in the first half of the paper, this is a promising result to develop new models that do not depend on the availability of large-scale training data. This approach can complement traditional models, which excel in the context of massive train data, and also would facilitate the development of further advanced due diligence models for any legal information need.

5.4 Analysis

We conduct further analysis, addressing the following questions: How does the choice of examples in few-shot prompting affect LLM performance? To what extent can LLMs pick up specific legal topics? Do the descriptions and examples provide sufficient guidance for the particular risk the topic targets? Do other closed-source models exhibit similar performance? To address these questions, we conducted three targeted experiments, including some on the selected topics also explored by Jang and Stikkel, 2024. We choose the three topics, 1086, 1244, and 1247, already shown in Table 5 above, based on their complexity: Topic 1086 was the least complex, where the Kira Baseline performed best. Topic 1247 showed moderate difficulty, leading to lower performance, while Topic 1244 was the most challenging, yielding the poorest results across models.

5.4.1 Prompt Sensitivity Analysis

Prompt consistency is a significant challenge in natural language processing, as minor variations in prompt structure can lead to vastly different outcomes (Roegiest, Chitta, et al., 2023). This study focused on a binary classification task to distinguish between relevant and non-relevant information. We investigated the robustness of our prompting mechanism using Gemma2, the best-performing open-source model from our preliminary evaluations. To assess the performance and consistency of the model, we employed four distinct sets of examples in the prompts, resulting in four unique prompt configurations, referred to as P1, P2, P3, and P4. Each prompt variation incorporated different examples to provide contextual information, which is the main factor differentiating each prompt. We then analyzed the performance of these prompts across all 50 topics, comparing their average metrics.

The results suggest that the model remains robust across modest variations in the examples included within the prompt. While some fluctuations in precision and F1-score are observed, the recall values remain consistently high, indicating the model's stability in identifying relevant sentences.

However, we acknowledge that the number of prompt configurations explored is limited to four manually selected sets. Although these were chosen to reflect reasonable diversity in content and phrasing, broader generalizability remains an open question. Further work

Prompt Sensitivity	**Precision**	**Recall**	**F1-Score**
P1	0.7974^{-} [0.75, 0.85]	$\mathbf{0.8734}^{-}$ [0.84, 0.91]	0.8134^{-} [0.78, 0.85]
P2	0.7983^{-} [0.75, 0.85]	0.8685^{-} [0.84, 0.90]	0.8118^{-} [0.78, 0.84]
P3	$\mathbf{0.8028}^{-}$ [0.76, 0.85]	0.8678^{-} [0.84, 0.90]	$\mathbf{0.8174}^{-}$ [0.78, 0.85]
P4	0.7845^{-} [0.73, 0.83]	0.8711^{-} [0.84, 0.90]	0.8048^{-} [0.77, 0.84]

Table 7: Prompt Sensitivity Analysis varying the examples for Gemma2: No significant differences were found among the configurations for Precision, Recall, or F1-Score

could evaluate a larger and more systematically sampled set of prompts to characterize the sensitivity of LLMs better to prompt variation in high-recall legal retrieval tasks.

5.4.2 Cross-Topic Analysis

LLMs demonstrate that the prompts accurately capture the specific topic, as evidenced by the significant performance drop in cross-topic evaluations (Table 8). In this setup, we define the "Prompt Topic" as the topic whose title, description, and examples are used to construct the prompt. The "Test Topic" refers to the topic on which the model is evaluated. In cross-topic experiments, we deliberately mismatch these, using the prompt of one topic while testing on a different topic, to examine whether LLMs rely on topic-specific cues or general patterns. Our goal is to assess how sensitive the models are to the intended task framing and whether performance holds when the prompt does not correspond to the evaluated topic.

While models perform well when the Prompt Topic matches the Test Topic, their ability to classify unseen topics is highly constrained, with F1 scores approaching zero in many cases. This suggests that the model's predictions are indeed guided by the topic context provided in the prompt, and not by generic patterns in the data.

As the different risks in each topic are closely related, models could perform well without precisely capturing the topic's legal meaning. This also validates our experimental subset for evaluating LLMs for due diligence passage retrieval.

The models tend to perform well by simply distinguishing relevant passages from arbitrary non-relevant ones, without fully capturing topic-specific nuances. This tendency, where models exploit superficial patterns such as length, phrasing, or distributional properties rather than true semantic alignment, is a well-known issue in many NLP datasets (Gururangan et al., 2018; McCoy, Pavlick, and Linzen, 2019). To mitigate this, we made particular efforts to sample non-relevant passages with a similar length and word distribution as the relevant ones, ensuring a fairer and more realistic evaluation setup.

5.4.3 Open-source and Closed-source Models

We anticipated closed-source models to outcompete open-source models, yet found that open-source models exhibit performance that meets and sometimes exceeds the closed-source models used in our LLM experiments. But is this due to the use of the smaller model GPT-4o-mini?

Model	1086			1244			1247		
	Prec	Rec	F1	Prec	Rec	F1	Prec	Rec	F1
Prompt for Topic 1086									
Dolphin-llama3	0.30	0.99	0.46	0.08	0.53	0.14	0.017	0.04	0.026
Gemma2	0.71	0.90	0.79	0.20	0.17	0.18	0.00	0.00	0.00
Llama3.1	0.74	0.75	0.75	0.07	0.04	0.05	0.00	0.00	0.00
Prompt for Topic 1244									
Dolphin-llama3	0.21	0.69	0.32	0.13	0.97	0.24	0.01	0.04	0.02
Gemma2	0.00	0.00	0.00	0.39	0.97	0.55	0.00	0.00	0.00
Llama3.1	0.00	0.00	0.00	0.45	0.97	0.62	0.00	0.00	0.00
Prompt for Topic 1247									
Dolphin-llama3	0.21	0.24	0.23	0.0161	0.048	0.024	0.40	0.89	0.55
Gemma2	0.11	0.0091	0.0169	0.00	0.00	0.00	0.82	0.77	0.80
Llama3.1	0.05	0.0091	0.0158	0.00	0.00	0.00	0.84	0.77	0.80

Table 8: Performance Results of Prompt Testing: Cross Topics Analysis (prompts matching the topic are underlined)

Including GPT-4o (Hurst et al., 2024) in our evaluation provides insights into whether an alternative closed-source model offers greater stability than GPT-4o-mini. Additionally, we evaluated the recent open-source DeepSeek-R1:8B (Guo et al., 2025) model on these topics. DeepSeek-R1 was released shortly before we finalized the study. Due to time and computational constraints, we restricted its evaluation to a subset of three representative topics (shown in Table 9), rather than including it in Table 6, which spans all 50 topics.

We show the results for three selected topics for the Title + Description + Examples prompt for all models in Table 9. The results for other prompt conditions are provided in Appendix C.

We observe that the GPT4o model performs marginally better than the smaller GPT4o-mini model and that the DeepSeek-R1:8B model is less effective than the GPT4o models. Again, the closed-source models do not consistently outperform the open-source models.

We observe that GPT-4o consistently delivers strong performance, particularly in F1 score and precision, in the three topics. However, it does not uniformly outperform the best open-source models, especially in recall, where models like Dolphin-Llama3 and Gemma2 show competitive or higher values on certain topics. This highlights that while closed-source models like GPT-4o achieve high precision and balanced F1 scores, open-source models can offer comparable recall performance, which is critical for high-recall tasks like due diligence. Therefore, the advantage of closed-source models is not absolute and varies depending on the metric and task focus.

This is an encouraging outcome, as it signals that the effectiveness of the models cannot be attributed solely to proprietary training and instruction-tuning. Instead, it highlights the value of detailed task descriptions that were originally crafted for human annotators and are now reused in prompt design for LLMs.

Model	**1086**			**1244**			**1247**		
	Prec	**Rec**	**F1**	**Prec**	**Rec**	**F1**	**Prec**	**Rec**	**F1**
Dolphin-llama3	0.30	0.99	0.46	0.14	0.98	0.24	0.40	0.89	0.55
Gemma2	0.71	0.90	0.79	0.39	0.98	0.55	0.83	0.77	0.80
Llama3.1	0.74	0.75	0.75	0.46	0.98	0.62	0.84	0.77	0.80
GPT4o-mini	0.90	0.64	0.75	0.60	0.94	0.73	0.95	0.70	0.81
GPT4o	0.91	0.79	0.85	0.59	0.91	0.69	0.94	0.78	0.85
DeepSeek-R1:8B	0.58	0.86	0.69	0.26	0.98	0.41	0.63	0.83	0.72

Table 9: Performance Metrics Across Few-shot Prompt Configurations (T+D+Examples) for Dolphin-llama3, Gemma2, Llama3.1, GPT4o-mini, GPT-4o and Deepseek-R1:8B across three topics 1086, 1244, 1247.

6 Discussion and Conclusions

We conclude this paper by discussing our findings and drawing conclusions. Our study successfully replicated the legal document retrieval research by Roegiest, Hudek, and McNulty (2018), providing a solid foundation for addressing challenges related to a large dataset of 15 million sentences. We confirmed that traditional machine learning models, such as CRF, can benefit from optimized feature engineering, similar to the proprietary in-house text preprocessing used in the original work.

We extend our reproducibility study with an analysis of recent Large Language Models (LLMs), evaluating their potential for legal due diligence tasks. Unlike traditional models that rely on extensive labeled training data, LLMs demonstrated the ability to classify legal text with minimal supervision using few-shot and zero-shot learning approaches. Our findings highlight that while traditional models excel with large-scale training data, LLMs offer a flexible alternative that can adapt across different topics and domains using prompt-based approaches. The extensive labeled training data, as available in the KIRA collection (Roegiest, Hudek, and McNulty, 2018), is costly to create and usually not readily available. It remains an open question how well the trained classifiers generalize to different applications, including other languages, countries, business practices, or other legal frameworks.

The KIRA collection contains very detailed topic descriptions for typical due diligence tasks. These were used by the legal and regulatory professionals annotating the original due diligence data. These descriptions were not used in any way in the trained models of (Roegiest, Hudek, and McNulty, 2018), which were exhaustively trained on the labeled corpus. Our LLM experiments do not utilize the labeled training data in any way. Interestingly, we demonstrate that these detailed task descriptions are crucial for creating effective prompts.

It is an attractive idea to closely couple the instructions of the human legal professional and the technology-assisted review models used by them, using identical instructions. Compared to annotating extensive corpora, the efforts involved in drafting precise instructions are minimal. This makes it easy to tailor the instruction to other languages, countries, business practices, or other legal frameworks. In addition, rather than relying on relatively generic due diligence topics, such as the 50 topics used in the KIRA data, one can envision updating the instructions to focus on finer-grained topics or tailoring them to the specific case at hand.

Our exploratory experiments with LLMs for due diligence demonstrate their promise and suggest several avenues for further analysis to enhance their effectiveness. In future research, we plan to investigate a larger set of models and transition from sentence-level data to cleaned-up passage-level data, providing more context for models. This is also of interest to further study modern models in terms of high recall and skewed class distributions.

Finally, the primary motivation of this reproducibility study was to promote further research on the challenging task of high-recall legal document and passage retrieval, and to thoroughly analyze these models using simplified approaches similar to those employed in other IR/NLP models. We have made all the code available on GitHub, enabling easy replication of our experiments in Python.

Acknowledgments and Disclosure of Funding

All our code and models are available at https://github.com/UAmsterdam/IRRJ_2025. The code of the original paper is available from https://github.com/zuvaai/science. Both repositories have an empty data directory. A license to use the original Kira dataset needs to be requested from Zuva. Upon showing this license, we will provide access to all data splits and model predictions used in the second set of experiments.

The experiments in this paper were carried out on the National Supercomputer Snellius, which was supported by SURF and the University of Amsterdam's HPC Board. Madhukar Dwivedi is supported by the University of Amsterdam (AI4FinTech program). Jaap Kamps is partly funded by the Netherlands Organization for Scientific Research (NWO CI # CISC.CC.016, NWO NWA # 1518.22.105), the University of Amsterdam (AI4FinTech program), and ICAI (AI for Open Government Lab). Views expressed in this paper are not necessarily shared or endorsed by those funding the research.

References

Achiam, Josh, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, et al. (2023). "GPT-4 Technical Report". In: *CoRR* abs/2303.08774. DOI: 10.48550/ARXIV.2303.08774. arXiv: 2303.08774.

Crammer, Koby, Ofer Dekel, Joseph Keshet, Shai Shalev-Shwartz, and Yoram Singer (2006). "Online Passive-Aggressive Algorithms". In: *Journal of Machine Learning Research* 7, pp. 551–585. URL: https://jmlr.org/papers/v7/crammer06a.html.

Dubey, Abhimanyu, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, et al. (2024). "The Llama 3 Herd of Models". In: *CoRR* abs/2407.21783. DOI: 10.48550/ARXIV.2407.21783. arXiv: 2407.21783.

Guo, Daya, Dejian Yang, Haowei Zhang, Junxiao Song, Ruoyu Zhang, Runxin Xu, et al. (2025). "DeepSeek-R1: Incentivizing Reasoning Capability in LLMs via Reinforcement

Learning". In: *CoRR* abs/2501.12948. doi: 10.48550/arXiv.2501.12948. arXiv: 2501.12948.

Gururangan, Suchin, Swabha Swayamdipta, Omer Levy, Roy Schwartz, Samuel Bowman, and Noah A. Smith (2018). "Annotation Artifacts in Natural Language Inference Data". In: *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 2 (Short Papers)*. Ed. by Marilyn Walker, Heng Ji, and Amanda Stent. New Orleans, Louisiana: Association for Computational Linguistics, pp. 107–112. doi: 10.18653/v1/N18-2017.

Hartford, Eric, Lucas Atkins, and Fernando Fernandes (2024). *Dolphin 2.9 Llama 3 8b*. url: https://huggingface.co/cognitivecomputations/dolphin-2.9-llama3-8b.

Hurst, Aaron, Adam Lerer, Adam P. Goucher, Adam Perelman, Aditya Ramesh, Aidan Clark, et al. (2024). "GPT-4o System Card". In: *CoRR* abs/2410.21276. doi: 10.48550/ARXIV.2410.21276. arXiv: 2410.21276.

Jang, Myeongjun and Gábor Stikkel (2024). "Leveraging Natural Language Processing and Large Language Models for Assisting Due Diligence in the Legal Domain". In: *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 6: Industry Track)*. Ed. by Yi Yang, Aida Davani, Avi Sil, and Anoop Kumar. Mexico City, Mexico: Association for Computational Linguistics, pp. 155–164. doi: 10.18653/v1/2024.naacl-industry.14.

Klaber, Ben (2013). "Artificial Intelligence and Transactional Law: Automated M&A Due Diligence". In: *International Conference on Artificial Intelligence and Law, DESI V Workshop*. url: https://users.umiacs.umd.edu/~oard/desi5/additional/Klaber.pdf.

Langford, John, Lihong Li, and Alexander Strehl (2025). *Vowpal Wabbit open source project*. url: https://vowpalwabbit.org/.

Ma, Xueguang, Liang Wang, Nan Yang, Furu Wei, and Jimmy Lin (2024). "Fine-Tuning LLaMA for Multi-Stage Text Retrieval". In: *Proceedings of the 47th International ACM SIGIR Conference on Research and Development in Information Retrieval*. SIGIR '24. Washington DC, USA: Association for Computing Machinery, pp. 2421–2425. isbn: 9798400704314. doi: 10.1145/3626772.3657951.

McCoy, R. Thomas, Ellie Pavlick, and Tal Linzen (2019). "Right for the Wrong Reasons: Diagnosing Syntactic Heuristics in Natural Language Inference". In: *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*. Ed. by Anna Korhonen, David Traum, and Lluís Màrquez. Florence, Italy: Association for Computational Linguistics, pp. 3428–3448. doi: 10.18653/v1/P19-1334.

Mesnard, Thomas, Cassidy Hardin, Robert Dadashi, Surya Bhupatiraju, Shreya Pathak, Laurent Sifre, et al. (2024). "Gemma: Open Models Based on Gemini Research and Technology". In: *CoRR* abs/2403.08295. doi: 10.48550/ARXIV.2403.08295. arXiv: 2403.08295.

Moriarty, Ryan, Howard Ly, Ellie Lan, and Suzanne K. McIntosh (2019). "Deal or No Deal: Predicting Mergers and Acquisitions at Scale". In: *2019 IEEE International Conference on Big Data (IEEE BigData), Los Angeles, CA, USA, December 9-12, 2019*. Ed. by Chaitanya K. Baru, Jun Huan, Latifur Khan, Xiaohua Hu, Ronay Ak, Yuanyuan Tian, et al. IEEE, pp. 5552–5558. doi: 10.1109/BIGDATA47090.2019.9006015.

Nocedal, Jorge (1980). "Updating quasi-Newton matrices with limited storage". In: *Mathematics of computation* 35.151, pp. 773–782.

Ollama (2024). *Ollama: Local Large Language Model Runner*. Accessed: 2024-06-27. url: https://github.com/ollama/ollama.

OpenAI (2024). *GPT-4o mini: advancing cost-efficient intelligence*. https://openai.com/index/gpt-4o-mini-advancing-cost-efficient-intelligence/.

Parikh, Akshat, Samit Poojary, and Aadit Gupta (2023). "AMP — Optimizing M&A Outcomes: Harnessing the Power of Big Data Analytics and Natural Language Processing". In: *International Journal of Data Science and Big Data Analytics* 3 (2). doi: 10.51483/IJDSBDA.3.2.2023.35-50.

Pradeep, Ronak and Jimmy Lin (2024). "Towards Automated End-to-End Health Misinformation Free Search with a Large Language Model". In: *Advances in Information Retrieval - 46th European Conference on Information Retrieval, ECIR 2024, Glasgow, UK, March 24-28, 2024, Proceedings, Part IV*. Ed. by Nazli Goharian, Nicola Tonellotto, Yulan He, Aldo Lipani, Graham McDonald, Craig Macdonald, et al. Vol. 14611. Lecture Notes in Computer Science. Springer, pp. 78–86. doi: 10.1007/978-3-031-56066-8_9.

Roegiest, Adam, Radha Chitta, Jonathan Donnelly, Maya Lash, Alexandra Vtyurina, and Francois Longtin (2023). "Questions about Contracts: Prompt Templates for Structured Answer Generation". In: *Proceedings of the Natural Legal Language Processing Workshop 2023*. Ed. by Daniel Preoțiuc-Pietro, Catalina Goanta, Ilias Chalkidis, Leslie Barrett, Gerasimos Spanakis, and Nikolaos Aletras. Singapore: Association for Computational Linguistics, pp. 62–72. doi: 10.18653/v1/2023.nllp-1.8.

Roegiest, Adam, Alexander K. Hudek, and Anne McNulty (2018). "A Dataset and an Examination of Identifying Passages for Due Diligence". In: *The 41st International ACM SIGIR Conference on Research & Development in Information Retrieval, SIGIR 2018, Ann Arbor, MI, USA, July 08-12, 2018*. Ed. by Kevyn Collins-Thompson, Qiaozhu Mei, Brian D. Davison, Yiqun Liu, and Emine Yilmaz. ACM, pp. 465–474. doi: 10.1145/3209978.3210015.

Sanh, Victor, Albert Webson, Colin Raffel, Stephen Bach, Lintang Sutawika, Zaid Alyafeai, et al. (2022). "Multitask Prompted Training Enables Zero-Shot Task Generalization". In: *International Conference on Learning Representations.* URL: https://openreview.net/forum?id=9Vrb9D0WI4.

Sherer, James A., Taylor M. Hoffman, and Eugenio E. Ortiz (2015). "Merger and acquisition due diligence: a proposed framework to incorporate data privacy, information security, e-discovery, and information governance into due diligence practices". In: *Richmond Journal of Law & Technology* 21.2, p. 5. URL: https://scholarship.richmond.edu/jolt/vol21/iss2/3.

Sherer, James A., Taylor M. Hoffman, Kevin M. Wallace, Eugenio E. Ortiz, and Trevor J. Satnick (2016). "Merger and acquisition due diligence part II-the devil in the details". In: *Richmond Journal of Law & Technology* 22.2, p. 4. URL: https://scholarship.richmond.edu/jolt/vol22/iss2/2/.

Shi, Yunxiao, Xing Zi, Zijing Shi, Haimin Zhang, Qiang Wu, and Min Xu (2024). "Enhancing Retrieval and Managing Retrieval: A Four-Module Synergy for Improved Quality and Efficiency in RAG Systems". In: *ECAI 2024 - 27th European Conference on Artificial Intelligence, 19-24 October 2024, Santiago de Compostela, Spain - Including 13th Conference on Prestigious Applications of Intelligent Systems (PAIS 2024).* Ed. by Ulle Endriss, Francisco S. Melo, Kerstin Bach, Alberto José Bugarín Diz, Jose Maria Alonso-Moral, Senén Barro, et al. Vol. 392. Frontiers in Artificial Intelligence and Applications. IOS Press, pp. 2258–2265. DOI: 10.3233/FAIA240748.

Touvron, Hugo, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, et al. (2023). "LLaMA: Open and Efficient Foundation Language Models". In: *CoRR* abs/2302.13971. DOI: 10.48550/ARXIV.2302.13971. arXiv: 2302.13971.

Wang, Shuai, Harrisen Scells, Shengyao Zhuang, Martin Potthast, Bevan Koopman, and Guido Zuccon (2024). "Zero-Shot Generative Large Language Models for Systematic Review Screening Automation". In: *Advances in Information Retrieval - 46th European Conference on Information Retrieval, ECIR 2024, Glasgow, UK, March 24-28, 2024, Proceedings, Part I.* Ed. by Nazli Goharian, Nicola Tonellotto, Yulan He, Aldo Lipani, Graham McDonald, Craig Macdonald, et al. Vol. 14608. Lecture Notes in Computer Science. Springer, pp. 403–420. DOI: 10.1007/978-3-031-56027-9_25.

Yu, Fangyi, Lee Quartey, and Frank Schilder (2022). "Legal Prompting: Teaching a Language Model to Think Like a Lawyer". In: *CoRR* abs/2212.01326. DOI: 10.48550/ARXIV.2212.01326. arXiv: 2212.01326.

Appendix A. Prompt Templates

A.1 Title Only

This configuration has the least context; we only provide the topic title generated by experts, omitting any descriptions or examples (as shown in Table 10). This method takes advantage of the model's intrinsic ability to interpret the task on its own and is, therefore, a direct test of the model's pre-trained knowledge.

```
Objective:
Your task is to determine if the provided text contains ‘relevant’ information concerning
‘topic title’. This involves identifying information directly related to the specified
topic, which in a legal or financial document might pertain to specific clauses, terms,
or conditions.
Instructions for Response Format:
Analyze the text provided and determine its relevance based on the specifics of ‘topic
title’ and its implications. Provide your analysis in the following format:
Answer: [Relevant/Not Relevant]

Text for Analysis:
‘input_sentence’
```

Table 10: Prompt template for Title Only analysis

A.2 Title + Description

This configuration provides more context than the Title Only setting by including expert-generated descriptions but no examples (as shown in Table 11). It tests whether these descriptions effectively guide the models in understanding the task, making it a valuable assessment of how well expert-written descriptions enhance the overall performance of LLM models.

```
Objective:
Your task is to determine if the provided text contains ‘relevant’ information concerning
‘topic title’. This involves identifying information directly related to the specified
topic, which in a legal or financial document might pertain to specific clauses, terms,
or conditions.
Topic Definition:
‘Topic description’

Instructions for Response Format:
Analyze the text provided and determine its relevance based on the specifics of ‘topic
title’ and provided ‘topic description’. Provide your analysis in the following format:
Answer: [Relevant/Not Relevant]

Text for Analysis:
‘input_sentence’
```

Table 11: Prompt template for Title + Description analysis

A.3 Title + Description + Examples

In this approach (Table 12), the prompts were given with additional context, including the topic title, description, and six examples—three labeled as ‘Relevant’ and three as ‘Not Relevant.’ By incorporating these examples, the few-shot prompts provide the model with specific guidance and improve its focus, enabling a more precise and informed analysis.

```
Objective:
Review the provided text to determine if it contains relevant information concerning
‘topic title’. Relevant information directly discusses risks or specifics related to
the topic title, pledged in financial transactions.
Topic Definition:
‘Topic description’

Examples: Here are examples for each class
Relevant:
‘[...]’
Not Relevant:
‘[...]’

Instructions for Response Format:
Analyze the text provided and determine its relevance based on the specifics of ‘topic
title’ and its implications. Provide your analysis in the following format:
Answer: [Relevant/Not Relevant]

Text for Analysis:
‘input_sentence’
```

Table 12: Prompt template for Title + Description + Examples (few-shot) analysis with three examples per class

Appendix B. Few Shot Examples

Table 13 shows detailed examples used in the prompt for Topic 1242.

Appendix C. Detailed Per Topic Results

Table 14 shows the results of different prompt configurations across all models, including GPT-4o and Deepseek-R1:8B, on three selected topics (1086, 1244, and 1247).

GPT-4o demonstrates slightly more consistent performance across different prompt configurations, maintaining stable F1 scores across the Title Only, Title + Description, and Title + Description + Examples settings. However, its improvements over GPT-4o-mini are marginal, with no significant performance gap between the two models. This suggests that GPT-4o shows better stability but does not provide a substantial advantage over GPT-4o-mini in legal classification tasks.

Deepseek-R1:8B demonstrated reasonable recall improvements in the Title + Description + Examples setting. Its overall performance remained below that of GPT-4o, particularly in precision, indicating a tendency towards overclassification. These findings highlight that LLM performance in legal due diligence remains highly dependent on structured prompting. While closed-source models like GPT-4o offer stability, their advantages over well-optimized open-source alternatives remain limited.

#	Relevant Examples	Non-Relevant Examples
	Prompt 1	
1	‘” Consolidated Fixed Charge Coverage Ratio ” shall mean , for any Test Period , the ratio of (a) the sum of (i) Consolidated Adjusted EBITDA for such Test Period minus (ii) the aggregate amount of Consolidated Capital Expenditures for such period (other than financed with the incurrence of Indebtedness (other than Loans hereunder or under the Term Loan Agreement)) to (b) Consolidated Fixed Charges for such Test Period .’	‘The Australian Security Agreements , upon execution and delivery thereof by the parties thereto , will create in favor of the Collateral Agent (or the Australian Security Trustee) , for the ratable benefit of the Secured Parties , a legal , valid , enforceable and perfected First Priority Lien in the ” Collateral ” (as defined in the relevant Australian Security Agreements) of the Loan Parties party to such documents to the extent set forth therein .’
2	‘” Fixed Charge Coverage Ratio ” shall mean , as of any date , the ratio of (i) EBITDAR to (ii) the sum of (A) Debt Service plus (B) Rents , in each case for the immediately preceding four fiscal quarters ended on or closest to such date ;’	‘In the event of any conflict between the accounts and records maintained by the Administrative Agent and the accounts and records of any Lender in respect of such matters , the accounts and records of the Administrative Agent shall control in the absence of manifest error .’
3	‘” Consolidated Interest Coverage Ratio ” means , as of any date of determination , the ratio of (a) Consolidated EBITDA for the period of the four prior fiscal quarters ending on such date to (b) Consolidated Interest Charges for such period .’	‘(b) neither the Administrative Agent nor any other Secured Party has any fiduciary relationship with or duty to any Grantor arising out of or in connection with this Agreement or any of the other Loan Documents , and the relationship between the Grantors , on the one hand , and the Administrative Agent and the other Secured Parties , on the other hand , in connection herewith or therewith is solely that of debtor and creditor ;’
	Prompt 2	
1	‘provided that with respect to cost savings or synergies relating to any Sale , Purchase or other transaction , the related actions are expected by the Borrower Representative to be taken no later than 18 months after the date of determination .’	‘In addition , each new Wholly-Owned Subsidiary that is required to execute any Credit Document shall execute and deliver , or cause to be executed and delivered , all other relevant documentation (including opinions of counsel) of the type described in Section 6 as such new Subsidiary would have had to deliver if such new Subsidiary were an Obligor on the Second Restatement Effective Date .’
2	‘” Interest Coverage Ratio ” means the ratio as of the last day of any Fiscal Quarter of (i) Consolidated Adjusted EBITDA for the four-Fiscal Quarter period then ending , to (ii) Consolidated Corporate Interest Expense for such four-Fiscal Quarter period .’	‘(b) Each of the Arranger and the Lenders authorises the Agent to perform the duties , obligations and responsibilities and to exercise the rights , powers , authorities and discretions specifically given to the Agent under or in connection with the Finance Documents together with any other incidental rights , powers , authorities and discretions .’
3	‘for the period of the four prior fiscal quarters of the Parent Borrower ending on the Calculation Date to (II) Consolidated Interest Expense paid or payable in cash during such period (together with any sale discounts given in connection with sales of accounts receivable and / or inventory by the Consolidated’	‘(b) Any such request shall be made to the Administrative Agent not later than 11 #C# 00 a.m. (Chicago , Illinois time) , twenty (20) Business Days prior to the date of the desired Borrowing or issuance (or such other time or date as may be agreed by the Administrative Agent and , in the case of any such request pertaining to Letters of Credit , the applicable Fronting Bank , in its or their sole discretion) .’

Table 13: In Context Learning Examples for Topic 1247 on *Coverage Ratio/Interest Cover*

Model/Prompt	**1086**			**1244**			**1247**		
	Prec	**Rec**	**F1**	**Prec**	**Rec**	**F1**	**Prec**	**Rec**	**F1**
Dolphin-llama3									
Title Only	0.14	0.75	0.24	0.15	0.94	0.25	0.72	0.60	0.65
T. + Description	0.68	0.87	**0.76**	0.33	0.96	**0.49**	0.88	0.66	**0.76**
T. + D. + Examples	0.30	0.99	0.46	0.14	0.98	0.24	0.40	0.89	0.55
Gemma2									
Title Only	0.39	0.60	0.47	0.20	1.0	0.38	0.92	0.68	0.78
T. + Description	0.90	0.72	**0.80**	0.33	0.98	0.49	0.91	0.76	**0.83**
T. + D. + Examples	0.71	0.90	0.79	0.39	0.98	**0.55**	0.83	0.77	0.80
Llama3.1									
Title Only	0.03	0.01	0.02	0.34	0.57	0.43	0.93	0.16	0.27
T. + Description	0.90	0.34	0.49	0.55	0.89	**0.67**	0.93	0.65	0.76
T. + D. + Examples	0.74	0.75	**0.75**	0.46	0.98	0.62	0.84	0.77	**0.80**
GPT4o-mini									
Title Only	0.20	0.76	0.32	0.33	0.96	0.49	0.92	0.72	0.81
T. + Description	0.83	0.79	**0.81**	0.47	0.94	0.63	0.92	0.67	0.77
T. + D. + Examples	0.90	0.64	0.75	0.60	0.94	**0.73**	0.95	0.70	**0.81**
GPT4o									
Title Only	0.28	0.67	0.39	0.38	0.96	0.55	0.95	0.67	0.79
T. + Description	0.92	0.78	**0.85**	0.37	0.96	0.53	0.91	0.76	0.83
T. + D. + Examples	0.91	0.79	**0.85**	0.59	0.91	**0.69**	0.94	0.78	**0.85**
DeepSeek-R1:8B									
Title Only	0.13	0.84	0.23	0.07	0.98	0.14	0.29	0.88	0.44
T. + Description	0.44	0.87	0.59	0.21	0.95	0.35	0.62	0.78	0.70
T. + D. + Examples	0.58	0.86	**0.69**	0.26	0.98	**0.41**	0.63	0.83	**0.72**

Table 14: Performance Metrics Across Different Prompt Configurations for Dolphin-llama3, Gemma2, Llama3.1, GPT4o-mini, GPT-4o and Deepseek-R1:8B across three topics 1086, 1244, 1247.

Information Retrieval Research 1 (2025) 247-280 Submitted 05/25; Revised 11/25; Published 12/25

An Eigensystem for Topic Term Weighting yields Fair and Effective Document Rankings

Massimo Melucci MASSIMO.MELUCCI@UNIPD.IT
Department of Information Engineering
University of Padova, 35131 Padova, Italy, EU

Editor: Johanne Trippas

Abstract

By incorporating fairness into the ranking function at retrieval time through topic term weights, this paper suggests a way to at the same time achieve effective and fair rankings in document collections. The topic term weights are calculated using the eigensystem of a matrix that linearly combines effectiveness and fairness matrices in a single matrix, whose main eigenvector provides the weights. The latter can then be utilized as a new topic representation, thus providing a fair ranking while minimizing the impact on effectiveness. The experiments described in the paper demonstrate that the proposed approach works.

Keywords: group fairness, probability ranking principle, eigensystem

1 Introduction

Many information management systems strive to suit the user's information needs by providing access to enormous data repositories and delivering optimal outcomes by organizing output data in several formats. Ranking is a common form used by recommender and Information Retrieval (IR) systems. Because of the introduction of normative conceptions such as fairness or equality in the late 2010s, all of these systems have struggled to balance both user satisfaction and producer or author visibility. Visibility is crucial for ensuring that diverse voices and perspectives are represented in the information ecosystem.

In IR, one finds that the principal task has traditionally been articulated in terms of relevance: retrieve those documents most likely to be relevant to a user's query or pertinent to a topic.[1] Much effort has been devoted to the construction of models that estimate, more or less precisely, this elusive probability of relevance. Within this framework, the representation of both documents and queries as weight distributions over terms is critical. To be specific, a document d is at retrieval time assigned a usually non-negative score measuring the likelihood that the IR system will retrieve it to respond to a given topic q. Let

$$y(q,d) = \sum_{t \in q \cap d} w(t,d)\kappa(t,q) \tag{1}$$

be the retrieval score given to the document to answer a topic, where $w(t,d)$ is a function capturing the association between term t and document, such as term frequency, probabilis-

[1] In this paper, we also use "topic" because the test collections utilized in the experiments include topics rather than real queries.

WWW: https://irrj.org DOI: 10.54195/irrj.23480

tic relevance score, or contextual embedding similarity, and $\kappa(t,q)$ has analogous meaning yet it is usually set to 1. Next, we rank the retrieved documents based on their score. Topic term weights are often set to a constant value, e.g., 1, because experimenters have no clue, nor do users provide information as regards the importance of topic terms.

One particular concern is that retrieval systems may systematically under-rank documents associated with certain categories, especially when these are underrepresented in data or associated with less frequent terms. We consider the problem of fair ranking—that is, ranking documents most likely to be relevant to a user's query or pertinent to a topic and fairly representative of the categories to which they refer. The problem bears some resemblance to the approaches for fairness in classification: one adjusts the influence of particular inputs to ensure parity across outputs. In IR, however, the problem is compounded by the necessity of preserving relevance.

The difficulty also recalls the remarks made by Robertson (1977) on the trade-off between optimal ranking and clustering. According to Robertson, ranking documents by decreasing their relevance to information contradicts the cluster hypothesis, which states that the best document clusters should be ranked first. This is because relevant documents tend to cluster together more than non-relevant documents. This clustering effect can lead to a situation where the most relevant documents are not necessarily the top-ranked ones, as they may be grouped with other relevant documents in the same cluster. As a result, this trade-off requires careful consideration when designing retrieval systems to balance the benefits of both ranking and clustering effectively.

To obtain a fair yet effective ranking, a suite of options is available to the designers of a system. One option is to directly permute document rankings and optimize a certain function of utility that balances both effectiveness and fairness. Another option is to indirectly obtain a new ranking by changing the term weights. It is in the manipulation of the term weight distributions that we encounter opportunities not only to improve retrieval effectiveness but also to simultaneously address broader concerns, notably that of fairness. Suppose we identify that certain terms—associated with marginalized topics, communities, or discourses—are chronically underweighted in a baseline model. By explicitly increasing the weights of such terms during re-ranking, we can enhance the visibility of documents pertaining to these underrepresented areas. A possible procedure may start from the ranking function (1) which can be expanded as:

$$y(q,d) = \sum_{t \in q \cap d} w(t,d)x(t,q) \tag{2}$$

where $x(t,q)$ is the weight assigned to the term. The crux here is that by using x, we can raise or lower the profile of particular topic terms, thus reshaping the final ranking also to the aim of making the ranking fairer than the initial ranking. As a consequence, the variation of a topic term weight should not only award the terms that promote relevant documents and demote non-relevant documents, it should also award the terms that promote underrepresented category documents and demote overrepresented category documents. The challenge is to balance promotions and demotions.

However, a naive reweighting that elevates fairness at the cost of user satisfaction is unlikely to be acceptable. Thus, the art lies in achieving a balance wherein gains in fairness do not substantially degrade effectiveness, or where modest losses in effectiveness are justified by substantial fairness improvements. Such intervention, to be clear, need not be ad hoc; on the contrary, a principled approach was adopted in this work.

Group fairness is concerned with the balanced exposure of groups. In the event of IR, group fairness is concerned with the balanced exposure of the groups of document author's organization to the end users which interact with a system to retrieved documents and meet information needs. The requirement of group fairness derives from the hypothesis that minority organization groups are less likely exposed to the end user by means of the retrieved documents.

We considered group fairness and addressed the problem of fair ranking as follows. An initial list of documents most likely to be relevant to a user's query or pertinent to a topic is ranked by some baseline method. By assigning weights to individual terms according to their topical relevance, one may modulate the influence of different query or topic components on the final ranking. This method may, under suitable conditions, yield improved effectiveness. But our claim here is more ambitious: that topic term weighting, suitably construed, can serve also as an instrument of fairness. The topic term weights thus serve a dual role: guiding the model toward content representation and simultaneously adjusting for societal disparities encoded in the data.

2 Previous Work on Fairness and Information Retrieval

The problem of fairness in IR is novel in the sense that explicit discussions of fairness have only relatively recently entered the vocabulary of IR researchers, in part due to a broader societal concern with bias, accountability, and transparency in algorithmic systems, as witnessed by Baeza-Yates (2018). But it is not novel in the sense that any systems that provide ranked outputs have always made distinctions, highlighted some items and omitted others, and thereby implicitly favored some items over others, as investigated in result diversification; see McDonald et al. (2022); Chapelle et al. (2011). The question, then, is not whether IR systems are fair, but rather: how do they treat the concept of fairness, and ought they be doing more?

Actually, the classical IR framework was never designed with fairness in mind, at least not as a first-class concern. The objective was to retrieve documents that are relevant to a given query or pertinent to a given topic; see Van Rijsbergen (1979); Salton (1968). Relevance was at the center of the whole process; see Robertson et al. (1982). The evaluation framework, exemplified by Cranfield-style test collections, built upon this centrality of relevance: precision, recall, average precision, and later, NDCG, all serve to measure relevance.

But of course, as the scope of retrieval systems has expanded—from digital libraries to web search, from enterprise to social media—the stakes have changed. In recommending job ads, loan offers, or news articles, the system is not only reflecting some notion of relevance, which itself is complex as well explained by Saracevic (1975, 2007a,b) and by Mizzaro (1997), but it is also participating in a social process, one in which the distribution of attention, opportunity, and information may have significant consequences. A discussion was provided by Balagopalan et al. (2023) which applied concepts from social sciences, information retrieval, and machine learning to empirically investigate if breaches of criteria for combining effectiveness and fairness had an influence, and if so, how to avoid such concerns.

Let us take an example. Suppose a job search platform retrieves and ranks positions for a user based on their profile, queries, topics, and past interactions. The system may optimize for click-through rate or predicted user satisfaction. But what if this optimization consistently results in showing high-paying technical roles to male users and lower-paying service jobs to female users, even when their qualifications are similar? Is the system "fair"? In the traditional sense of relevance, perhaps it is—if those results align with inferred user preferences. But from a broader societal view, the answer is more troubling.

This disjunction—between individual relevance and societal fairness—is at the heart of the fairness-in-IR discussion. It invites a number of further questions. Can we define fairness in retrieval systems in a rigorous way? If so, how might we measure it? And—most crucially—how might it trade off against other objectives, notably relevance?

Several proposals have emerged in recent literature. One line of work attempts to define group fairness in ranked outputs: ensuring that members of different groups (say, based on gender or ethnicity) are represented equitably in the top-k results; see Binns (2020); Jaenich et al. (2024); Melucci (2024a,b, 2025a,b); Morik et al. (2020); Sakai et al. (2023). Another strand focuses on individual fairness, the idea that similar individuals (in some feature space) should receive similar outcomes; see Biega et al. (2018); Binns (2020); Draws et al. (2021); Ilvento et al. (2020); Kletti et al. (2022); Xiao et al. (2017). Both of these concepts, while intuitive, face considerable practical challenges. Group fairness may require knowledge of sensitive attributes (which may not be available or legally permissible to use); individual fairness depends critically on the definition of similarity (which is itself a value-laden and context-sensitive choice).

In retrieval settings, the notion of fairness becomes further complicated by the ranked nature of outputs. Not only must one consider which items are retrieved, but also where they appear in the ranking. Users typically examine only the top few results—so the fairness implications of a system are concentrated in those positions. A group may be underrepresented although the documents that belong to the group are ranked not far from the top-k hits. As a result, measures of fairness must be sensitive to rank position, much as traditional effectiveness measures are.

Here, one might consider analogues of well-known metrics. For instance, Normalized Discounted Cumulative Gain (NDCG) gives more weight to items at higher ranks—see Jarvëlin and Kekäläinen (2002). One might define Discounted Cumulative Fairness, where the representation of groups in top ranks is weighted more heavily. But this, too, introduces tension. A system that seeks to maximize NDCG may produce rankings that are highly effective but group-imbalanced; a system that enforces strict fairness constraints may compromise on user satisfaction. This tension is real, and it cannot be wished away.

Some researchers have proposed re-ranking methods to address this: first, produce a ranking based on relevance; then, adjust it to improve fairness. Adjusting a ranking for fairness may require demoting highly relevant items, which can degrade performance—we adopted this approach in our previous work; see Melucci (2024a,b, 2025a,b). Alternatively, one may attempt to build fairness into the scoring function itself—this, however, raises difficult questions of how to balance relevance and fairness at the model level. It is indeed this approach and the difficulties thereof we addressed in this work. Another approach was reported by Bigdeli et al. (2021), where the authors empirically investigated the tradeoff between fairness and effectiveness and saw if it is possible to increase fairness while preserving equivalent retrieval utility. They investigated if this is feasible by rewriting the input query using a bias-aware pseudo-relevance feedback system.

Recent work by Jaenich et al. (2025) proposes generative query reformulation to improve fairness without compromising relevance—an approach closely aligned with the goals of this paper. Instead of modeling fair reranking as this paper proposes, Jaenich et al. (2025) found that Large Language Models (LLMs) may provide a significant improvement in some cases yet at the expenses of high computational costs.

3 Are Test Collections Fair?

The focus on relevance has driven the design of experiments in IR and, in particular, the preparation of test collections. The test documents are harvested from various sources and assembled in a repository according to some categories and objectives. At the time of document harvest, a set of test topics is selected. The experiments consist of document retrieval, i.e., an IR system retrieves documents and selects those likely relevant to a topic. At retrieval time, the system ranks the retrieved documents by likelihood of relevance to topics. A test collection also includes relevance assessments, i.e., some users judge document relevance to topics.

The preparation and the use of a test collection can affect IR fairness in different ways. At harvest time, designers and experimenters may affect the distribution of documents, topics, and relevance assessments because of data availability, source authoritativeness, technological reasons, language barriers, or cultural bias. At retrieval time, a system may affect the distribution of the retrieved documents because of stopword removal, stemming, or distribution of document and topic terms. When ranking documents, a system may affect the distribution of ranked documents because of fairness-unaware term weighting. The relevance assessments of a test collection may also affect the distribution of ranked documents by categories because of the assessors' bias towards unprotected categories and against protected categories.

Consider a topic q and a document subset X. In this paper, we define the score of X with respect to q as

$$y(q, X) = \sum_{d \in X} y(q, d) \ .$$

The score of X measures the degree to which a user will find relevant information from the documents in X if s/he examines all the documents. In other terms, the score of X is a measure provided by the system of the "quantity" of relevance found in X. Following the foundations of probability theory, the probability of X with respect to a topic is defined as

$$\Pr(X) = \frac{y(q, X)}{\sum_d y(q, d)} \ .$$

Let Y be another document subset. In this paper, we define

$$\Pr(X, Y) = \Pr(X \cap Y) \qquad \Pr(X|Y) = \frac{\Pr(X, Y)}{\Pr(Y)} \ .$$

Let $\Pr(C|X)$ be a measure of category exposure within X and can therefore be used to measure fairness. Let $\mathcal{C}$ be a set of categories.

The Gini index of mutability

$$G(\mathcal{C}|X) = 1 - \sum_{C \in \mathcal{C}} \Pr(C|X)^2$$

measures fairness for each distribution conditioned to X. Gini's measure of mutability was reported more than a century ago in (Gini, 1912, pages 142–144), but no English translation has been available until Ceriani and Verme (2024) published "an abridged and commented translation of this second part of" Gini's book. The translation is crucial since mutability is commonly misunderstood with variability and the former is often ignored. We believe it is beneficial to deliver a quick presentation.

When a category, such as gender, is observed in a group of people, an unordered list of category values is eventually produced. An unordered series cannot have extreme values, nor can median or average values be derived; this is why variance and quantiles cannot be calculated. Nonetheless, an unordered series can have its relative frequency estimated.

Let X be a set of individuals, and $\mathcal{C} = C_1, ..., C_m$ be a collection of categories such that $C_k \subseteq X$ and $C_\ell \cap C_k = \emptyset$. Let $f_k = |C_k|/|X|$, where $n = |X|$. When an individual belongs to C_k, one deviation $1 - f_k/n$ from the relative frequency and m deviations $f_\ell/n, \ell = 1, ..., k-1, k+1, ..., m$ from the other relative frequencies are observed. The sum of the variances caused by witnessing the k-th value in one individual is

$$1 - f_k/n + f_1/n + \cdots + f_{k-1}/n + f_{k+1}/n + \cdots + f_m/n = 2\frac{n - f_k}{n} \; .$$

Because the f_k individuals display the k-th value, the overall deviation will be $2 f_k \frac{n - f_k}{n}$. After seeing the attribute in each individual, the overall deviation is

$$\frac{2}{n} \sum_{k=1}^{m} f_k (n - f_k) \; .$$

Because each individual is considered twice, once for negative deviations and once for positive deviations, the average deviation can be written as

$$\begin{aligned} G(\mathcal{C}|X) &= \frac{1}{n^2} \sum_{k=1}^{m} f_k (n - f_k) \\ &= 1 - \sum_{k=1}^{m} \left(\frac{f_k}{n} \right)^2 \; . \end{aligned} \tag{3}$$

Equation (3) is Gini's index of mutability for the set X with respect to the collection of categories $\mathcal{C}$; see the original book by Gini (1912) as well as Ceriani and Verme (2024)'s article for a commented translation.

The test collections from the Text Retrieval Conference (TREC) Fair tracks in 2021 and 2022 were used to measure the extent to which fairness can be affected; see Ekstrand et al. (2022a, 2023) and Section 5. At TREC, evaluation takes place in two steps, i.e., training and

Track	Category	Group	Pr(C)
2021	gender	Female	0.056
2021	gender	Male	0.239
2021	gender	Third	0.000
2021	gender	Unknown	0.706
2021	locations	Africa	0.022
2021	locations	Antartica	0.002
2021	locations	Asia	0.099
2021	locations	Europe	0.212
2021	locations	LAC	0.031
2021	locations	N. America	0.187
2021	locations	Oceania	0.026
2021	locations	Unknown	0.422
2022	gender	Man	0.231
2022	gender	Non-binary	0.000
2022	gender	Woman	0.055
2022	gender	Unknown	0.714
2022	locations	ANZ	0.032
2022	locations	E. Africa	0.002
2022	locations	N. Africa	0.001
2022	locations	S. Africa	0.003
2022	locations	W. Africa	0.002
2022	locations	C. America	0.005
2022	locations	N. America	0.228
2022	locations	S. America	0.008
2021	locations	Antartica	0.000
2022	locations	C. Asia	0.001
2022	locations	E. Asia	0.018
2022	locations	S. Asia	0.021
2022	locations	W. Asia	0.010
2022	locations	S.-E. Asia	0.009
2022	locations	Melanesia	0.000
2022	locations	Micronesia	0.000
2022	locations	Caribbean	0.003
2022	locations	E. Europe	0.023
2022	locations	N. Europe	0.114
2022	locations	S. Europe	0.021
2022	locations	W. Europe	0.062
2022	locations	UNK	0.258
2022	locations	Unknown	0.179

continue in the next column

Table 1: This is the initial distribution of the test document across categories and groups. This distribution is the result of harvesting the data included in the test collection. It is worth noting that the proportions differ by one or even two orders of magnitude. LAC refers to Latin America and the Caribbean, and ANZ to Australia and New Zealand.

testing. At training time, the participants design and train their own system configurations by using the full, past data. At testing time, the participants submit the runs by using a new set of topics without knowledge of the relevance judgments, though.

The test collections show that *the distribution of documents appears as biased towards males, Europe, and Northern America (NA).* Let C be a subset of documents of a certain category. Document harvest creates the initial distribution $\Pr(C)$ reported in Table 1. Besides the fact that the test collections are biased, the order of magnitude of $\Pr(C)$ of males, NA, and European location documents influences the subsequent phases. Note that the unknown locations and genders take a quite significant portion of the collection. In principle, they should not be considered as a category because the unknown category documents cannot contribute to the calculation of the index of mutability. The values of the Gini index reported in this paper consider the proportion of unknonws—the purified values of the proportion of unknowns can be obtained with the following expression:

$$1 - \frac{\left(\sum_{C \in \mathcal{C} \setminus \{U\}} \Pr(C|X)^2\right)^2}{(1 - \Pr(C|U))^2} - \Pr(C|U)^2 \qquad \Pr(C|U) < 1$$

where U is the subset of items of the unknown group.

After document retrieval, *the distribution of retrieved documents remains as biased towards males, Europe, and NA*, thus implying that the state-of-the-art search algorithms reproduce the bias observed after collections are harvested. Let B be a subset of documents retrieved to match q. Retrieval decomposes $\Pr(C)$ into the $\Pr(C|B)$'s in Table 2. Retrieval creates document sets in which the distribution by category differs a little from the distribution $\Pr(C)$. In particular, the exposure of Africa and unknown groups has been made a little

Track	Phase	Category	Group	Pr($C\|B$)
2021	train	locations	Africa	0.017
2021	train	locations	Asia	0.087
2021	train	locations	Europe	0.09
2021	train	locations	LAC	0.023
2021	train	locations	N. America	0.192
2021	train	locations	Oceania	0.024
2021	train	locations	Unknown	0.567
2021	test	gender	Female	0.005
2021	test	gender	Male	0.017
2021	test	gender	Unknown	0.978
2021	test	locations	Africa	0.003
2021	test	locations	Asia	0.045
2021	test	locations	Europe	0.134
2021	test	locations	LAC	0.004
2021	test	locations	N. America	0.162
2021	test	locations	Oceania	0.024
2021	test	locations	Unknown	0.627
2022	train	gender	Man	0.065
2022	train	gender	Unknown	0.926
2022	train	gender	Woman	0.01
2022	train	locations	ANZ	0.028
2022	train	locations	Caribbean	0.003
2022	train	locations	C. America	0.002
2022	train	locations	C. Asia	0.001
2022	train	locations	E. Africa	0.0
2022	train	locations	E. Asia	0.022
2022	train	locations	E. Europe	0.02
2022	train	locations	Melanesia	0.0
2022	train	locations	Micronesia	0.0
2022	train	locations	M. Africa	0.0
2022	train	locations	N. Africa	0.0
2022	train	locations	N. America	0.281
2022	train	locations	N. Europe	0.12
2022	train	locations	Polynesia	0.0
2022	train	locations	S. America	0.004
2022	train	locations	S.-E. Asia	0.009
2022	train	locations	S. Africa	0.002
2022	train	locations	S. Asia	0.012
2022	train	locations	S. Europe	0.013
2022	train	locations	UNK	0.26
2022	train	locations	Unknown	0.155
2022	train	locations	W. Africa	0.0
2022	train	locations	W. Asia	0.009
2022	train	locations	W. Europe	0.058
2022	test	gender	Man	0.025
2022	test	gender	Unknown	0.97
2022	test	gender	Woman	0.004
2022	test	locations	ANZ	0.027
2022	test	locations	Caribbean	0.003
2022	test	locations	C. America	0.002
2022	test	locations	C. Asia	0.001
2022	test	locations	E. Africa	0.001
2022	test	locations	E. Asia	0.016
2022	test	locations	E. Europe	0.011
2022	test	locations	Melanesia	0.0
2022	test	locations	Micronesia	0.0
2022	test	locations	N. Africa	0.0
2022	test	locations	N. America	0.227
2022	test	locations	N. Europe	0.109
2022	test	locations	S. America	0.003
2022	test	locations	S.-E. Asia	0.01
2022	test	locations	S. Africa	0.002
2022	test	locations	S. Asia	0.008
2022	test	locations	S. Europe	0.01
2022	test	locations	UNK	0.22
2022	test	locations	Unknown	0.302
2022	test	locations	W. Africa	0.001
2022	test	locations	W. Asia	0.009
2022	test	locations	W. Europe	0.038

continue in the next column

Table 2: This is the breakdown of retrieved documents by category and group. Each topic's top 1,000 ranked documents were retrieved. Because some topic sets lack another field to extract topic terms, the documents were retrieved to answer questions based solely on the topic title. The sum of document scores was calculated by category. Because the top 1,000 ranked documents produce the highest scores, the reported proportions can be regarded as a reliable estimate of the proportion obtained if the sum of all document scores were calculated. LAC refers to Latin America and the Caribbean, and ANZ to Australia and New Zealand.

larger than before retrieval against the exposure of the other categories. However, males, NA, and Europe still remain the majority because of the prior probability determined at collection harvest time. Note that Table 2 makes a distinction between "train" and "test" because retrieval took place at two different phases during the TREC Fair tracks. Table 3 summarizes Table 2 by using the Gini index.

After document ranking, *the distribution of retrieved documents remains as biased, and the Gini index of mutability does not increase or significantly decrease.* Let D be a subset of documents retrieved at certain ranks, such as the retrieved documents ranked between

Track	Phase	Category	$G(C\|B)$
2021	train	locations	0.729
2021	test	gender	0.065
2021	test	locations	0.653
2022	train	gender	0.207
2022	train	locations	0.845
2022	test	gender	0.088
2022	test	locations	0.832

Table 3: This table reports the Gini index of mutability with respect to each category measured for the retrieved documents from each subcollection.

the first and the tenth position of the first retrieved page. [2] Ranking allocates the retrieved documents into ranks at which the distributions $\Pr(C|B, D)$ by category and pages may differ from $\Pr(C|B)$. Figure 3 of Appendix A depicts the Gini indexes of the 100-document rankings segmented in 10-document pages such that page p includes the retrieved documents ranked from $10(p-1)+1$ to $10p$. Ten-document pages are common in many applications.

After relevance assessment, *the trade-off between fairness and effectiveness arises.* Let A be a subset of documents relevant to q. Assessment decomposes $\Pr(C|B, D)$ into the $\Pr(C|A, B, D)$'s. Figure 4 of Appendix A depicts the proportion of relevant documents that are ranked in each page, the latter being called "precision at page p" whereas Figure 3 of Appendix A depicts the Gini index. The heatmaps depicting the distribution of accuracy across the 10 document pages reveal that the top pages have the highest precision values and the lowest Gini index of mutability. However, the trade-off between accuracy and Gini index is less pronounced in the location category than in the gender category. Note that, the heatmap was necessary because of δ, which is absent from the recovery effectiveness measurement. Regarding the latter, a one-dimensional heatmap was used instead of other chart types because the x-axis represents the page, just like with equity heatmaps.

As a summary, a document page on average includes three to four NA documents, three European documents, one to two Asian documents, and the remaining documents are from the other locations. In parallel, five or six documents refer to males, whereas other genders are basically absent from the ranking pages. Ranking confirms the unbalanced distribution of documents by categories; for example, the top ten ranked documents on average are four North American documents, three European documents, and the remaining three are distributed over the other four locations. After assessment, the likelihood of relevance results as unbalanced over the locations. NA and Europe are the most likely locations of relevant documents in the top-ranking pages. The retrieval phase little reduces the imbalance against some unprotected categories.

4 Eigensystem for Topic Term Weighting

Given a document collection and the indices thereof, a way to achieve fair rankings is allowing the system to re-rank the retrieved documents. Re-ranking retrieved documents has the downside of reducing effectiveness, since the top-ranked relevant documents might be moved away from the top ranks. An alternative approach—which is adopted in this

[2] In IR, Search Engine Retrieved Page (SERP) means the dynamic webpage listing the links to the retrieved documents.

paper—is to incorporate fairness in the ranking function at the retrieval time. The way is paved by the topic term weights.

The selection of the topic terms and the weights thereof can, in theory, impact ranking and, as a consequence, category exposure. The topic terms' impact on ranking by means of the number of terms, the document term weights, and the topic term weights. The number of terms and the document term weights depend on the test collection and on the weighting scheme adopted by a retrieval system—therefore, they can hardly be changed to accomplish the goal of fair ranking. A simple example, which might even seem trivial, arises when the documents of a privileged category are longer than those of a disadvantaged category. Since term weights increase with the length of the documents in which they appear, those from the privileged category will be shown before those from the disadvantaged category.

However, a system could in principle adjust topic term weights if some information about category exposure and distribution were available—in this paper, the topic term weights are automatically adjusted to promote the retrieved documents of unprotected groups by minimizing the impact on effectiveness. The retrieval document score can be utilized to this end in combination with category-document distribution, since document scores are sums of products between document term weights and the corresponding topic term weights. Accordingly, the document term weights remain unchanged, whereas the topic term weights can be adjusted to incorporate some information about category exposure, thus making protected categories exposed as unprotected ones.

Topic term weights can be adjusted to change the score of the documents matching the topic term. In particular, the weight of a topic term can be increased if the term occurs in documents of protected categories more frequently than in documents of unprotected categories. The scores of protected documents can thus be increased to take fairness into account. If topic term weights were adjusted to take fairness into account, the ranking would also be adjusted to increase the exposure of the documents of unprotected categories.

```
for all topic $q$ do
    $e \leftarrow 1^k$ -- unary representation of $q$
    $y_e \leftarrow Be$ -- rank documents matching $e$
    $B \leftarrow$ term-document occurrence matrix
    $Q \leftarrow$ equation (4)
    $x \leftarrow$ main eigenvector of $B'B - B'QB$
    $y \leftarrow Bx$ -- rank documents matching $x$
    $C \leftarrow$ retrieved document-group matrix
    compute fairness and effectiveness measures
end for
```

Figure 1: The computational approach of Eigensystem for Topic Term Weighting (ET) is summarized in this pseudo code.

ET aims to maximize both fairness and effectiveness by automatically reweighting topic terms and re-ranking documents that are retrieved to answer the topic of which terms are equally weighted. Consider Eq. (2): ET finds the function called x in a principled way. To this end, a k-dimensional vector space over the real field represents k topic terms. The main eigenvectors of a symmetric matrix provide an alternative k-dimensional vector to represent topics. Moreover, the main eigenvectors of the symmetric matrix maximize a function measuring effectiveness and fairness. The function takes a quadratic form whose matrix is the parameter. The computational approach is summarized in Figure 1.

In mathematical terms, consider $n \in \mathbb{N}$ retrieved documents, $k \in \mathbb{N}$ topic terms, $m \in \mathbb{N}, m > 1$ categories, and $r \in \mathbb{N}, r > 1$ relevance grades. Let $B \in \mathbb{R}^{n\times k}$ be the term-to-document matrix where a B's element is the weight of a term in a document, i.e., $w(t, d)$. Let $C \in \mathbb{R}^{m\times n}$ be the documents-to-categories matrix where an element is the degree of membership of a document to a category. Suppose the C's column is non-negative and sums to 1. Let CB be the terms-to-categories matrix, which is in $\mathbb{R}^{m\times k}$. Let

$$x \in \mathbb{R}^{k\times 1}$$

be the topic term weight vector such that

$$\|x\| = 1 \ .$$

Given a topic and a set of retrieved documents, let Bx be the retrieved document scores, which are in $\mathbb{R}^{n\times 1}$. Let

$$y = Bx$$

be the retrieved document normalized scores. Without loss of generality, y can be assumed as non-negative. Using L_1 normalization, let

$$y = Bx \,/\, 1'Bx$$

where $1 \in \{1\}^{n\times 1}$. Therefore,

$$y \geq 0 \qquad 1'y = 1 \ .$$

Using L_1 normalization, suppose the rows of C are normalized. The sum of scores of the retrieved documents of each category is given by

$$p = Cy$$

where $p \in [0, 1]^{m\times 1}$ and sums to 1, i.e., p is a probability distribution, since the C's columns are non-negative and sum to 1 as y does. The Gini index of mutability becomes

$$G = 1 - \sum_{j=1}^{m} p_j^2 \ .$$

The index ranges between 0 and $1 - 1/m$. Let's replace p with Cy, and let c_j be the j-row of C:

$$p_j^2 = (c_j y)(c_j y) = (c_j y)'(c_j y) = (y' c_j')(c_j y) = y' c_j' c_j y \ .$$

Therefore,

$$\sum_{j=1}^{m} p_j^2 = \sum_{j=1}^{m} y' c_j' c_j y = y' \left(\sum_{j=1}^{m} c_j' c_j \right) y = y' Q y$$

where

$$Q = \sum_{j=1}^{m} c_j' c_j \tag{4}$$

As $y = Bx$, the Gini index becomes a function of x defined as

$$G(x) = 1 - x'B'QBx \ .$$

The latter is a symmetric quadratic form since $B'QB$ is symmetric. Moreover, G is a continuous function in a closed interval; therefore, a maximal point does exist according to Weierstrass' theorem. The maximum of G is achieved when x is one of the main eigenvectors of $I - B'QB$. As a consequence, a ranking of n retrieved documents exhibits the highest fairness if the topic terms are weighted according to x and then is obtained by Bx. Let

$$F = y'y$$

be the sum of squared retrieved document scores. As $y = Bx$,

$$F = F(x) \qquad F(x) = x'B'Bx$$

can be viewed as a measure of the likelihood that an n-document ranking represents relevant information to the topic. The main eigenvector of $B'B$ determines the n-document ranking for which F is maximum. If y is a measure of probability of relevance of the n items, the maximal value of F is a necessary condition of the Probability Ranking Principle (PRP), which has been used, in one form or another, by various people since M. E. Maron and J. L. Kuhns. W. S. Cooper gave a formal statement of the principle as reported by S. E. Robertson; see Robertson (1977).

> If a reference retrieval system's response to each request is a ranking of the documents in the collections in order of decreasing probability of usefulness to the user who submitted the request, where the probabilities are estimated as accurately as possible on the basis of whatever data has been made available to the system for this purpose, then the overall effectiveness of the system to its users will be the best that is obtainable on the basis of that data.

Apart from the ranking by non-decreasing probability of usefulness or relevance, the optimal probability estimation is one key concept of the principle. The estimation of the probability that the i-th item is relevant or useful is provided by (1) or (2), depending on whether the topic term weights are subject to estimation or not, respectively.

Whereas $w(t, d)$ can be referred to as relevance and then to effectiveness, $x(t, q)$ can be leveraged to integrate fairness and effectiveness when ranking documents. The potential provided by x is then what the PRP requires to make a ranking optimal, since x can integrate w, thus providing a—hopefully—better estimation of the probability of usefulness to the user.

The objective of a fair ranking is to maximize both effectiveness and fairness; therefore, $F + G$ should be maximized, i.e., both F and G should be. As both F and G are functions of x, then $F + G$ is also a function of x, i.e.,

$$F + G = (F + G)(x) \ .$$

Therefore,

$$(F + G)(x) = F(x) + G(x) = x'B'Bx + 1 - x'B'QBx = x'(B'B + I - B'QB)x \ .$$

It follows that the maximal point of $F + G$ is one of the main eigenvectors of

$$B'B + I - B'QB \; .$$

Therefore, finding the maximum point of $(F + G)(x)$ means finding the maximum point of $x'(B'B + I - B'QB)x$ where I is the $k \times k$ identity matrix. The latter is a quadratic form in the operator $B'B + I - B'QB$ whose maximum point is one out of the main eigenvectors. For any z such that $z'z = 1$, we have that

$$z'(B'B + I - B'QB)z = z'(B'B - B'QB)z + 1 \; .$$

The latter reaches its maximum value if z is one out of the main eigenvectors called Eigensystem for Topic Term Weightings (ETs) of $B'B - B'QB$.

Actually, any vector of the subspace spanned by the ETs are maximal points of $F + G$. Indeed, a quadratic form may in general have more than one maximal point corresponding to the ETs, since multiple eigenvectors may share the same largest eigenvalue. In particular, there might be more than one ETs because $B'B - B'QB$ might not be primitive, and one cannot leverage Perron and Frobenius' theorem stating that a primitive matrix has one main eigenvector; see Seneta (2006).

The multiplicity of the ETs is of little importance, though, since one eigenvector out of them is sufficient to maximize $F + G$. Nevertheless, the ETs are still mutually orthogonal, and therefore they represent alternative topic term reweighting schemes. Finally, note that, as the scale of the eigenvalues of $B'B$ may largely differ from those of $B'QB$, the matrices $B'B$ and $B'QB$ should be divided by the respective main eigenvalue.

5 Experiments

5.1 Data and Measures

The test collections from the TREC Fair tracks in 2021 and 2022 were used to evaluate the algorithms. The experimental data were obtained from Wikimedia[3]; see Ekstrand et al. (2022a, 2023). The public availability of the experimental collections makes the utilization for an empirical study such as that reported in this section acceptable despite the unavoidable limitations of any dataset, as explained by Ekstrand et al. (2023). Apart from the experimental documents, the test collections also comprise a series of experimental inquiries, the metadata (information about category membership), and evaluations of the documents' pertinence. Table 4 reports on the test collections' main statistics.

	2021	2022
Number of documents	6,280,328	6,475,401
Average document size (bytes)	6,748	3,197
Number of train topics	47	46
Number of test topics	22	20
Average number of train topic words	5.38	61.3
Average number of test topic words	2.72	3.35

Table 4: The main statistics of the test collections

[3] https://en.wikipedia.org/wiki/WikiProject

The 2021 and 2022 Fair Ranking Tracks were an ad hoc retrieval methodology where participants were given a collection of documents and a series of topics to rank documents based on relevance to a specific WikiProject.[4] The challenges shared a topic set, corpus, fundamental problem structure, and fairness aim. The topics were prepared from WikiProject subjects, documents being English Wikipedia articles, rankings being lists of articles that editors may consider relevant to the topics—NIST assessors annotated the retrieved documents with binary relevance scores for given topics—and articles being fairly exposed based on geographical location and gender.

As for the 2021 track, the categories were defined by gender and geographic locations, the latter being used for training. Within TREC, "training" refers to the experimental phase for tuning the participating systems and precedes "test," which is the actual measurement of the effectiveness of the systems for the reference year, test data, and gender, which was only used for the test data.

The 2022 Fair Ranking Track required the participants to deal with a larger number of categories other than geographic location and gender. As for the 2022 track, the categories were defined by geographic locations, gender, age of the topic, occupation, alphabetical order of name, age of the article, page popularity, and replication of articles in other languages as detailed by Ekstrand et al. (2022b). In order to compare the tracks, the experiments were only carried out with geographic locations and genders for the 2022 track as well.

The experiments were carried out with the following input:

- the experimental topics completed with the relevance assessments,
- the experimental document identifiers completed with category memberships, i.e.:
  - gender-based category with male, female, or non-binary, also known as "third";
  - geographical location-based with different continents or subcontinent regions;
- the experimental full-text documents.
- the document retrieval scores returned by an experimental search system to compute A on the basis of both: (Elasticsearch v 7.17.1 and the standard configuration thereof were utilized.)
  - either title-only topics or full topics, and
  - either title-only document or full document;

For each topic, the list of relevant documents was retrieved from the test collection. A term-based representation of the topic was built by concatenating the words found in the topic fields, such as title and keywords. If there were at least two topic terms, a list of retrieved documents was ranked by (1). The $n = 100$ documents retrieved for each topic were ranked by Fielded Best Match 25 (BM25F); as said, ten-document pages are common in many applications, and Figure 3 of Appendix A can still be readable if ten pages are observed. Each document was scored by

$$\sum_{t \in d \cap q} \mathrm{BM25F}(t, d)$$

[4] https://en.wikipedia.org/wiki/WikiProject

where

$$\mathrm{BM25F}_d = \sum_{\ell=1}^{L} \sum_{t=1}^{k} \mathrm{IDF}_t \mathrm{SAT}_{d,t}$$

and

$$\mathrm{IDF}_t = \log 1 + \frac{N - n_t + \frac{1}{2}}{n_t + \frac{1}{2}} \qquad \mathrm{SAT}_{d,t} = \frac{f_{d,t}(k_1 + 1)}{f_{d,t} + k_1 \left(1 - b + b\frac{s_{d,\ell}}{s_\ell}\right)}$$

and ℓ refers to a field such as title or abstract, $f_{d,t}$ is the frequency of t in d, n_t is the number of documents indexed by t, $s_{d,\ell}$ is the size of ℓ of d, $k_1 = 1.2$ and $b = 0.75$ are hyperparameters, IDF stands for Inverse Document Frequency, SAT stands for "saturation," and BM25F stands for Fielded Best Match 25. Note that the topic term weight is assumed to be 1, thus not rescaling BM25F.

In parallel, an efficient implementation of B and C was obtained by in-memory, sparse matrices. The ranking obtained for the topic was browsed, and the Compressed Sparse Row (CSR)-based representation of each matrix was allocated. Using the CSR-based representation, the ℓ-th matrix non-zero entry v of row i and column j was allocated into three array entries as follows:

$$\mathrm{value}[\ell] = v, \mathrm{row}[\ell] = i, \mathrm{column}[\ell] = j \ .$$

In this way, the CSR-based matrices for A, B, and C were allocated where $n = 100$, k topic terms, and m categories.

The eigensystem was then computed, thus obtaining an $\mathrm{ET}(q)$ of a given topic. To be specific, using the ad-hoc subroutines for the CSR-based representations, the products $R = B'B$ and $S = B'QB$ were computed. Both R and S were divided by the respective main eigenvalue to make their scale neutral. Each document was then rescored by

$$\sum_{t \in d \cap q} \mathrm{BM25F}(t, d)\mathrm{ET}(t, q) \ .$$

where $\mathrm{ET}(t, q)$ is the eigenvector component of a term of the topic that determined the ranking. Note that the topic term weight is $\mathrm{ET}(t, q)$, thus rescaling BM25F according to the ET.

For each $\delta \in \mathbb{R}$ selected from a predefined set of values that was decided at the experimentation time, the eigensystem of $Q = R - \delta\, S$ was computed. The Q's main eigenvector, x, was multiplied by B, thus obtaining the n-dimensional vector $y = Bx$ of document scores recomputed by using non-unitary topic term weights. In this way, this ranking was compared with the ranking obtained by Be, where $e \in 1^k$ is the n-dimensional vector y_e of document scores recomputed by using unitary topic term weights. Note that the entries of the main eigenvector of Q might be non-positive. As a consequence, the y might not be non-negative. To work around the issue of negativity of Q, G was obtained as follows: First, y was translated to $y + \min y$ and then L_1-normalized, thus obtaining a non-negative, normalized retrieval score vector. Then, Q was computed.

The following measures were computed for each topic, i.e., for each baseline ranking and the re-rankings thereof: [5]

[5] A re-ranking is a ranking of documents rescored by an ET.

- Gini's index of mutability (G),
- Precision at 10-document page (P), and
- a Global Measure, i.e., $M = G \times P$.

The latter function is non-decreasing for both P and G—it aims at measuring the extent to which reranking documents may not decrease if not even increase effectiveness or fairness. The sum of G and P is not appropriate because they are derived from different measurement units and phenomena. Instead, there are some justifications for multiplying different quantities, such as those used in Macroeconomics. Other fairness measures are occasionally used, such as Attention-Weighted Rank Fairness (AWRF); see Raj and Ekstrand (2022). Other measures of effectiveness, such as NDCG and Average Precision (AP), are also available. The former was defined for non-binary relevance grades, which are not applicable in our case. The latter considers relevant document ranks—we used it in the previous work; see Melucci (2024a,b, 2025a,b). We preferred to investigate the same issues from a different perspective and then use different measures. AWRF assumes an attention model, but we wanted to avoid additional assumptions and take an assumption-free approach. A comparison, albeit indirect, can be made by observing the results in Melucci (2025b).

5.2 Main Results

Overall, ET can yield fairer and more effective rankings than the baseline depending on δ and on category. Figures 2 and 5(a) of Appendix A depict the baseline values of M and the values of M for each δ and page and both tracks as a whole. As a summary, moderate to high delta values (1.00–1.75) yield the most improvement in the global measure under gender bias correction. Moreover, peak gains are concentrated around the middle-ranking pages (5–7). Finally, too high or too low delta settings tend to be less effective, especially on the tails of the ranking. As for location, bias mitigation for location yields strong and consistent improvements, especially from delta 1.25 onward. In addition, the ranking system appears to benefit more clearly from fairness adjustments for location than for gender, likely due to higher initial disparity. Pages 5 and 6 are again the sweet spot for fairness or utility gains. In order to understand what may affect the global measure, the results have been reported by measure, i.e., G and P. Both measures increase as δ does, and in particular the mid-ranked pages show the largest values. Both measures cannot be increased at the top-ranked pages, where effectiveness and fairness remain relatively low. Figures 5(b) and 5(c) of Appendix A depict the values of G, P for each δ and page and both tracks as a whole.

Figures 6 and 7 of Appendix A show that *ET can improve fairness to varying degrees in most cases.* The variations of the Gini index of mutability clearly depend on the fairness, i.e., bias against some groups caused by the harvest and retrieval phases. This bias can lead to unequal outcomes, which the ET method seeks to address by redistributing weights more equitably. By analyzing the Gini index, we can better understand the impact of these adjustments on fairness across different demographic groups. For example, gender appears to be the most biased category, and as a result, the little variations of G achieved by ET should be considered as quite large. In contrast, categories with relatively high Gini index values can become more equitable than the baseline to varying degrees. This

track	phase	category	δ	training M	testing M
2021	test	gender	1.25	0.046	0.039
2021	test	locations	0.0	0.466	0.453
2021	train	locations	0.25	0.461	0.458
2022	test	gender	1.25	0.156	0.100
2022	test	locations	1.5	0.467	0.465
2022	train	gender	1.25	0.322	0.287
2022	train	locations	2.0	0.528	0.524

Table 5: The maximum δ for each track, phase, and category was determined for the training topic set and used to calculate $M = G \times P$ for the test set. Except in one case, using $\delta > 1$ enhances effectiveness and fairness.

shift towards greater fairness can lead to improved outcomes for underrepresented groups, fostering inclusivity and fairness in decision-making processes. By concentrating on these adjustments, a more equitable environment that serves the interests of all parties involved can be established.

As shown in Figures 6 and 8 of Appendix A, *ET may decrease efficacy for most cases*, even for only the top pages, whereas effectiveness can enhance for the mid or bottom pages. Heatmaps illustrate that precision at top pages decreases as δ grows. As δ grows, the decrease in precision slows or even becomes an increment when browsing from top to bottom pages. This shows that, while the top pages may lose precision, the middle and bottom pages may benefit from changes to the browsing parameters. As a result, this variety emphasizes the necessity for personalized solutions that improve user experience across various levels of content. This necessity for personalized solutions highlights the importance of adaptive algorithms that can cater to individual user preferences and behaviors.

Considering known groups only slightly changes the overall results. In Section 3, we discussed the significant proportion of papers whose group is known and hypothesized that this proportion may affect the metrics of fairness. The issue is similar to the lack of relevance judgments, making it difficult to develop effective relevance feedback algorithms in IR. Despite a significant number of unknown-group documents, we found that the overall efficacy of ET is barely affected. Figure 9 of Appendix A depicts heatmaps of the results obtained by omitting these documents. The consistent performance of ET suggests that the presence of unknown-group documents does not significantly compromise the fairness metrics we are investigating. Therefore, the conclusion should not change. This finding demonstrates the robustness of our approach in handling varying levels of group knowledge within the dataset.

Topic term weighting based on ET and $\delta > 0$ is a good decision in terms of global measure, to some extent depending on the track, phase, and category. Cross-validation has been used to determine the best *delta* for each track, phase, and category. To accomplish this, the topic set was divided into training and test sets using the $3/4, 1/4$ rule. The training set included 75% of the topics, while the test set included the remaining 25%. The training set was selected at random and repeated twenty times. The fairness and effectiveness measures, G and P, were calculated for each page and δ. The global measure was then computed and averaged across the pages to produce a global measure for ranking purposes. Table 5 shows that δ should be greater than 1 to achieve the best performance in terms of M, except for one case.

δ	Fairness: 2021 train locations										Effectiveness: 2021 train locations									
	1	2	3	4	5	6	7	8	9	10	1	2	3	4	5	6	7	8	9	10
.25	.02		.01			.04		.04	.03	.02	.0	.0							.01	.03
.50	.04		.04		.04				.02		.0									
.75					.04			.02	.04	.01	.0	.01							.03	
1.00	.02				.05				.0		.0		.04		.02			.04		
1.25									.01	.0	.0				.04					
1.50	.04		.03								.0	.01			.01				.02	
1.75	.02							.02	.0	.0	.0	.01			.0				.01	
2.00	.02		.04					.0	.0	.0	.0	.0			.01			.03	.01	

Table 6: The table reports the p-values for each δ and page for the year 2021, the training phase, and the category "geographical location". The blank cells indicate non-significance. If a cell is not blank then the number is the p-value observed–the lower the p-value the more significant the variation of either G or P caused by δ with respect to the baseline for the given track, phase, and category.

The darker areas of the heatmaps do *not always correspond to significant variations in fairness or effectiveness.* To provide a measure of the statistical significance of the results, the Wilcoxon signed-rank test was conducted, and only the cases where the p-value was less than or equal to 0.05 were considered significant. The Gini index and precision were observed for each track, phase, category, topic, and ten-document page for $\delta = 0$ and then for each $\delta > 0$ as mentioned previously. So, a page of ten documents obtained with $\delta = 0$ was compared to the corresponding one with a certain $\delta > 0$. See Table 6.

The original bias against a category may impact on the significance of the variations and as a consequence fairness can hardly be improved if the initial conditions as rather unfavourable. The following table shows that the significant variations are less frequently significant with gender than in the case of locations:

Fairness: 2021 test gender										
δ	1	2	3	4	5	6	7	8	9	10
0.25			0.04		0.04					
0.50										
0.75			0.04		0.03					
1.00					0.02					
1.25					0.02					
1.50										
1.75										
2.00										

Effectiveness: 2021 test gender										
δ	1	2	3	4	5	6	7	8	9	10
0.25										
0.5										
0.75										
1.00	0.0						0.02			
1.25	0.0					0.05	0.05		0.02	
1.5	0.0						0.05		0.02	
1.75	0.0						0.05		0.02	
2.0	0.0						0.05		0.02	

As for the 2021 test set and the location category and the 2022 sets, a very few differences were statistically significant as reported in Appendix B.

The experimental results presented in this section are expressed using two distinct measures of effectiveness and fairness, namely G and P, which are then summarized into an overall measure, $M = G \times P$. This is not the only possible suite of measures. In conclusion to this section, the results obtained with the measurement proposed recently by Sakai et al. (2023) are presented in order to provide a comprehensive overview and, hopefully, make them comparable with other experiments. In the work by Sakai et al. (2023), a measure called the Global Fairness and Relevance (GFR) is introduced, defined as follows, and which is reproduced in this work with the same notation as the authors of this measure:

$$\mathrm{GFR}(L) = w_0\mathrm{RELEVANCE}(L) + w_1\mathrm{GF}(L) \tag{5}$$

where RELEVANCE(L) is a measure of the utility of the ranked list L and Global Fairness (GF) measures the user-perceived difference of fairness between a "fair" ranking and the actual ranking. In particular,

$$\mathrm{GF}(L) = \sum_{k=1}^{|L|} \mathrm{DECAY}(L,k)\mathrm{DISTRSIM}(L,k)$$

where DECAY measures the event a user accesses the k-th item in L and DISTRSIM how far the distribution of groups is from an ideal, uniform distribution. Expression (5) has been computed for each topic of every track, phase, category, and δ, and for each of the following methods implementing the three functions occurring in (5):

- DECAY: Expected Reciprocal Rank (ERR) and Ranked Biased Precision (RBP),
- DISTRSIM was defined as

  $$1 - \mathrm{DIVERGENCE}(p||p^*)$$

  where p is the observed group distribution and p^* is the fair group distribution and DIVERGENCE was either Jensen-Shannon Divergence (), Root Normalised Order-aware Divergence (RNOD), or Normalised Match Distance (NMD).[6]

We were interested to the trend of GFR as δ increases to check whether GFR tends to increase, decrease or keep stable if the importance to the fairness operator in the calculation of the eigentopics increases. The average of the GFR values obtained for each topic was then calculated for each track, phase, category, and δ and classified by the aforementioned methods used to implement the three functions of (5). Table 7 summarizes the results where the average and the standard deviation values of GFR have been computed over all the possible combinations of the methods used for decay and distribution similarity.

Table 8 breaks the GFR values down the decay and distribution similarity methods. Overall, GFR makes an improvement of fairness *and* effectiveness visible for different decay and distribution similarity methods. The variations of GFR while δ increases depend on the variations of the measures of effectiveness and fairness. For example, as for the ERR-based decay method the variations of precision and fairness are reported below on the left and on the right, respectively:

δ	All	2021	2022	Gndr	Locat	Train	Test	All	2021	2022	Gndr	Locat	Train	Test
0.0	0.495	0.386	0.576	0.282	0.507	0.608	0.41	0.162	0.135	0.181	0.039	0.249	0.19	0.14
0.25	0.481	0.365	0.568	0.272	0.491	0.571	0.414	0.19	0.167	0.208	0.05	0.27	0.26	0.138
0.5	0.48	0.363	0.567	0.265	0.494	0.57	0.412	0.191	0.169	0.207	0.049	0.269	0.265	0.135
0.75	0.468	0.346	0.56	0.257	0.477	0.566	0.395	0.185	0.171	0.197	0.05	0.273	0.247	0.139
1.0	0.442	0.294	0.553	0.17	0.464	0.574	0.343	0.215	0.232	0.203	0.121	0.311	0.243	0.195
1.25	0.432	0.286	0.543	0.152	0.449	0.568	0.331	0.209	0.223	0.198	0.101	0.305	0.239	0.186
1.5	0.433	0.28	0.548	0.15	0.45	0.567	0.333	0.212	0.221	0.206	0.101	0.312	0.24	0.192
1.75	0.434	0.278	0.552	0.15	0.45	0.569	0.333	0.212	0.221	0.206	0.101	0.313	0.24	0.192
2.0	0.431	0.278	0.547	0.15	0.449	0.561	0.334	0.209	0.219	0.201	0.101	0.305	0.241	0.185

[6] The details have been described by Sakai et al. (2023) in their paper.

δ	All	2021	2022	Gndr	Locat	Train	Test
0.0	0.33 ± 0.10	0.27 ± 0.08	0.38 ± 0.01	0.15 ± 0.08	0.39	0.41 ± 0.06	0.28 ± 0.10
0.25	0.34 ± 0.10	0.27 ± 0.07	0.39 ± 0.01	0.15 ± 0.08	0.39	0.42 ± 0.04	0.28 ± 0.10
0.5	0.34 ± 0.11	0.27 ± 0.07	0.39 ± 0.01	0.15 ± 0.09	0.39	0.42 ± 0.05	0.28 ± 0.10
0.75	0.33 ± 0.11	0.26 ± 0.08	0.38 ± 0.01	0.15 ± 0.09	0.38	0.41 ± 0.05	0.27 ± 0.10
1.0	0.33 ± 0.12	0.26 ± 0.09	0.38 ± 0.02	0.13 ± 0.10	0.39	0.41 ± 0.06	0.26 ± 0.11
1.25	0.32 ± 0.13	0.25 ± 0.08	0.37 ± 0.01	0.11 ± 0.10	0.38	0.41 ± 0.05	0.25 ± 0.11
1.5	0.32 ± 0.12	0.25 ± 0.08	0.38 ± 0.01	0.11 ± 0.10	0.38	0.41 ± 0.05	0.26 ± 0.12
1.75	0.32 ± 0.12	0.24 ± 0.08	0.38 ± 0.01	0.11 ± 0.10	0.38	0.41 ± 0.06	0.26 ± 0.12
2.0	0.32 ± 0.12	0.24 ± 0.08	0.38 ± 0.01	0.11 ± 0.10	0.38	0.40 ± 0.06	0.26 ± 0.12

Table 7: The average GFR ± the standard deviations thereof have been computed over the mean GFRs obtained for each decay or distribution similarity method and classified by track, phase, and category. These averages measures the scale for each track, phase, and category. The standard deviations suggest that there is variability from a method to another, being the methods used for gender showing less variations between methods than the other criteria.

The variations of GFR then depend on the weights assigned to precision and GF in Expression (5). The need of deciding the weights can be considered the sympton of a GFR's weakness. First, the sum of two different measures can often be problematic if these quantities measure different phenomena because of different measure units, for instance. Second, the indications provided by GFR strongly depend on weights, thus asking a question left to the future work as Sakai et al. (2023) admit.

6 Final Remarks

Group fairness is considered in this paper, but there is the question of whose fairness we are talking about. In other circumstances, fairness may be defined in terms of what is displayed (e.g., job ads or products). In such a case, search result diversity would be an acceptable phrase if the items were not affiliated with real organizations who profit from exposure. To be precise, fairness should be defined in terms of the producers or authors, particularly in systems that serve multiple stakeholders. There is a third possibility as well: that fairness pertains to subjects of content, such as people mentioned in news articles or search results, the latter being best termed as a privacy issue. Finally, topics might be pertinent to fairness: some topics might not be acceptable in some contexts, whereas they are in other contexts. In sum, the multi-stakeholder nature of IR further complicates fairness definitions.

So where does this leave us? First, we must recognize, at the risk of looking generic, that fairness is not a property of an algorithm alone, but of a system in context. Despite the trend to consider the distinct phases of search, a fair system is a product of design choices, data distributions, user behaviors, and societal norms. No amount of technical tinkering will yield a "perfectly fair" system if the underlying assumptions are flawed or the inputs are biased. Second, fairness cannot be divorced from trade-offs. We must make explicit our priorities—between relevance and (fair) representation, between user satisfaction and (fair) societal impact. Third, evaluation becomes a central challenge. Unlike relevance,

(a) ERR

δ	All	2021	2022	Gndr	Locat	Train	Test
0.0	0.328	0.261	0.379	0.161	0.378	0.399	0.275
0.25	0.336	0.266	0.388	0.161	0.380	0.416	0.276
0.5	0.335	0.266	0.387	0.157	0.381	0.417	0.273
0.75	0.327	0.259	0.378	0.154	0.375	0.407	0.267
1.0	0.329	0.263	0.378	0.145	0.388	0.409	0.269
1.25	0.321	0.254	0.370	0.126	0.377	0.403	0.258
1.5	0.323	0.250	0.377	0.125	0.381	0.404	0.262
1.75	0.323	0.250	0.378	0.126	0.382	0.404	0.262
2.0	0.320	0.248	0.374	0.125	0.377	0.401	0.259

(b) RBP

All	2021	2022	Gndr	Locat	Train	Test
0.34	0.278	0.387	0.145	0.404	0.416	0.283
0.344	0.273	0.397	0.146	0.399	0.424	0.284
0.342	0.273	0.394	0.141	0.401	0.423	0.281
0.333	0.261	0.387	0.138	0.39	0.413	0.273
0.323	0.247	0.381	0.111	0.391	0.413	0.257
0.32	0.246	0.376	0.103	0.384	0.413	0.251
0.32	0.241	0.379	0.102	0.384	0.409	0.253
0.32	0.239	0.381	0.102	0.384	0.41	0.253
0.318	0.239	0.377	0.102	0.381	0.407	0.251

(c) JSD

δ	All	2021	2022	Gndr	Locat	Train	Test
0.0	0.327	0.261	0.377	0.153	0.379	0.403	0.271
0.25	0.334	0.264	0.387	0.155	0.38	0.411	0.277
0.5	0.329	0.26	0.38	0.151	0.375	0.401	0.274
0.75	0.322	0.256	0.371	0.147	0.369	0.395	0.267
1.0	0.317	0.244	0.371	0.127	0.375	0.402	0.253
1.25	0.309	0.242	0.36	0.115	0.363	0.396	0.245
1.5	0.311	0.24	0.365	0.114	0.366	0.398	0.247
1.75	0.313	0.24	0.367	0.114	0.368	0.4	0.247
2.0	0.311	0.239	0.364	0.114	0.365	0.397	0.246

(d) RNOD

All	2021	2022	Gndr	Locat	Train	Test
0.338	0.285	0.377	0.154	0.396	0.409	0.284
0.343	0.284	0.388	0.154	0.394	0.427	0.28
0.344	0.286	0.388	0.149	0.399	0.434	0.277
0.334	0.272	0.381	0.146	0.389	0.42	0.27
0.332	0.274	0.376	0.132	0.398	0.418	0.269
0.328	0.265	0.375	0.118	0.391	0.417	0.261
0.329	0.262	0.379	0.117	0.394	0.415	0.265
0.329	0.261	0.38	0.117	0.393	0.415	0.264
0.325	0.26	0.375	0.117	0.388	0.411	0.261

(e) NMD

δ	All	2021	2022	Gndr	Locat	Train	Test
0.0	0.338	0.262	0.395	0.152	0.398	0.41	0.283
0.25	0.342	0.261	0.403	0.152	0.395	0.422	0.282
0.5	0.342	0.263	0.402	0.147	0.399	0.426	0.28
0.75	0.333	0.251	0.395	0.145	0.389	0.414	0.273
1.0	0.329	0.247	0.39	0.125	0.395	0.412	0.266
1.25	0.324	0.242	0.385	0.112	0.387	0.412	0.258
1.5	0.324	0.235	0.391	0.111	0.388	0.406	0.262
1.75	0.324	0.234	0.392	0.111	0.388	0.407	0.262
2.0	0.321	0.232	0.387	0.111	0.384	0.404	0.259

Table 8: In these tables, GFR has been detailed by decay and distribution similarity methods. There are no big differences except for scale.

where judgments can be solicited from assessors (albeit imperfectly) as noted by (Voorhees and Harman, 2005, page 24) for example, fairness is often a normative question as noted by (Balagopalan et al., 2023, page 2658), Fang et al. (2024), Zehlike et al. (2022a,b) for example. What counts as a fair outcome may vary across contexts, cultures, and communities. Thus, fairness evaluation may need to rely not only on quantitative metrics but also on qualitative assessments, stakeholder consultation, and ethical reflection.

Regarding evaluation, two questions naturally arise. First of all, can document collections be fair in general? Since document collections are built from broader sources, such as catalogs, portals, or subsets of the web, they will necessarily reflect the content present in those same sources. Therefore, if the content of the sources contains biases, the collections will also contain them. For example, in the Fair Track collections, there is a clear bias in favor of males, Europe, and North America. Does this bias represent the actual content? Unfortunately, yes. The statistics presented in Section 3 speak for themselves. These statistics also suggest that bias might not be eliminated by an automated system that retrieves and ranks documents in response to a topic, nor by a user who evaluates the retrieved documents as relevant or not, since users, like the authors of the documents, are affected by the same prejudices.

Nevertheless, the approaches to balance effectiveness and group fairness, like the one illustrated in this paper, ignore the context from which groups and categories are normed in the sense that only the frequency distribution of items over groups counts. Such a view may inevitably be seen as specific or even restrictive to the eyes of social experts; however, the balance is also a matter of statistics and computation. Computation plays a crucial role in quantifying disparities, yet it often fails to capture the nuanced realities of social dynamics. Therefore, integrating qualitative insights alongside quantitative methods is essential for a more comprehensive understanding of fairness and effectiveness in decision-making processes.

We are aware that the conclusions drawn from this paper may be conditioned by the context and the objectives of the experimental data. This specific test collections are a niche use case. The definition of bias would likely change in the case of general-purpose search in which not only Wikipedia articles are retrieved and typical web queries are sent to a search engine.

Fairness in information retrieval requires us to broaden our view of what retrieval systems are for and whom they serve. It asks us to reflect not only on what our systems retrieve but on what they leave out—-and whom they may leave behind. In doing so, we may discover that fairness, like relevance before it, is not a static concept but a moving target—one that will evolve as our systems and our societies do.

Acknowledgments and Disclosure of Funding

I thank the editor and the anonymous reviewers for the comments. No external funding was received in support of this work.

References

Ricardo Baeza-Yates. Bias on the web. *Communication of the ACM*, 61(6):54–61, 2018.

Aparna Balagopalan, Abigail Z. Jacobs, and Asia J. Biega. The role of relevance in fair ranking. In *Proceedings of SIGIR*, pages 2650–2660, 2023. doi: 10.1145/3539618.3591933.

Asia J. Biega, Krishna P. Gummadi, and Gerhard Weikum. Equity of attention: Amortizing individual fairness in rankings. In *Proceedings of SIGIR*, pages 405–414. ACM, 2018. doi: 10.1145/3209978.3210063.

Amin Bigdeli, Negar Arabzadeh, Shirin Seyedsalehi, Morteza Zihayat, and Ebrahim Bagheri. On the orthogonality of bias and utility in ad hoc retrieval. In *Proceedings of SIGIR*, SIGIR '21, pages 1748–1752, New York, NY, USA, 2021. Association for Computing Machinery. ISBN 9781450380379. doi: 10.1145/3404835.3463110.

Reuben Binns. On the apparent conflict between individual and group fairness. In *Proceedings of FAccT*, 2020.

Lidia Ceriani and Paolo Verme. Gini on mutability. *METRON*, 82(3):269–292, 2024. doi: 10.1007/s40300-024-00279-2.

Olivier Chapelle, Shihao Ji, Ciya Liao, Emre Velipasaoglu, Larry Lai, and Su-Lin Wu. Intent-based diversification of web search results: metrics and algorithms. *Information Retrieval*, 14(6):572–592, 2011.

Tim Draws, Nava Tintarev, Ujwal Gadiraju, Alessandro Bozzon, and Benjamin Timmermans. This is not what we ordered: Exploring why biased search result rankings affect user attitudes on debated topics. In *Proceedings of SIGIR*, pages 295–305, 2021. doi: 10.1145/3404835.3462851.

Michael D. Ekstrand, Graham McDonald, Amifa Raj, and Isaac Johnson. Overview of the trec 2021 fair ranking track. In *The Thirtieth Text REtrieval Conference (TREC 2021) Proceedings*, 2022a.

Michael D. Ekstrand, Graham McDonald, Amifa Raj, and Isaac Johnson. TREC 2022 Fair Ranking Track Participant Instructions. `https://fair-trec.github.io/docs/Fair_Ranking_2022_Participant_Instructions.pdf`, 2022b.

Michael D. Ekstrand, Graham McDonald, Amifa Raj, and Isaac Johnson. Overview of the trec 2022 fair ranking track. In *The Thirtieth Text REtrieval Conference (TREC 2021) Proceedings*, 2023.

Yi Fang, Ashudeep Singh, and Zhiqiang Tao. Fairness in search systems. *Foundations and Trends in Information Retrieval*, 18(3):262–416, 2024. doi: 10.1561/1500000101.

Corrado Gini. *Variabilità e mutabilità.* Tipografia di Paolo Cuppin, 1912.

Christina Ilvento, Meena Jagadeesan, and Shuchi Chawla. Multi-category fairness in sponsored search auctions. In *Proceedings of FAccT*, FAT* '20, pages 348–358. ACM, 2020. ISBN 9781450369367. doi: 10.1145/3351095.3372848.

Thomas Jaenich, Graham McDonald, and Iadh Ounis. Fairness-aware exposure allocation via adaptive reranking. In *Proceedings of SIGIR*, SIGIR '24, pages 1504–1513, New York, NY, USA, 2024. Association for Computing Machinery. ISBN 9798400704314. doi: 10.1145/3626772.3657794.

Thomas Jaenich, Graham McDonald, and Iadh Ounis. Fair exposure allocation using generative query expansion. In Claudia Hauff, Craig Macdonald, Dietmar Jannach, Gabriella Kazai, Franco Maria Nardini, Fabio Pinelli, Fabrizio Silvestri, and Nicola Tonellotto, editors, *Advances in Information Retrieval*, pages 267–281, Cham, 2025. Springer Nature Switzerland. ISBN 978-3-031-88717-8.

K. Jarvëlin and J. Kekäläinen. Cumulated gain-based evaluation of IR techniques. *ACM Transactions on Information Systems*, 20(4):422–446, 2002.

Till Kletti, Jean-Michel Renders, and Patrick Loiseau. Introducing the expohedron for efficient pareto-optimal fairness-utility amortizations in repeated rankings. In *Proceedings of WWW*, pages 498–507, 2022. doi: 10.1145/3488560.3498490.

Graham McDonald, Craig Macdonald, and Iadh Ounis. Search results diversification for effective fair ranking in academic search. *Information Retrieval*, 25(1):1–26, March 2022. doi: 10.1007/s10791-021-09399-z.

Massimo Melucci. A model of the relationship between the variations of effectiveness and fairness in information retrieval. *Discover Computing*, 27(3), 2024a.

Massimo Melucci. On the trade-off between ranking effectiveness and fairness. *Expert Systems with Applications*, 241, 2024b.

Massimo Melucci. Preference eigensystems for fair ranking. *Expert Systems with Applications*, 269:126324, 2025a. doi: 10.1016/j.eswa.2024.126324. https://www.sciencedirect.com/science/article/pii/S0957417424031919.

Massimo Melucci. Three methods for fair ranking of multiple protected items. *Scientific Reports*, to appear, 2025b.

Stefano Mizzaro. Relevance: The whole history. *Journal of the American Society for Information Science*, 48(9):810–832, 1997.

Marco Morik, Ashudeep Singh, Jessica Hong, and Thorsten Joachims. Controlling fairness and bias in dynamic learning-to-rank. In *Proceedings of SIGIR*, pages 429–438. ACM, 2020. doi: 10.1145/3397271.3401100.

Amifa Raj and Michael D. Ekstrand. Comparing fair ranking metrics, 2022. Arxiv. https://arxiv.org/abs/2009.01311.

S.E. Robertson, M.E. Maron, and W.S. Cooper. Probability of relevance: A unification of two competing models for document retrieval. *Information Technology: Research and Development*, 1:1–21, 1982.

Stephen E. Robertson. The probability ranking principle in information retrieval. *Journal of Documentation*, 33(4):294–304, 1977.

Tetsuya Sakai, Jin Young Kim, and Inho Kang. A versatile framework for evaluating ranked lists in terms of group fairness and relevance. *ACM Transactions on Information Systems*, 42(1), 2023. doi: 10.1145/3589763.

G. Salton. *Automatic Information Organization and Retrieval.* Mc Graw Hill, New York, NY, 1968.

T. Saracevic. Relevance: a review of and a framework for thinking on the notion in information science. *Journal of the American Society for Information Science*, 26:321–343, 1975.

T. Saracevic. Relevance: a review of the literature and a framework for thinking on the notion in information science. Part II: nature and manifestations of relevance. *Journal of the American Society for Information Science and Technology*, 58(3):1915–1933, 2007a.

T. Saracevic. Relevance: a review of the literature and a framework for thinking on the notion in information science. Part III: Behavior and effects of relevance. *Journal of the American Society for Information Science and Technology*, 58(13):2126–2144, 2007b.

E. Seneta. *Non-negative Matrices and Markov Chains*. Springer, 2006.

Cornelis Joost Van Rijsbergen. *Information Retrieval*. Butterworths, London, 1979.

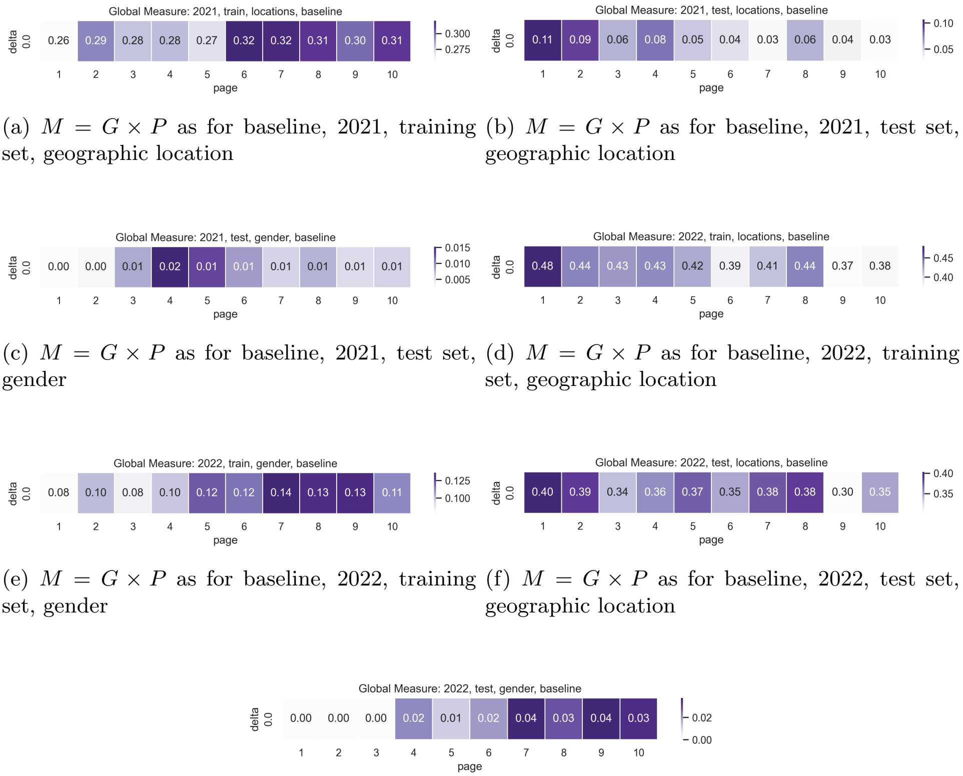
E.M. Voorhees and D.K. Harman, editors. *TREC: Experiment and Evaluation in Information Retrieval*. The MIT Press, Cambridge, MA, 2005.

Lin Xiao, Zhang Min, Zhang Yongfeng, Gu Zhaoquan, Liu Yiqun, and Ma Shaoping. Fairness-aware group recommendation with Pareto-efficiency. In *Proceedings of RecSys*, pages 107–115, 2017.

Meike Zehlike, Ke Yang, and Julia Stoyanovich. Fairness in ranking, part i: Score-based ranking. *ACM Computing Surveys*, 55(6), December 2022a. ISSN 0360-0300. doi: 10.1145/3533379.

Meike Zehlike, Ke Yang, and Julia Stoyanovich. Fairness in ranking, part ii: Learning-to-rank and recommender systems. *ACM Computing Surveys*, 55(6), December 2022b. ISSN 0360-0300. doi: 10.1145/3533380.

Appendix A. Heatmaps



(a) $M = G \times P$ as for baseline, 2021, training set, geographic location

(b) $M = G \times P$ as for baseline, 2021, test set, geographic location

(c) $M = G \times P$ as for baseline, 2021, test set, gender

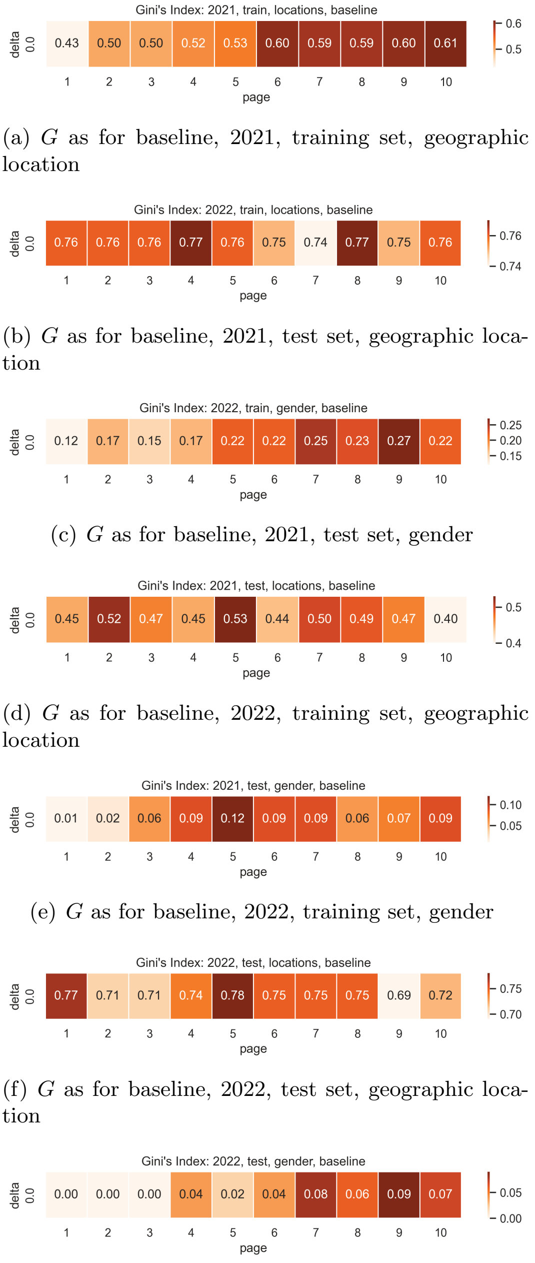
(d) $M = G \times P$ as for baseline, 2022, training set, geographic location

(e) $M = G \times P$ as for baseline, 2022, training set, gender

(f) $M = G \times P$ as for baseline, 2022, test set, geographic location

(g) $M = G \times P$ as for baseline, 2022, test set, gender

Figure 2: The heatmaps of this figure depicts the *baseline* global measure $M = G \times P$ by track (2021, 2022), phase (train, test), category (location, gender). It is the baseline $M = G \times P$ because ET was not activated in this case, i.e. $\delta = 0$. These heatmaps should be read in conjunction with Figures 3 and 4. In particular, Figure 4 not surprisingly shows that the baseline precision is higher at the top pages than at the bottom pages for all tracks, phases, and categories. Therefore, the variations of the baseline global measures mostly depend on the variations of Gini's Index as depicted by Figure 3. In sum, the rankings tend to be biased with respect gender to a larger extent than location.



(a) G as for baseline, 2021, training set, geographic location

(b) G as for baseline, 2021, test set, geographic location

(c) G as for baseline, 2021, test set, gender

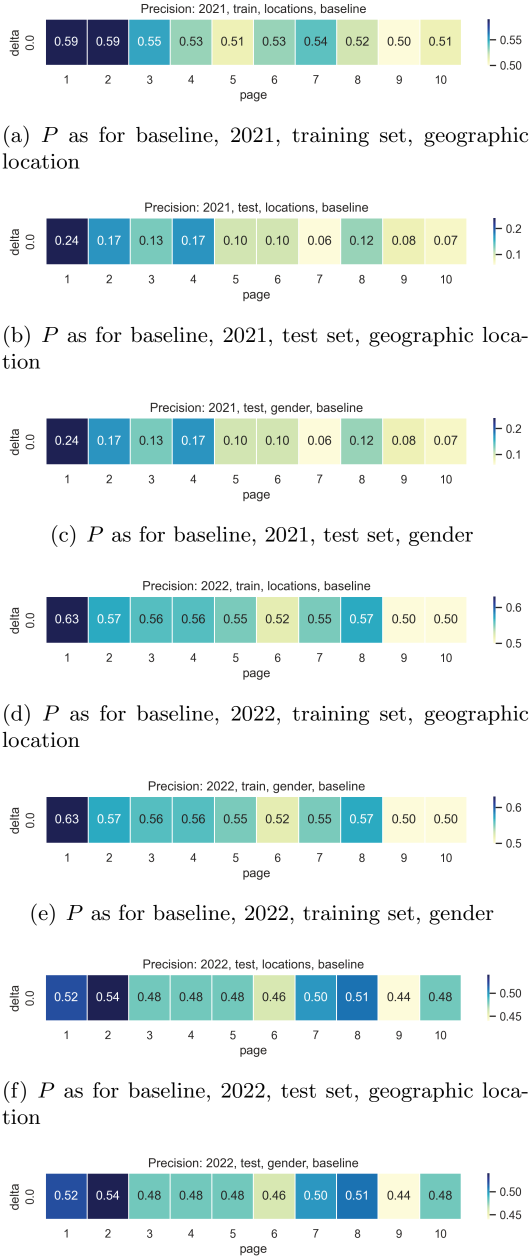
(d) G as for baseline, 2022, training set, geographic location

(e) G as for baseline, 2022, training set, gender

(f) G as for baseline, 2022, test set, geographic location

(g) G as for baseline, 2022, test set, gender

Figure 3: The rankings provided by a retrieval system are generally biased against certain groups, particularly non-males and non-Western regions, with the exception of location-based groups in the 2022 track.



(a) P as for baseline, 2021, training set, geographic location

(b) P as for baseline, 2021, test set, geographic location

(c) P as for baseline, 2021, test set, gender

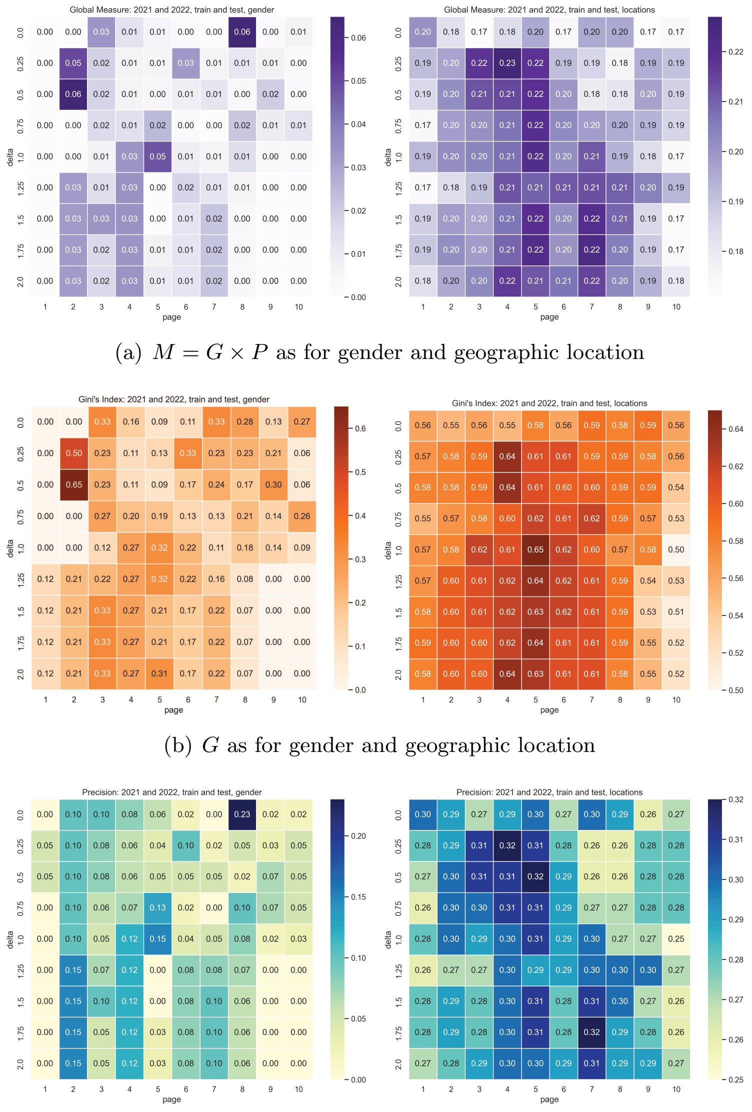
(d) P as for baseline, 2022, training set, geographic location

(e) P as for baseline, 2022, training set, gender

(f) P as for baseline, 2022, test set, geographic location

(g) P as for baseline, 2022, test set, gender

Figure 4: Retrieval effectiveness is relatively high, with minor variations across the 10-document pages. However, the top two pages are the most dense of relevant documents.

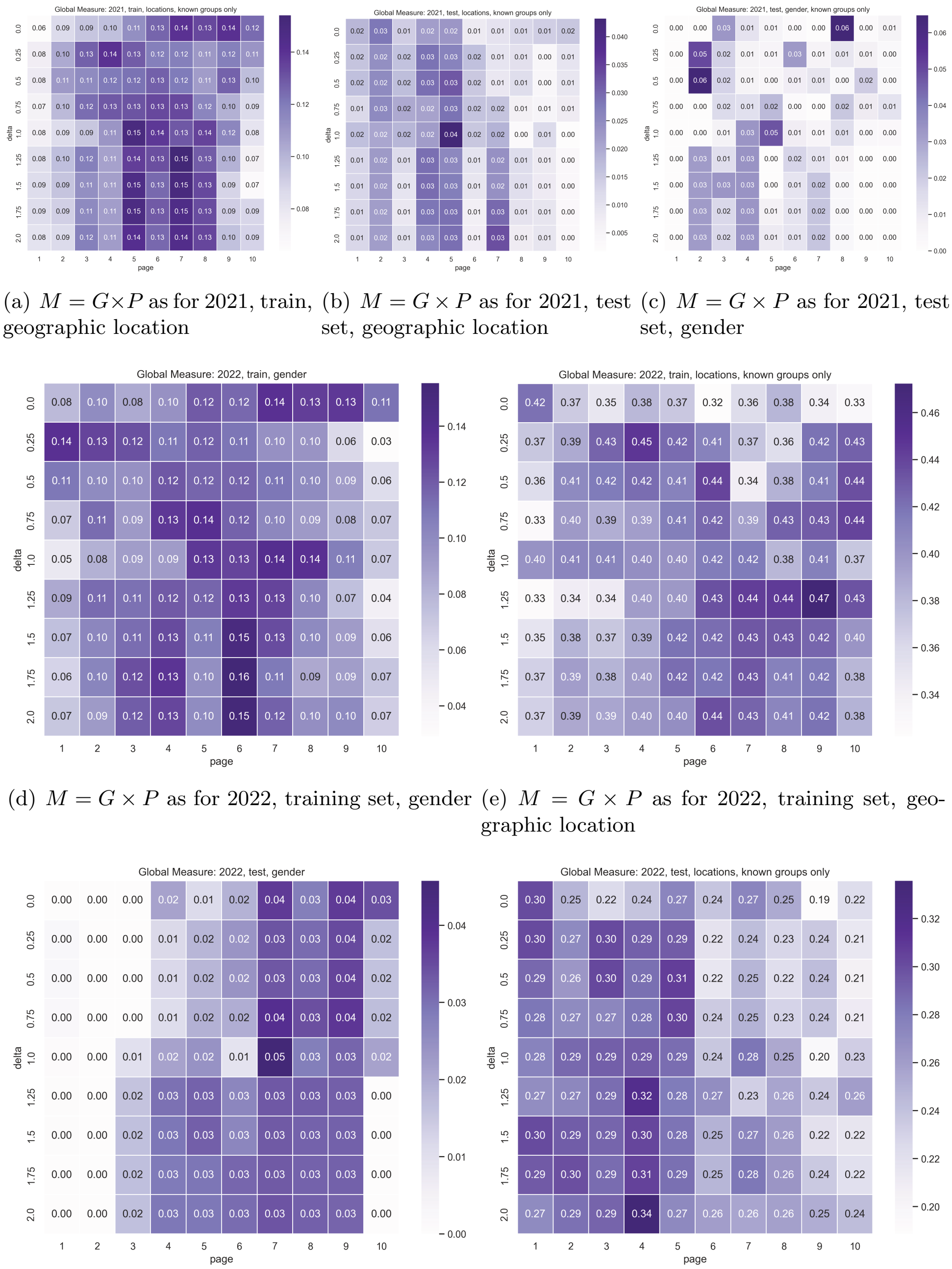


(a) $M = G \times P$ as for gender and geographic location

(b) G as for gender and geographic location

(c) P as for gender and geographic location

Figure 5: The figure shows heatmaps of global measures $M = G \times P$ for each 10-document page, ranging from 0 to 2. The heatmaps show that precision is no longer the highest at the top pages, indicating that the global measures observed may depend on both variations of Gini's Index and precision. Improving $M = G \times P$ requires increasing δ, indicating that ET is effective in improving fair, effective retrieval and ranking, as well as controlling variations of fairness and effectiveness. Some improvement can also be seen in gender, which is the most difficult bias due to the high prevalence of males in the test collection.



(a) $M = G \times P$ as for 2021, train, geographic location

(b) $M = G \times P$ as for 2021, test set, geographic location

(c) $M = G \times P$ as for 2021, test set, gender

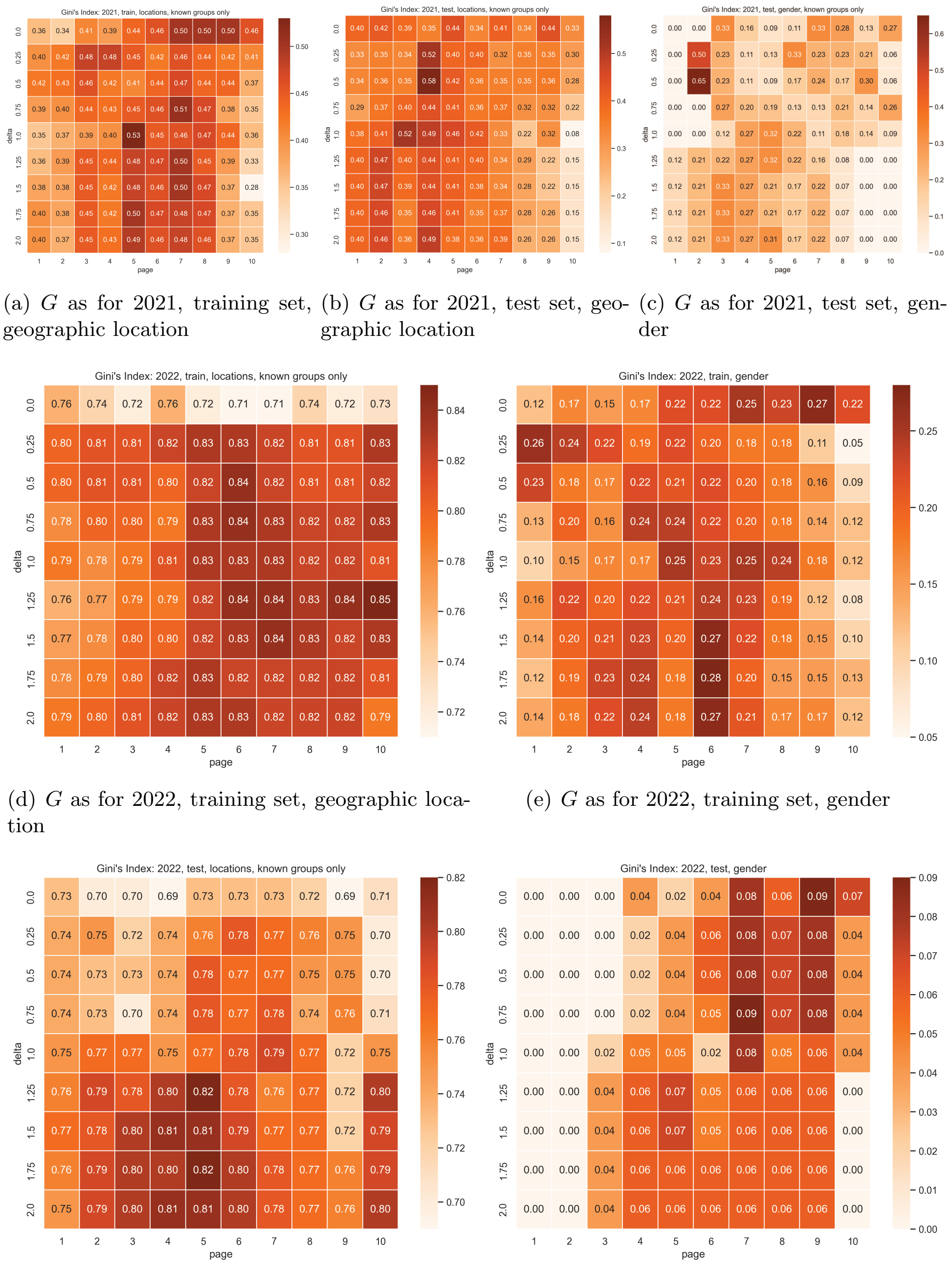
(d) $M = G \times P$ as for 2022, training set, gender

(e) $M = G \times P$ as for 2022, training set, geographic location

(f) $M = G \times P$ as for 2022, test set, gender

(g) $M = G \times P$ as for 2022, test set, geographic locations

Figure 6: In contrast to Figure 5(a), these heatmaps show $M = G \times P$ values by track, phase, and category.



(a) G as for 2021, training set, geographic location

(b) G as for 2021, test set, geographic location

(c) G as for 2021, test set, gender

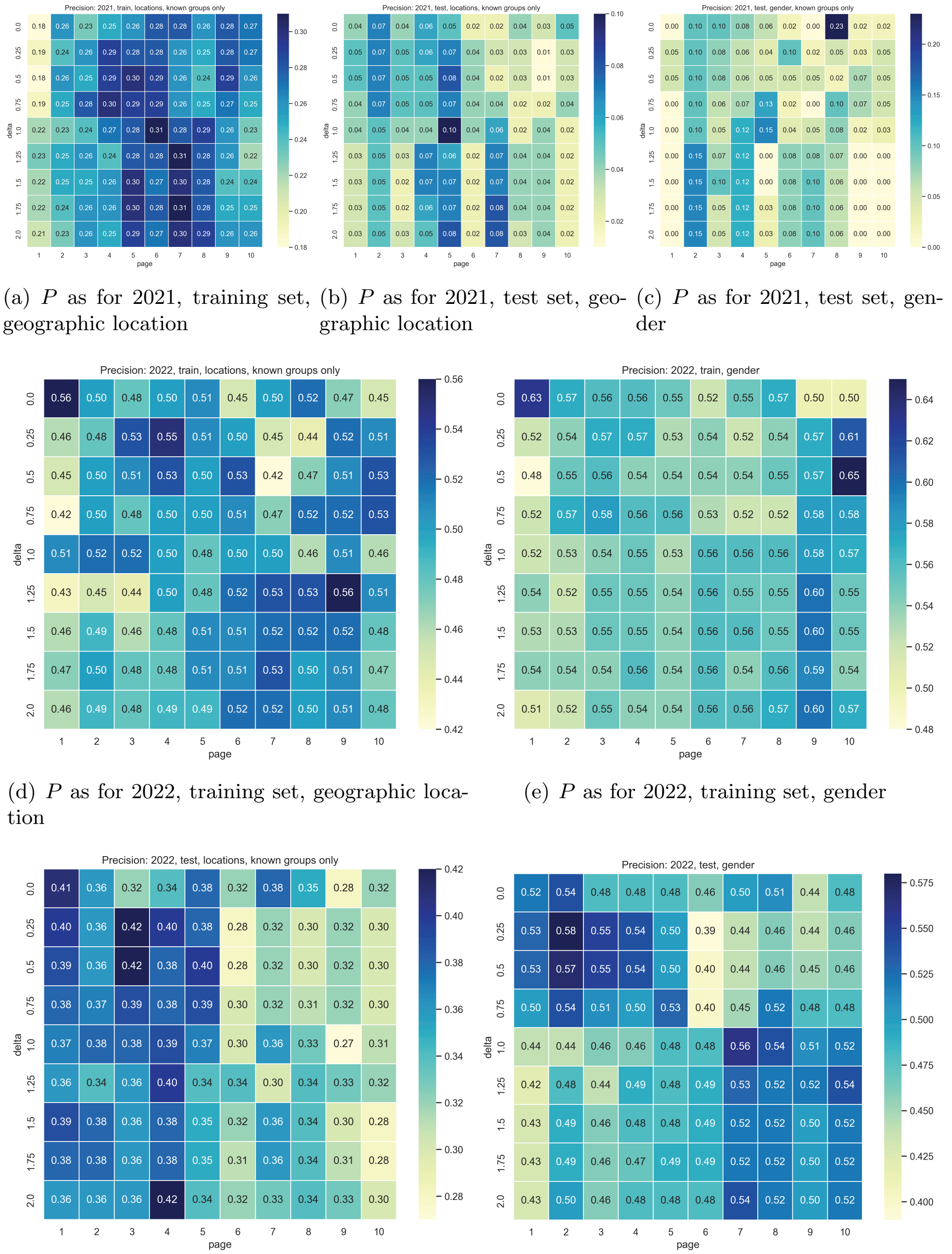
(d) G as for 2022, training set, geographic location

(e) G as for 2022, training set, gender

(f) G as for 2022, test set, geographic location

(g) G as for 2022, test set, gender

Figure 7: Unlike Figure 5(b), these heatmaps show G values by track, phase, and category.



(a) P as for 2021, training set, geographic location

(b) P as for 2021, test set, geographic location

(c) P as for 2021, test set, gender

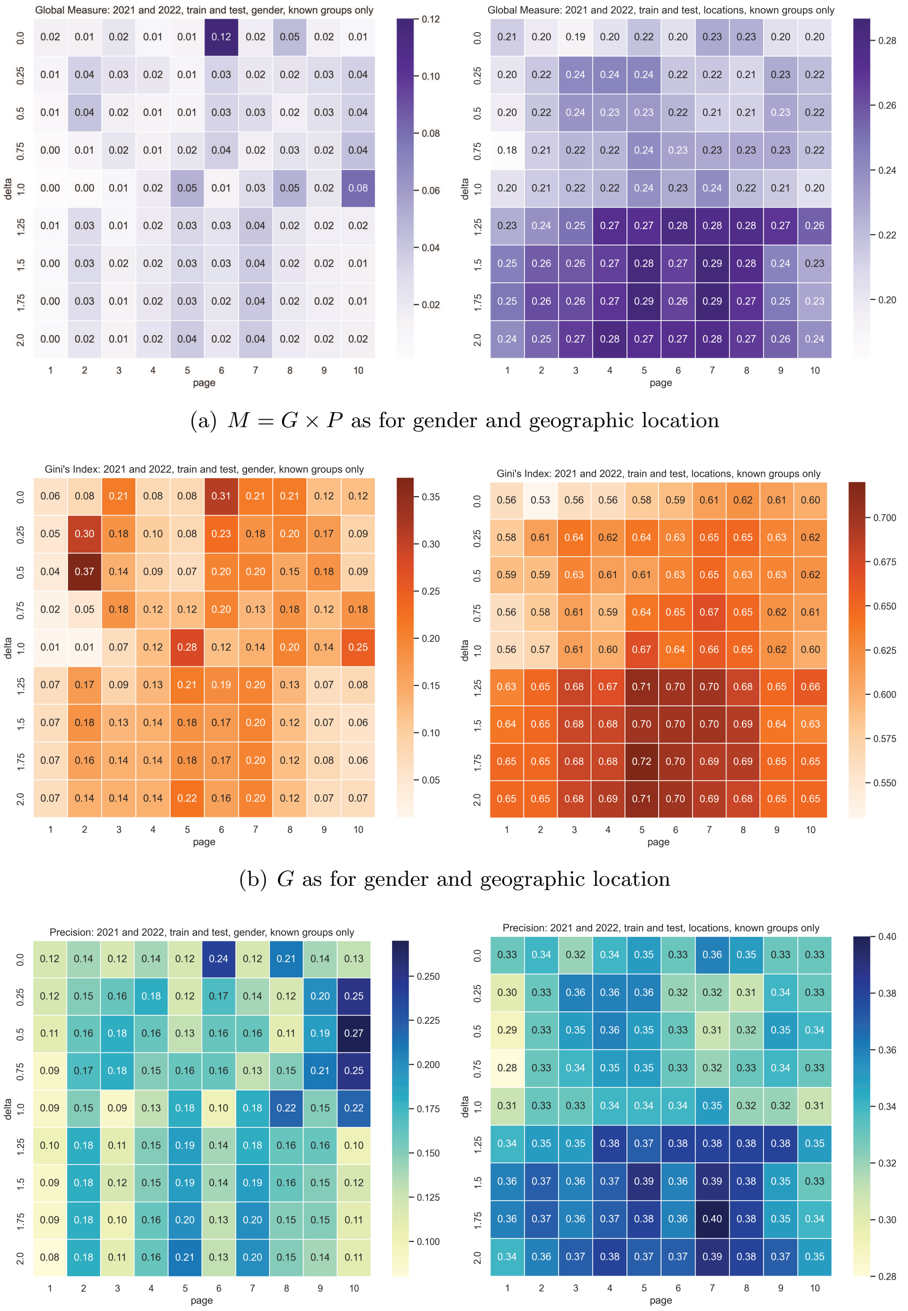
(d) P as for 2022, training set, geographic location

(e) P as for 2022, training set, gender

(f) P as for 2022, test set, geographic location

(g) P as for 2022, test set, gender

Figure 8: These heatmaps show P values by track, phase, and category, rather than combining tracks and phases as in Figure 5(c).



Global Measure: 2021 and 2022, train and test, gender, known groups only

Global Measure: 2021 and 2022, train and test, locations, known groups only

(a) $M = G \times P$ as for gender and geographic location

Gini's Index: 2021 and 2022, train and test, gender, known groups only

Gini's Index: 2021 and 2022, train and test, locations, known groups only

(b) G as for gender and geographic location

Precision: 2021 and 2022, train and test, gender, known groups only

Precision: 2021 and 2022, train and test, locations, known groups only

(c) P as for gender and geographic location

Figure 9: Unknown group membership appear to have a minor impact. Compare heatmaps to Figures 5(a), 5(b), and 5(c).

Appendix B. Statistical Significance Testing

Fairness: 2021 test locations										
δ	1	2	3	4	5	6	7	8	9	10
0.25										
0.50										
0.75	.02									
1.00										
1.25										
1.50										
1.75										
2.00										

Effectiness: 2021 test locations										
δ	1	2	3	4	5	6	7	8	9	10
0.25										
0.50										
0.75										
1.00	.0						.01			
1.25	.0						.05			
1.50	.0								.04	
1.75	.0								.04	
2.00	.0								.02	

Fairness: 2022 train gender										
δ	1	2	3	4	5	6	7	8	9	10
0.25										
0.50										
0.75										
1.00										
1.25										
1.50										
1.75										
2.00										

Effectiness: 2022 train gender										
δ	1	2	3	4	5	6	7	8	9	10
0.25	.02									.01
0.50	.0								.03	.0
0.75	.01								.05	
1.00	.01								.02	
1.25	.03								.0	
1.50	.02								.01	
1.75	.03								.01	
2.00	.01								.01	

Fairness: 2022 train locations										
δ	1	2	3	4	5	6	7	8	9	10
0.25										
0.50										
0.75	.02	.01								
1.00		.02								
1.25										
1.50										
1.75	.01									
2.00		.01								

Effectiness: 2022 train locations										
δ	1	2	3	4	5	6	7	8	9	10
0.25	.01									.0
0.50	.02					.04				.0
0.75	.0									.01
1.00										
1.25	.0	.04							.0	.02
1.50	.0									
1.75	.01									
2.00	.01									

Fairness: 2022 test gender										
δ	1	2	3	4	5	6	7	8	9	10
0.25										
0.50										
0.75										
1.00										
1.25										
1.50										
1.75										
2.00										

Effectiness: 2022 test gender										
δ	1	2	3	4	5	6	7	8	9	10
0.25										
0.50										
0.75										
1.00		.04								
1.25										
1.50										
1.75										
2.00										

Fairness: 2022 test locations										
δ	1	2	3	4	5	6	7	8	9	10
0.25				.03		.02			.03	
0.50				.02					.04	
0.75									.04	
1.00		.0	.0	.0		.01				
1.25	.05	.0	.0	.0	.01	.0			.05	
1.50	.01	.0	.0	.0	.02	.01			.04	
1.75	.04	.0	.0	.01	.01	.0	.04			
2.00		.0	.0	.0	.03	.0			.05	

Effectiness: 2022 test locations										
δ	1	2	3	4	5	6	7	8	9	10
0.25										
0.50			.04							
0.75										
1.00										
1.25										
1.50										
1.75										
2.00										

Information Retrieval Research 1 (2025) 281-312 Submitted 07/25; Revised 10/25; Published 12/25

Exploring Embedding Interpretability by Correspondences Between Topic Models and Text Embeddings

Meng Yuan MENG.YUAN@UNIMELB.EDU.AU
School of Computing and Information Systems
University of Melbourne, Melbourne, Australia

Lida Rashidi LIDA.RASHIDI@RMIT.EDU.AU
School of Computing Technologies
RMIT University, Melbourne, Australia

Justin Zobel JZOBEL@UNIMELB.EDU.AU
School of Computing and Information Systems
University of Melbourne, Melbourne, Australia

Editor: Solomon Atnafu

Abstract

Text embeddings have become essential for representing documents in Information Retrieval (IR), yet their high-dimensional nature often limits interpretability. To bridge this gap, we introduce a novel mapping framework that aligns embedding dimensions with topics derived from both probabilistic and neural models. Using three standard collections and three embedding methods, we demonstrate that embedding features consistently map to a subset of coherent topics, even as the total number of topics varies. We further quantify this correspondence with a Mean Mapping Specificity Improvement Rate, showing that mapped topics exhibit significantly higher specificity than the global topic set if the embedding dimensions are set properly. A stability analysis over varying embedding dimensions confirms the stability of the mapping across random feature samples. Our contributions are three-fold: A general-purpose mapping method that visualizes and formalizes correspondences between embedding features and topic representations; Empirical evidence that text embeddings and topic models are not independent descriptors but can mutually validate each other's semantic structures; A numeric indicator that captures the degree to which embedding features correspond to high-quality topics, providing a new tool for evaluating embedding interpretability and guiding dimensionality reduction choices. These findings suggest that topic-embedding mapping can serve both as a diagnostic for embedding quality and as a means to visualise embedding dimensions more human-interpretable, advancing the practice of collection description in IR.

Keywords: Embedding Interpretability, Language Model Explanability, Topic Modelling

1 Introduction

Text representation is a key task for collection description and organisation in information retrieval (IR) and related areas. Word embedding has become a widely adopted strategy for representing words with fixed-length vectors. However, the representation of longer texts and collections of texts is more challenging than word representations (Incitti et al., 2023).

WWW: https://irrj.org DOI: 10.54195/irrj.23703

In this work, we study the interpretability of text embeddings for documents by linking them to a fundamentally different but more interpretable collection descriptor, topic modelling.

Topic modelling and text embedding are two approaches to semantic representation of document collections. Although text embedding has been extensively examined as an effective method for capturing semantic characteristics of words and individual documents, topic modelling is more typically used to extract shared semantic information among documents in a collection. The common structure of embedding models usually includes a neural network, an encoding method for the input documents, and a predefined length of output vectors. Some of these models have achieved strong performance in IR tasks, including document classification, query formulation, and recommendation systems (Mikolov et al., 2013; Vulić and Moens, 2015; Ganguly et al., 2015). Due to the effectiveness of embedding techniques in such applications, a variety of refinements have been proposed for conventional topic models using text embeddings (Wu et al., 2020). The applications of text embedding in topic modelling include smoothing the bag-of-words representations as inputs (Seifollahi et al., 2021; Gupta and Patel, 2021); capturing rich semantic features of short texts (Meddeb and Romdhane, 2022; Zuo et al., 2023; Murshed et al., 2022); and constructing new topic model structures with pre-trained language models (Sia et al., 2020; Meng et al., 2022; Jin et al., 2021).

The many approaches for incorporating text embeddings into topic modelling reflect the relatedness of these two techniques. In this paper we argue that the document representations and topical characteristics of the collection loosely correspond to each other and should not be considered as independent sources of information. Instead, we can exploit one to increase the interpretability of the other. If we take a more general view of text embeddings, noting that the primary objective is to learn representations of individual documents, the features also reflect collection-wise semantic characteristics of the collections that were used for training. Likewise, an objective of topic modelling is to capture collection-wise features to describe the documents as a whole. Therefore, studying the connection between the two document representations helps us to understand how the individual dimensions of an embedding capture the semantics of the collection.

A challenge in exploring correlations between topic modelling and text embeddings is that they are assessed under different principles. In IR, a typical evaluation of a model involves a pipeline for assessment and validation. For example, to evaluate the performance of a new text embedding model, it is usually compared with a state-of-the-art baseline model within a series of tasks. However, the use of a pipeline may be an obstacle to understanding the correspondences and differences between distinct kinds of algorithm. There are two reasons why a baseline model is not suitable for this research. First, interpretability of text embedding is not a property that has been defined and quantified explicitly, and a single measure of embedding is insufficient to understand the interpretability of text embeddings in general. Second, a lack of interpretability does not suggest any direct impact on the performance of a model in standard tasks. Therefore, we seek a different approach to study interpretability in a way that can assess text embeddings and topic modellings within a unified narrative of collection description.

To address the challenges in the interpretability of text embeddings discussed above, we design experiments with the following objectives.

- To investigate the relationship between topic modelling and text embedding as document representations for retrieval purposes;
- To improve the interpretability of text embeddings by exploiting the mapping to topics;
- To explore novel approaches to evaluate the effectiveness of collection descriptors beyond a standard model-testing pipeline.

The contributions of this work are three-fold: First, we propose a new mapping method between topic models and text embeddings that visualises the correspondence between topics and embeddings in a range of ways. Second, we demonstrate that topic models and text embeddings are not independent sources of information as document representations, and further that their correspondence can be used to verify the performance of each other. Third, we propose a numeric indicator ΔS, to quantify the degree of correspondence between topic models and text embedding with our proposed mapping method.

The paper is organised as follows: In Section 2, we review relevant literature on topic models, topic specificity, and text embedding. In Section 3, we describe our proposed mapping strategy and discuss the theoretical implications with respect to collection description. The experiments and results are presented in Section 4. Finally, we conclude in Section 5.

2 Related work

We first discuss several state-of-the-art text embedding methods used for collection description and organisation, as well as strategies used for studying embedding interpretability. We also introduce the general principles of topic modelling and how the quality of topics generated by such models is evaluated; and specifically discuss a recent measure, topic specificity, which we introduced in earlier work (Yuan et al., 2023).

2.1 Text embedding

Text embedding aims to transform natural language into high-dimensional vectors. The embeddings have different levels of granularity, with the first embeddings that were proposed focusing on vectorised representation of individual words (Mikolov et al., 2013). More recent embeddings focused on longer text representation (Le and Mikolov, 2014). To differentiate between these two types of embedding, we refer to the former as *word embedding* and those representing longer texts as *text embedding.*

Early embeddings were mostly designed to represent words, where the objective is to represent words as vectors in a high-dimensional semantic space; the number of dimensions of the semantic space can be fixed or user-specific depending on the construction of the system (usually a neural network) used to train the embeddings. These embeddings can be categorised as language models, as they aim to optimise based on the likelihood of generating target sequences of words. As a word embedding is usually dependent on the contextual features around the target word, there is a trend when using large collections to obtain pre-trained embeddings. These can then be further tuned for a specific application. Examples of word embeddings include *Word2Vec* (Kim et al., 2017; Mikolov et al., 2013), GloVe (Pennington et al., 2014), and BERT (Devlin et al., 2019). These embeddings are pre-trained representations of words that capture semantic features of natural languages (Şenel et al., 2022).

In more recent work, new proposed embedding approaches aim to transform sentences or even passages into vectors of the same length (Le and Mikolov, 2014; Reimers and Gurevych, 2019; Ganesh et al., 2016; Incitti et al., 2023). Le and Mikolov (2014) introduced *Doc2Vec* embedding, which takes the general structure of *Word2Vec* embeddings and adds an additional word-like vector to represent paragraph-wise information. There are also text embeddings that use a pre-trained language model; for example, Reimers and Gurevych (2019) proposed *Sentence-BERT*, which adds a pooling process on top of the BERT structure and yields a fixed-length vector to represent the input sentence. Despite the strong performance of these embeddings in IR tasks such as query expansion, the interpretability remains a challenge in understanding what is captured by the embeddings, in part because text embeddings are constructed from word embeddings that are already difficult to interpret.

Sentence embeddings have been widely applied in IR. *Word2Vec* (Mikolov et al., 2013), which generates semantic representations of words using a skip-gram model, has been used to improve query formulation and refinement (Ganguly et al., 2015). Vulić and Moens (2015) built cross-language retrieval models with their bilingual skip-gram model. The success of text embeddings is because the focus of study in IR is usually on documents; sentence embeddings can convert documents with varying lengths into fixed-length representations. The document embeddings can then be used to calculate distances between documents in multiple downstream tasks, such as clustering (Kim et al., 2017), collection organisation and retrieval (Zuccon et al., 2015; Tanabe et al., 2018; Guo et al., 2022; Zhan et al., 2020), and recommendation (Karvelis et al., 2018; Yang et al., 2016; Hassan et al., 2018).

2.2 Interpretability of embeddings

Interpretibility refers to the users' ability to comprehend decisions made by a machine learning model and to predict the outputs of this model given a certain task (Miller, 2019). Murdoch et al. (2019) defines interpretable machine learning as 'extraction of relevant knowledge from a machine-learning model concerning relationships either contained in data or learned by the model'. In the context of this work, the interpretability of embeddings describes the extent to which users can understand in real-world terms what the dimensions of the embedding capture.

Studying the interpretability of embeddings helps us understand how embeddings capture the semantics of texts and may, for example, assist with model tuning for a specific purpose. To make an embedding interpretable, an intuitive solution is to learn what the individual dimensions represent and what features they capture in terms of the source text. The desirability of interpretability became apparent because it was challenging to explain the performance of the early word embedding models such as Word2vec (Mikolov et al., 2013) and GloVe (Pennington et al., 2014), as well as contextualised embeddings such as ELMo (Peters et al., 2018) and BERT (Devlin et al., 2019). These models learn the semantic features of raw texts and output fixed-sized vectors to represent words. The output embeddings serve as base representations for more complex models (Şenel et al., 2022). Studying the interpretability of word embeddings helps us to understand these more complex models.

Techniques for studying the interpretability of embedding can be categorised into two types: task-based interpretation and vector-space interpretation. An example of task-based interpretation is intrusion detection (Prouteau et al., 2022), which involved an evaluation framework for the interpretation of word embedding with a word intrusion task. For the dimension of interpretation of embedding, the set of words with high weights in this dimension must be semantically coherent. Therefore, interpretability is measured by the number of intruder words in the set. More recently this approach has been extended to use word embeddings rather than raw words to achieve task-based interpretation of longer text embeddings. For example, Jha et al. (2023) proposed Contrastive Long Document Encoder (CoLDE), a transformer-based framework to capture document similarity at three different levels of granularity within or across documents. The fine-grained similarity scores generated by the framework not only aid better interpretability of the documents but also support improved understanding of long-text embedding behaviours. CoLDE and other interpretable encoders help with answering one question: When the embeddings of two documents are similar, to what extent can humans understand why they are similar? However, the focus is on interpreting the similarity between documents rather than on interpreting the embedding space.

There is also research that approaches the interpretability challenge by exploring the vector space of the embeddings. Ethayarajh (2019) studied contextualised word embeddings by comparing the geometry of embedding spaces, where the intuition is that for a word in different contexts, the embeddings are different. They used cosine similarity to compute the distance between the two embeddings of a word in different contexts. The significance of this work is that it provides a quantitative measure of semantic difference when the meaning of a word is extended or varied. It served as a foundation to many quantitative studies on embedding interpretability for more complex embeddings driven by large language models (LLMs), and also as an inspiration of this work, on quantifying semantic difference among various embedding approaches at the document-collection level.

Despite the differences in these approaches to evaluation of interpretability, these techniques tend to consider the interpretability of an embedding as binary: that is, if an embedding does not satisfy the criteria of the assessment, it is considered as a non-interpretable embedding method. Nauta et al. (2023) viewed interpretability as a multifaceted characteristic and argued that a quantitative way of measuring interpretability should result in a multidimensional view that indicates the degree to which certain properties are satisfied. However, despite the survey's identification of a promising research direction for visualizing multidimensional embedding interpretability, follow-up work has been scarce, primarily because there is a lack of a suitable semantic proxy for studying embeddings of long texts that is both intuitively human-readable and rigorously quantifiable.

In addition to evaluating interpretability with intrinsic measures, interpretability can also be evaluated by downstream tasks that depend on the interpretability of embeddings. Tasks that require an understanding of text semantics can be used as a proxy to evaluate the performance of text embeddings. For example, Şenel et al. (2022) used sentiment analysis, question classification, and news classification to evaluate their proposed interpretable embedding, bidirectional imparting, or BiImp. However, it is not clear how interpretability contributes to the performance of embeddings in the tasks. Tao et al. (2023) proposed a deep learning-based framework for interpretable text classification, IDeL. The framework consists of three main components: feature penetration, instance aggregation, and feature perturbation. They also evaluated models constructed by this framework by two tasks, fake news detection and social question categorisation. Methods similar to this one can be considered as a combination of the task-based approach and vector space analysis approach to interpretability.

New embedding approaches have been proposed that are inherently interpretable by controlling the semantic features captured in each dimension. The Word2Sense model uses the LDA topic model (discussed further below) to extract the distributions of word sense and form an alignment to the dimensions of the output embeddings (Panigrahi et al., 2019). Subramanian et al. (2017) proposed sparse interpretable neural embeddings which transforms traditional embeddings such as word2vec into interpretable embeddings based on the denoising k-sparse autoencoder.

While research of this kind is focused on generating interpretable embeddings, there is also work that seeks to use interpretable embeddings for downstream tasks such as text classification. Singh et al. (2023) proposed a novel approach to augmenting interpretable models for text classification with LLMs. They used LLMs for training but not inference so that the models can be trained in a decoupled way and have shown that this method can be used to solve practical tasks such as language zone identification in neuroscience.

A key feature that interpretable embeddings share is vector sparsity. Since the shared goal of these embeddings is to generate interpretable document representations and collection descriptions, higher dimensional vectors can capture more details of the collection than vectors with smaller dimensions. To describe the same amount of information, the interpretable embedding vectors are usually sparser than traditional non-interpretable embeddings. When embeddings are learnt, they are learnt for a certain task. For a language model, the task is to generate word sequences that maximise the likelihood of a target sentence. Therefore, the features learnt need to address the aspect of natural language. The underlying assumption is that the embeddings remove and reorganise the original features in the natural-language representation and generate new features that better suit the task. In contrast, non-interpretable embeddings are usually pre-trained for general purpose use. They should reflect the characteristics of the source collection without obvious bias towards any aspect of the semantics.

Overall, while these interpretability studies have considered significant aspects of how embeddings encode information, a unified, scalable framework for long texts is still absent.

Our approach addresses this gap by combining vector-space and task-based perspectives into a visual framework of embedding interpretability and can serve as a basis for more transparent and tunable embedding methods. Unlike studies that focus on improving or benchmarking embedding models, our framework is embedding-agnostic: it operates on any document-level vector representation and evaluates how such embeddings can be interpreted through their correspondences with topic models, rather than on their downstream performance.

2.3 Topic modelling

Topic modelling is an approach to learning representative themes from collections of documents. Early topic modelling approaches include statistical models such as latent semantic indexing (LSI) (Deerwester et al., 1990), non-Negative Matrix Factorisation (NMF) (Gillis and Vavasis, 2014; Wang and Zhang, 2012), and singular value decomposition (SVD) (Zheng and Wang, 2022; Steinberger and Ježek, 2005). These models follow a shared principle: the difference between term frequency within a document and term frequency in the collection as a whole is used as an indication of topicality. The limitation of these models is that topic modelling based on term frequency is coarse, so the performance of such models can be poor when topics are used in downstream tasks in which specific semantics is required.

Latent Dirichlet Allocation (*LDA*) (Blei et al., 2003), which is a representative and widely used probabilistic topic model, addresses the limitation of statistical models. In contrast to statistical models, *LDA* does not rely solely on the decomposition of the document term matrix; instead, it assumes that the distribution of terms is conditioned on a set of latent topics and that the distribution of topics is conditioned on the collection of documents (Blei et al., 2003). For many years *LDA* has been the most robust topic modelling approach as evaluated by topic coherence (Newman et al., 2010; Mimno et al., 2011; Morstatter and Liu, 2017), topic diversity (Bu et al., 2021), and perplexity (Blei et al., 2003).

Recent research seeks to use word embeddings and other forms of neural networks to improve traditional topic modelling methods, including *LDA*, by providing richer forms of input documents and topical structures. Cheng et al. proposed bi-term topic modelling (BTM), which learns term co-occurrence generation patterns instead of single-term sequence patterns in order to resolve the sparsity issue of term co-occurrence in short texts. BTM replaces the original term representations with embeddings that consist of multiple words and has achieved better topic coherence and diversity compared to several baseline models.

There has been a trend of using pre-trained language models on top of the conventional topic models to solve some particular problems: Han et al. proposed the unified neural topic model by utilising contrastive learning with conventional term-based representations and pre-trained language models to detect keywords from semantically coherent clusters; Gaussian *LDA* (Das et al., 2015) replaced the original term representation in conventional *LDA* with word embeddings drawn from a multivariate Gaussian distribution; and in the CluWords topic model (Viegas et al., 2019), the term representation in Non-negative Matrix Factorisation (NMF) is replaced by the representation of the nearest term from a pre-trained word embedding model to form a meta-representation of the documents, which enhances the representation of both syntactic and semantic information. The state of the art of embedding-enhanced topic models, *BERTopic*, generates topics by clustering the document embedding representations learnt from a pre-trained language model and calculating class-based TF-IDF vectors as the topic representation for each cluster (Grootendorst, 2022).

2.4 Topic specificity

Our recent work on topic modelling evaluation provides a theoretical foundation for assessment of topic quality from the perspective of document representation. To evaluate the degree to which a topic is representative of a subset of documents in a collection, Yuan et al. (2023) proposed the *topic specificity* measure. Topic specificity can be used on any topic model that represents documents as a vector of percentages or weightings of topics, including conventional topic models such as *LDA* and neural topic models such as *BERTopic*.

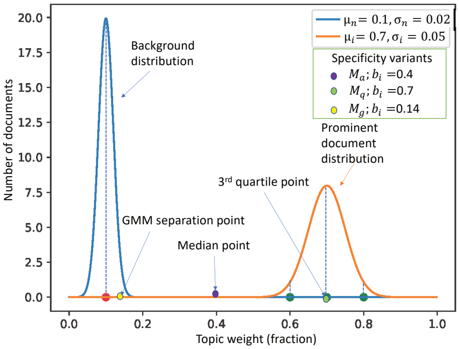


Figure 1: An illustration of the *bimodality* of topic weight distribution over a collection. This figure is reproduced from Yuan et al. (2023).

The derivation of this measure is associated with the phenomenon of *bimodality* that generally exists in topic models. Figure 1 shows how bimodality appears when an ideal topic is generated from a document collection: when a topic is used to describe a collection, the topic only shows high weights in those truely relevant documents while existing as background noise in the rest of the collections. By plotting topic weight of individual documents against the number of documents, a topic with good quality usually shows two peaks in the distribution plot - we call this shape *bimodality*.

Topic specificity is a quantification of the *bimodality* phenomenon. The intuition behind topic specificity is that a topic that is descriptive should have high weights in a small subset of the documents and low weights for the remainder of the documents. If we consider the topic weights of a single topic over the collection as a distribution, we should be able to find a threshold (the base) that differentiates the described subset from the background collection. Adopting the notation of the original paper, the specificity S of an arbitrary topic t_i is defined as

$$S(t_i) = \frac{\mu_i}{1 - b_i} \tag{1}$$

where μ_i is the average distance between prominent documents and the base and b_i is the base of the weight distribution of the topic. High specificity means that the topic clearly discriminates between the subset in which it is prominent and the remainder of the collection.

The base b_i can be calculated using the median of the topic's weight distribution over the collection, the third quartile of the distribution, or a Gaussian Mixture Model (GMM). The way we choose is by GMM since this is a complete mathematical simulation of distribution mixtures without any manual intervention. Hence, for a background distribution $X_n \sim \mathcal{N}(\mu_n, \sigma_n^2)$, the base b_i and the mean μ_i are defined as

$$b_i = 2\sigma_n + \mu_n \tag{2}$$

$$\mu_i = \frac{\sum_{d \in D_i}(f_i^d - b_i)}{|D_i|} \tag{3}$$

where D_i is the prominent document set in a document collection D for topic t_i and f_i^d is the weight of topic t_i for each document $d \in D_i$.

The range of topic specificity scores is 0–1, where 0 means that the topic is not descriptive of any document, and 1 means that the topic is only representative of a specific document. Values between 0.2 and 0.4 are generally observed to be high topic-specificity scores.

Specificity is more sensitive to noise in the topics than other coherence-based measures, which makes it suitable for use in contexts where robust sensitivity to topic quality is required. In addition, topic specificity quantifies how individual documents are described by a limited number of topics, directly capturing the 'aboutness' of topics on collection level. These properties make topic specificity a valid measure of topic quality for our use case and a reliable proxy for embedding semantic interpretation.

3 Methodology

We describe our approach in this section. First, we describe the document collections and models used for the experiments. Then, we use an illustration of embedding and topic model representations of document collection to explain the logic behind the correspondence study. Next, we describe our proposed mapping approach, a quantitative measure of mapping quality, and a feature sampling approach for the stability analysis of the mapping performance.

3.1 Collections

In our experiments, we used three document collections. The smallest document collection contains 5,000 documents and the largest 20,000. The following describes these collections and their corresponding topics:

wiki: A sample of 5,000 documents from a Wikipedia dump. We randomly selected 1,000 documents from five categories each to create our labelled dataset. The categories chosen are culture, geography, health, history, and mathematics.[1]

20NG: Contains 18,846 documents that were collected from the BBC News website. The documents are grouped into 20 news categories (Lang, 1995). [2]

wsj: Contains 98,733 Wall Street Journal articles collected as part of the TREC disks (Voorhees and Harman, 2005). Due to computational limitations, we randomly selected 20,000 documents from the collection. This is an unlabelled collection, and hence the natural number of topics is unknown. [3]

[1] Categories are selected as in `https://en.wikipedia.org/wiki/Wikipedia:Contents/Categories`. Pages were crawled on 2 September 2022. The dataset is available upon request.

[2] 20 News Group dataset is available at `http://people.csail.mit.edu/jrennie/20Newsgroups/`

[3] The Wall Street Journal (WSJ) collection used in TREC evaluations is part of the proprietary Tipster and TREC disks distributed by the Linguistic Data Consortium (LDC). It is not publicly available and requires a license agreement for access.

3.2 Model selection

For text embeddings, we select the models that are particularly designed for sentence or long-text vectorisation, namely *Doc2Vec* and *Sentence-BERT*. Note that there are many embedding approaches that have outstanding performance in IR tasks, but they are mostly term-based embeddings and not suitable for the task at hand. For fair comparison and a deeper illustration of the alignment between topics and embedding features, we generate three types of sentence-based embeddings in our experiments.

***Doc2Vec* Le and Mikolov (2014).** An extension of the *Word2Vec* model, which was described earlier in Section 2.1. In addition to word-level modelling, *Doc2Vec* adds another vector that captures the sequence characteristics of the sentence in the document as part of the input. To train *Doc2Vec* embeddings, the documents need to be tokenised and lemmatised first and the number of dimensions needs to be pre-defined. The default number of dimensions is 400.

***Sentence-BERT*.** An extension of the conventional BERT model, which was described in Section 2.1. It uses a pre-trained language model as the input to generate document-level embedding vectors. Unlike *Doc2Vec*, the dimensionality of *Sentence-BERT* embeddings is fixed and depends on the pre-trained language model used for training. The number of default dimensions is 384.

***RepLLaMa*.** A retrieval-enhanced embedding method built on the LLaMA transformer architecture (Ma et al., 2024). During fine-tuning, RepLLMa is trained with a document-level retrieval objective so that its hidden-state representations capture rich semantic alignments across documents. The model outputs fixed-length vectors (e.g., 4096 dimensions). For fair comparison, the dimensions of embeddings are reduced to a range between 10 and 200 using random projection. This adjustment prevents performance differences from being confounded by representational capacity and aligns the embedding dimensionality with the scale of the topic space. That is, for the selection of collections, the number of topics is much lower than the dimension of *RepLLaMa*, 4096. In addition, using random projection preserves pairwise distances between documents. According to the Johnson-Lindenstrauss lemma (Johnson and Lindenstrauss, 1984; Bingham and Mannila, 2001; Achlioptas, 2003), dimensionality reduction for extremely sparse high dimensional space may be achieved without substantial loss of geometric information.

3.3 Text embedding features

As introduced in Section 2.1, text embeddings are document representations that interpret documents of varying lengths as vectors of fixed dimensions. Intuitively, for a set of text embeddings learnt from the same collection, we can interpret the dimensions of the embeddings as the signature facet shared by all documents from a document collection. If we learn a set of embeddings for a document collection, we can represent the collection as a set of feature vectors. Figure 2 shows an illustration of a document-feature matrix that represents the entire collection. The rows are text embeddings of the documents; the columns represent the semantic aspects shared by all the representations of the documents. In other words, the column vectors are the semantic characteristics of the collection described by the document embeddings. From now on, we refer to these collection representations learnt

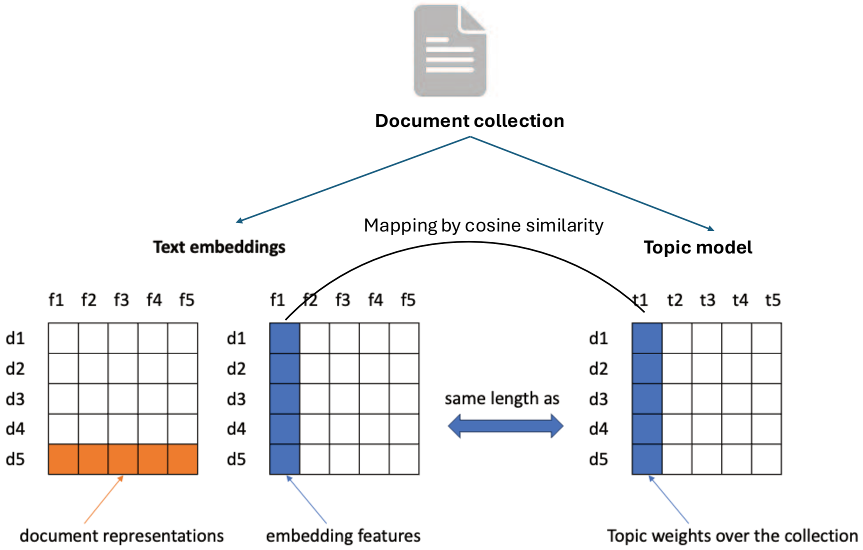


Figure 2: Both the text-embedding matrix and the topic model are derived from the same document collection. Unlike conventional use of text embeddings where document vectors (rows) are compared, this mapping operates on the embedding-feature vectors (columns) of the embedding matrix. Each embedding feature forms a vector over the document collection, having the same dimensionality (number of documents) as a topic vector from the topic model. To align the two representations, we compute the cosine similarity between each embedding feature and each topic vector, identifying the closest pairs as feature–topic matches that characterize semantic correspondences between the embedding and topic spaces.

by embeddings as *embedding features*. In previous attempts to create interpretable embeddings, the dimensions of the embedding vector are usually encoded as collection-wise topics or concepts. This is a reasonable assumption based on the expected utility of the embeddings. If the assumption is correct, we should be able to observe these topics from the collection regardless of the embedding method, that is, for the embeddings that learn the features implicitly from the collection, the dimensions should also reflect a semantic aspect of the collection.

Assuming that these embedding features are representations of collection topics implies that the embedding features should have a large overlap with other forms of topic representation, such as LDA topics. However, LDA topics and embedding features are produced by completely different principles and mechanisms: the document-topic representations are derived from probabilistic inference based on the topic-term distributions of topic models while the embedding features are learnt by the likelihood of a target sequence of words described in a fixed number of dimensions.

A key factor that differentiates these two types of collection representation is interpretability; topics can be interpreted as the distribution of topic weights over the collection, whereas the embedding features are simply real-valued numbers in a high-dimensional space.

This can be verified by a bimodality check: If we take the weights of a topic in a document collection and plot them as a distribution, a good topic should have a bi-modal shape. However, for dense embeddings, if we take the values of an embedding feature in the same collection and plot the distribution, the embedding features usually have a normal distribution. The bimodal distribution is what the sparse embeddings are intended to achieve. In other words, a single feature should describe a small subset of the documents, while being irrelevant to most of the documents in the collection. These features are also intended to be disjoint, so that they only matter to certain documents if the documents contain the semantic information described by the features.

Hence we conclude that topics and embeddings are not equivalent – an unsurprising conclusion perhaps, but also one that provides a basis for the next steps in our investigation.

3.4 Topic-embedding mapping approach

We propose mapping methods between embedding features and topic models. For a document collection $\mathcal{D}$, we learn an embedding representation $\mathcal{V}$ with shape $|\mathcal{D}| \times |\mathcal{V}|$. Then we train a topic model T and infer the document-topic representation with shape $|\mathcal{D}| \times |\mathcal{T}|$. As explained in the previous section, there are $|\mathcal{V}|$ embedding features and $|\mathcal{T}|$ topical representations learnt to describe the same collection.

Mathematically, both $\mathcal{V}$ and $\mathcal{T}$ can be regarded as projections of the same set of documents $\mathcal{D}$ into different lower-dimensional subspaces. Each subspace encodes the same underlying semantic structure, though derived through different optimization objectives. To quantify the similarity between these subspaces, the comparison metric should depend on the directional alignment of their vectors rather than their absolute magnitudes, since both embeddings and topic weights can be arbitrarily scaled. Therefore, cosine similarity is used, as it measures the angular proximity between vectors, capturing semantic alignment while being invariant to scale. This makes it particularly suitable for comparing representations that express relative importance or association strength, as in the case of embedding features and topic distributions.

The *mapping* between topics and embedding features is the process of finding the dimensions of the embedding space and the topical space that are close to each other. For a set of candidate embedding features and a set of candidate topics, we define a feature-topic pair as a mapping if the candidate embedding feature and the topic have the largest cosine similarity of any pair formed in the candidate sets. According to the mapping, we can categorise the topics of a topic model into two types: *mapped topics*, which occur in at least one pair of the mapping; and *un-mapped topics*, which are topics that are not in any pair of the mapping.

The mapping is conducted on the basis of embedding features. For each feature from the embedding representation $v \in \mathcal{V}$, first compute the cosine similarity with each topic from the topic model $t \in \mathcal{T}$; then map feature v with topic t if it has the largest cosine similarity with feature v, that is, $\text{argmax}_{t \in \mathcal{T}}\, cosine(v, t)$; and add feature v to the mapping set of topic t and remove v from the embedding representation, that is, $\mathcal{V} \leftarrow \mathcal{V} - v$. This mapping is performed without applying a similarity threshold, as the goal is to assign every embedding feature to the most related topic, thereby ensuring complete feature coverage for interpretability analysis rather than filtering out weak associations. Theoretically, one topic can be mapped to zero or more embedding features. The one-to-many nature of the mapping contributes to the interpretability of embedding features.

Feature Vector Sparsity Justification. One of the common observations w.r.t. embedding features is that they are highly sparse. Intuitively, it seems inappropriate for a very sparse vector to be mapped to a dense topic vector. However, feature vectors and topic vectors are representations of different spaces; the sparsity of feature indicates the effectiveness of this feature in representing a particular aspect of a small subset of the collection. The features are supposed to be sparse for the optimal descriptiveness of the collection. An equivalent quantifier of the sparsity of embedding features is the bimodal shape of the topics of a topic model. An informative topic should be able to differentiate the documents that belong to it from the rest of the collection. If we consider the topic weight distribution over all documents in a collection, it is highly skewed; for most of the documents, the topic is irrelevant and therefore has a weight very close to zero; for a small subset of documents, the topic is the dominant topic and hence has very large weight. Now, if we take the topic weights as a vector, it is also a sparse vector. But the representation of the topic, the distribution of word weight in a topic, is always a dense vector, regardless of the specificity of the topic.

3.5 Mean mapping specificity difference

In the previous section, we explained how the mapping mechanism relates the embedding features to topics. If there is a good mapping, this can aid in the interpretability of text embedding. To demonstrate how this mapping can aid interpretability, we use a variant of topic specificity, which measures the extent to which a topic represents a particular characteristic of the document collection.

To measure the quality of a topic model with specificity, we calculate the mean specificity scores over a set of topics generated by a topic model. Intuitively, if the features learnt from the text embedding approach are more representative than the conventional topical features, the mapped topics of these features should have higher specificity than the overall specificity of all topics. Hence, we can quantify the increase in topic specificity by computing the difference between the mean specificity of topic sets that were originally generated by the topic model and the topic sets after mapping using the embedding features.

Using a topic model $\mathcal{T}$ as a reference, the quality of a set of embedding features can be expressed by the mean mapping specificity $\mathcal{S}'$ as follows:

$$\mathcal{S}' = \frac{\sum_{t \in \mathcal{T}'} \mathcal{S}_t \times \mathcal{N}_t}{\sum_{t \in \mathcal{T}'} \mathcal{N}_t} \tag{4}$$

where $\mathcal{N}_t$ is the number of features of the embedding representation mapped to topic t, $\mathcal{T}' \subseteq \mathcal{T}$ is the set of topics that have at least one mapped embedding feature, and $\mathcal{S}_t$ denotes the specificity score for topic t.

We can determine how using only the mapped topics can impact the specificity of the topic model. At one extreme, all features can be assigned to the topic with the highest specificity; conversely, at the other extreme, all features can be mapped to the topic with the lowest specificity. These two extremes create a boundary around how the mean specificity of the topic model can vary after the mapping is performed. Therefore, we can derive a measure of the alignment between text embeddings and topics within a bounded range of $[-1, 1]$. We call this measure the *Mean Mapping Specificity Improvement Rate*, denoted by ΔS, and define it as follows:

$$\Delta\mathcal{S}(\mathcal{T}', \mathcal{T}) = \begin{cases} \dfrac{\mathcal{S}' - \mathcal{S}}{\max_{t\in\mathcal{T}}(\mathcal{S}_t) - \min_{t\in\mathcal{T}}(\mathcal{S}_t)}, & \text{if } \max_{t\in\mathcal{T}}(\mathcal{S}_t) \neq \min_{t\in\mathcal{T}}(\mathcal{S}_t); \\ 0, & \text{otherwise.} \end{cases} \tag{5}$$

where $\mathcal{S}$ and $\mathcal{S}'$ denote the mean specificity over the topics in $\mathcal{T}$ and $\mathcal{T}'$, respectively. ΔS quantifies the degree of alignment between embedding features and topic specificity: a positive value indicates that more features are mapped to topics with above-average specificity, while a negative value suggests that features are more likely mapped to topics with below-average specificity. When all topics have identical specificity (i.e., zero variance), the denominator becomes zero; in such cases, the measure is set to 0, as no improvement in specificity can be defined.

3.6 Embedding feature sampling

To assess the stability of ΔS under different random projections, we vary the number of projected dimensions and repeat the projection with different random seeds. This measures how robust the mapping from embedding features to topics is when the embedding space is randomly re-parameterized.

We use the same three corpora and embedding methods (*Sentence-BERT*, *Doc2Vec*, *RepLLaMa*) as before. Each original embedding matrix is $E \in \mathbb{R}^{D\times N}$, where D is the number of documents and N the full embedding dimension. Let $\mathcal{F} = \{f_{\min}, f_{\min} + s, \ldots, f_{\max}\}$ denote the range of target dimensions (e.g. $f_{\min} = 10$, $f_{\max} = N$, step $s = 10$).

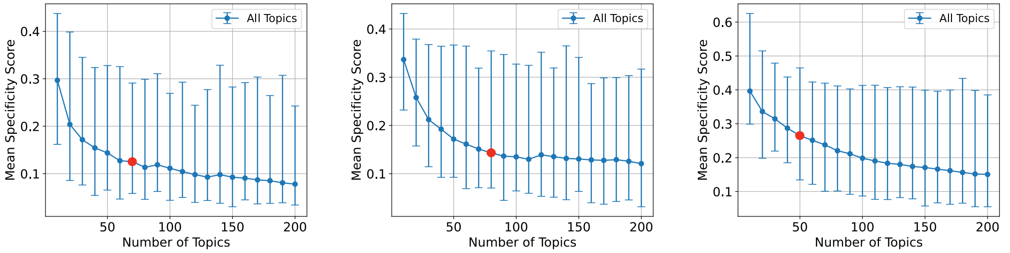
For each $f \in \mathcal{F}$, we perform B independent random projections:

1. Generate a random projection of E into f dimensions using a different random seed for each draw, producing $E_f^{(i)} = \text{RandomProj}(E; f, \text{seed} = i)$.
2. Apply the mapping procedure from Section 3.4 to align $E_f^{(i)}$ with the chosen topic model, yielding a set of (feature, topic) mappings.
3. Compute ΔS on the mapped topics and record the score for this projection.

This yields B ΔS values for each f. We summarize the results by reporting the mean and standard deviation of ΔS over the B random seeds, thereby quantifying mapping stability as a function of projection dimension.

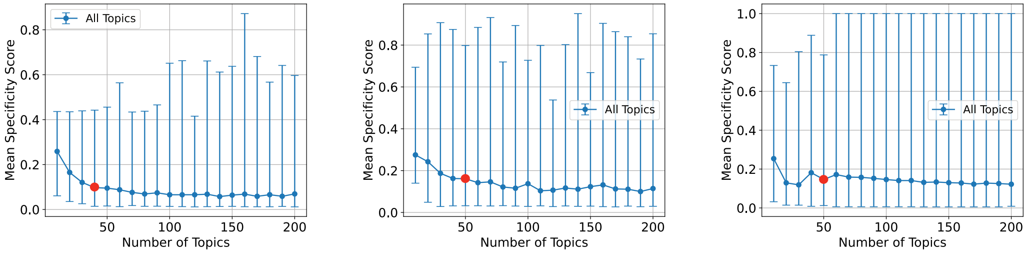
4 Experiments

With the proposed mapping method and the feature sampling approach described in Section 3, we design a set of experiments to reveal how topic models contributes to the interpretability of text embeddings. As for all the visualisation tasks of semantics of texts, the experiment processes include both quantitative measurements and visual interpretations. It ensures the interpretability study to generate not only numbers but also human-interpretable outcomes.



(a) Optimal $k = 70$ for 20NG. (b) Optimal $k = 80$ for WIKI. (c) Optimal $k = 50$ forWSJ.

Figure 3: Average topic specificity over all topics when ranging over number of topics for *LDA*. The curve's turning point marks where the mean specificity score levels off, indicating that adding more topics beyond this point yields diminishing returns and primarily introduces noise, as evidenced by subsequent drops in specificity. While the choice is somewhat arbitrary within this acceptable range, we found it sufficient for our purposes. We elaborate on this decision in the section *Experiments*.



(a) Optimal $k = 40$ for 20NG. (b) Optimal $k = 80$ for WIKI. (c) Optimal $k = 50$ for WSJ.

Figure 4: Average topic specificity over all topics when ranging over number of topics for *BERTopic*. The optimal number of topics is determined in the same way as for *LDA*. (In Plot (c) on WSJ, the top topic always scores 1—reflecting a single very specific core topic—but this does not affect our elbow choice, since we rely on the mean curve rather than the max/min scores.)

The experiments can be roughly divided into three components: Determination of the optimal topic model; Visualisation of topic evolution and mapping; Stability analysis of embedding semantics. These components are described in sections 4.1, 4.2, and 4.3.

4.1 Determining optimal numbers of topics

To show that our proposed mapping approach is valid for most topic models, the two we select are the most representative in their categories: *LDA* is the most commonly used statistical topic model and the foundation of many derivatives; *BERTopic* is a representative of the state-of-the-art neural topic models using pre-trained embeddings as input.

Due to the different nature of these topic models, they usually converge to different numbers of topics, k. Hence, we use a slope-plot with respect to topic specificity to determine the optimal number of topics for each selected topic model. First, we do a grid-search over the topic models by varying the number of topics from 10 to 200 with a step of 10 for both models in all collections. Next, for each run of the model, we calculate the specificity score per topic and plot the mean specificity per topic model run against the number of topics. As the number of topics increases, the topic model will have more noise topics with low specificity scores; even the topic with high quality will break, and eventually all topics end up with low specificity scores. Therefore, we end up with a decline of specificity scores. The last thing is to determine the optimal number of topics using the slope. The tuning point of the curve represents the optimal number of topics for the chosen topic model and collection. It reflects a balance where the number of coherent topics is sufficiently objective without being excessive, while noisy topics begin to separate out from the bulk of meaningful topics.

The precise location of the turning point is not critical, especially when the slope of the curve is relatively flat. In such cases, there often appears to be a range of values that can reasonably be considered part of the turning region. Selecting any representative value within this range is sufficient, as the exact number of noisy topics has limited impact on the subsequent mapping. In contrast, the number and quality of coherent topics play a more important role in determining the effectiveness and interpretability of the embedding-to-topic alignment.

The results of *LDA* and *BERTopic* of each collection are shown in Figures 3 and 4. For *LDA*, the optimal numbers of topics for collection 20NG, wiki, and wsj are chosen at 70, 80, and 50, respectively. For *BERTopic* model, the optimal numbers of topics are 40, 80, and 50. We will use these model runs for the following mapping experiments.

4.2 Embedding interpretability by mapping

We now examine the correspondence between the topics and the embedding features. An intuitive assumption of embedding features is their high semantic granularity compared to topics. Therefore, for a topic of good quality, the amount of information represented by the topic can be decomposed into multiple embedding features. We argue that mapping can help us identify topics that are more representative of individual documents, since topics with mapped features are closer to text embeddings than are other topics. Our aim is to demonstrate that, regardless of the number of features, which can be smaller or larger than the number of topics, features tend to map to topics with higher specificity, and at times multiple features are mapped to a topic that best captures document-level characteristics.

Intuitively, when there is an equal number of features and topics involved in the mapping, some topics will be left out due to their lack of similarity to any of the features. However, when there are fewer features than topics, it is possible that each feature is mapped to a different topic. In that case, the mapping would fail to distinguish the ‘good’ topics from the rest. This phenomenon is consistent for various number of topics and embedding features on all three collections. Note that topics with high specificity scores that are not mapped to any features are close to an already mapped topic, and therefore the mapping approach could not allocate a feature to them.

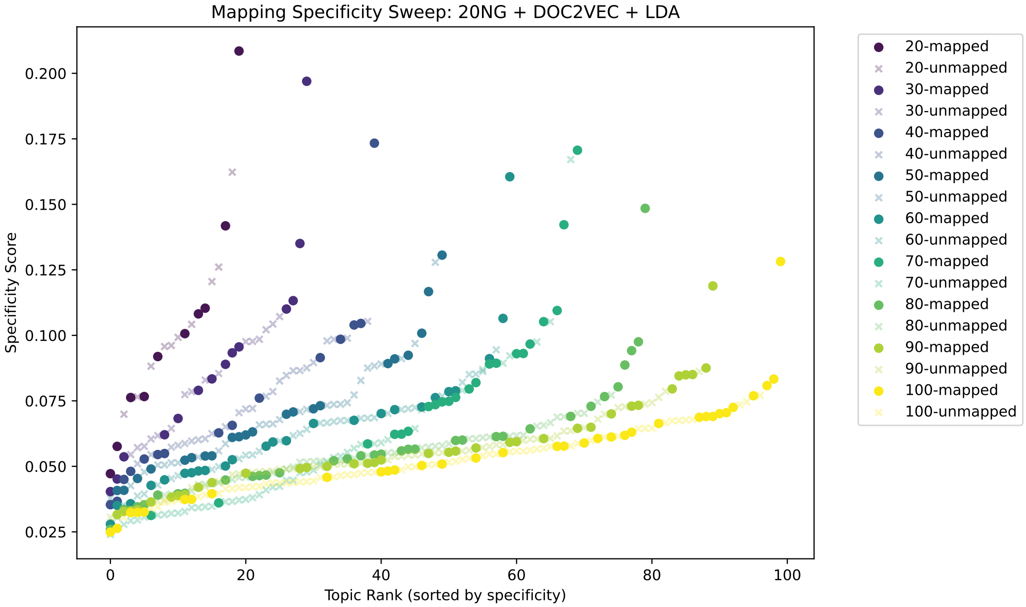


Figure 5: An example of using the mapping to visualise topic evolution. A series of *LDA* models of 20NG collection are mapped to the *Doc2Vec* embedding features of the same collection. When the number of topics is small, topic specificities are very high, but the mapping indicates the topics are likely to be too broad compared to the semantic dimensions captured by the embeddings. As the number of topics increases, topic specificities are dropping gradually due to the decomposition of large topics, yet the features tend to map to the 'good' topics towards the right upper corner of the plot.

Figure 5 illustrates use of the mapping framework to visualize topic evolution, in which we apply *LDA* models to the 20NG collection and map them to the *Doc2Vec* embedding features derived from the same corpus. When the number of topics is small, the resulting topics exhibit high specificity scores. However, the mapping reveals that these topics are often overly broad relative to the semantic dimensions captured by the embeddings. As the number of topics increases, specificity scores gradually decline due to the decomposition of large, general topics into more fine-grained ones. Nevertheless, the mapped features tend to concentrate in the upper-right region of the plot, indicating that they consistently align with a subset of coherent topics even as the overall topic structure becomes more granular.

We present mapping results for both topic models in Table 1. Based on the optimal number of topics selected by the elbow method in the previous step, we random project each embedding and reduce them to the same number of features for each collection. The mapping is then conducted between the topic models and embedding features for each collection. To illustrate the trend, Figure 6 provides an example of the mapping between *LDA* and *Sentence-BERT* and the mapping between *BERTopic* and *RepLLaMa*.

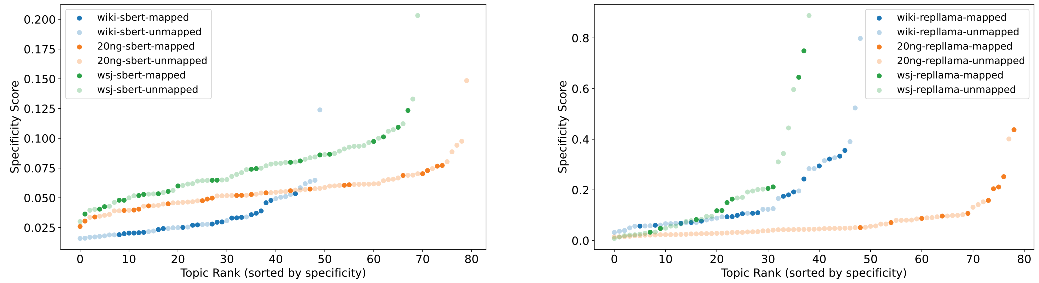
These results demonstrate the interpretability of embeddings in two ways. First, mapped topics usually have higher specificity scores, suggesting better descriptiveness of document

Collection	Topic model	Embedding	Mapped	Unmapped	Δ
WIKI	*LDA*	*Doc2Vec*	0.105	0.064	0.040
		Sentence-BERT	0.111	0.062	0.049
		RepLLaMa	0.118	0.066	**0.053**
	BERTopic	*Doc2Vec*	0.140	0.129	0.011
		Sentence-BERT	**0.158**	0.109	0.049
		RepLLaMa	0.134	0.118	0.016
20NG	*LDA*	*Doc2Vec*	0.097	0.051	0.046
		Sentence-BERT	0.095	0.045	0.051
		RepLLaMa	0.118	0.052	0.066
	BERTopic	*Doc2Vec*	0.200	0.059	0.141
		Sentence-BERT	0.191	0.058	0.132
		RepLLaMa	**0.208**	0.070	**0.138**
WSJ	*LDA*	*Doc2Vec*	0.105	0.078	0.027
		Sentence-BERT	0.129	0.075	0.054
		RepLLaMa	0.133	0.080	0.053
	BERTopic	*Doc2Vec*	0.183	0.143	0.041
		Sentence-BERT	0.264	0.153	0.111
		RepLLaMa	**0.341**	0.135	**0.206**

Table 1: Average topic–specificity scores (mapped vs. unmapped). Differences of average specificity of mapped and unmapped topics are shown in column Δ.

collection. Since the embedding features are mapped to these topics, the mapping results can be considered as validation of the embedding descriptiveness. Second, mapping multiple embedding features to the same topic suggests greater coverage of the semantics of a topic than an embedding feature. As embeddings generally have many more dimensions than the number of topics in a topic model, the granularity of information is higher in the embeddings. Consequently, to describe a concept in a document collection, text embeddings can use multiple features to decompose the concept. Therefore, a potential way to utilise the mapping results is to summarise what the embedding features describe by visualising the topics to which they are mapped. The topics can also be used to group the embedding dimensions into categories and serve as labels for better interpretability.

As a qualitative case study, we provide examples of mapped and unmapped topics in Appendix A to illustrate how the proposed mapping framework manifests in practice. For instance, in the WSJ collection, topics such as *"president, executive, chief, vice, officer, chairman, name, director, board"* and *"share, stock, trading, dividend"* are frequently mapped to Sentence-BERT features, indicating coherent, collection-relevant semantics that align with the quantitative ΔS results. In contrast, unmapped topics such as *"food, restaurant, franchise"* or *"insurance, policy, premium"* exhibit weaker internal coherence and



(a) Mapping between *LDA* and *Sentence-BERT* embedding features.

(b) Mapping between *BERTopic* and *RepLLaMa* embedding feartures.

Figure 6: Example visualisations of topic-model-to-embedding mappings. The number of scatter points for each collection corresponds to its optimal number of topics.

lower specificity scores. These examples complement the quantitative analysis by providing intuitive evidence that mapped topics tend to be more interpretable and semantically consistent, supporting our claim that the mapping can serve as a diagnostic tool for embedding interpretability.

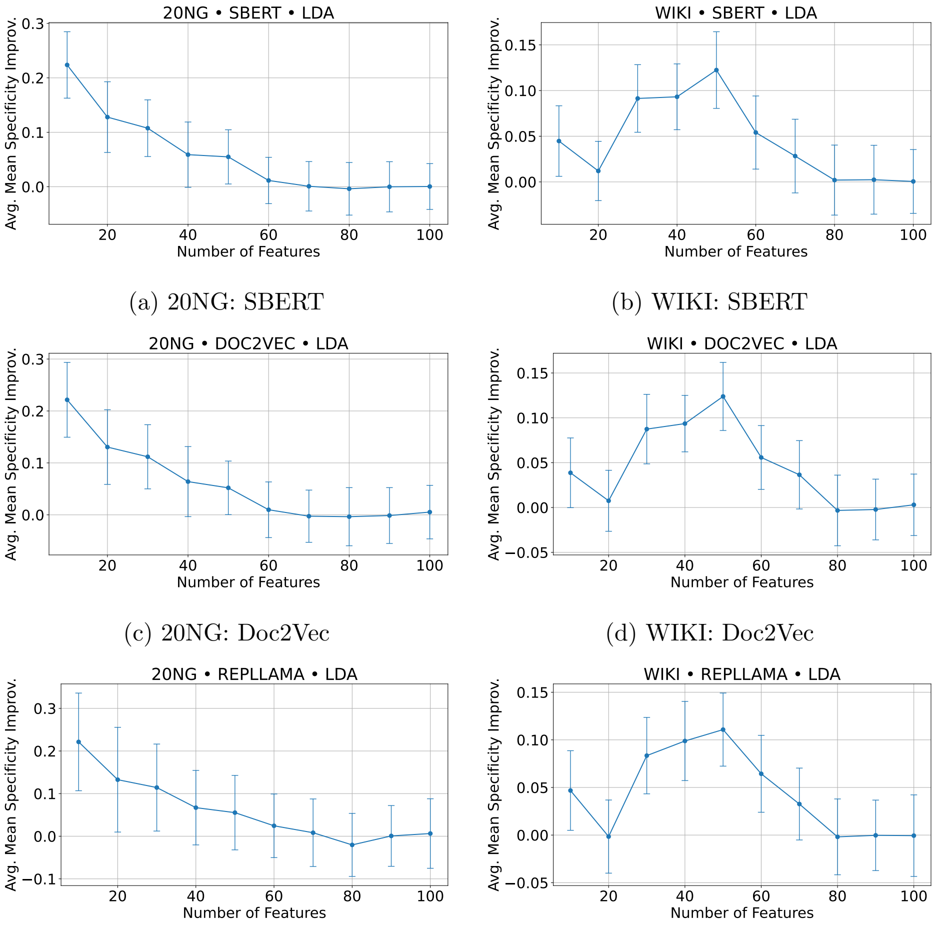
4.3 Stability analysis via feature sampling

The consistency of the mapping outcome is essential to the interpretability of embedding features. With the goal of generating human-interpretable representations of embedding features, the mapping to a topic model should generate robust and reproducible results so that the mapped topics can describe semantics of the embedding features being represented. To address this concern, and to conclude our exploration on the correspondence between topic models and text embedding, we seek to show the stability of the mapping approach by varying the number of sampled features via the method described in Section 3.6.

After sampling features with random projection for 100 runs on each settings of the three embeddings, we found that the topic model might be the ceiling of the interpretability of embedding features. Surprisingly, the optimal number of features is highly dependent on the topic model and less relevant to the embedding method – that is, there is likely to exist a limit on the granularity of semantic interpretability that a topic model can offer for a given document collection, regardless of the embedding methods used.

As the pattern seems consistent among these collection and topic model combinations, we present two of them as an example. Figure 7 show the mean and a min-max range of ΔS over 100 samples of embedding features with varying feature dimensions from 10 to 100 on 20NG collection and wiki collection. Each subplot represents the trend of ΔS of a embedding method. It is clear that the trend is nearly identical among embedding methods while differing between collections. For 20NG, the peak of mapping performance occurs when embedding features are reduced to 10 dimensions, compared to 50 dimensions for wiki. Note that the topic models used for this mapping experiment are consistent throughout all variations on the number of features.

Our stability analysis reveals that each topic model enforces its own semantic grain size on a collection, meaning that the optimal embedding dimensionality reflects the model's capacity to distinguish specific topics without introducing noise. As for future study, we will explore whether structured dimensionality-reduction methods, such as PCA or encoder-based models, can sharpen this stability peak and further improve interpretability. We also plan to apply our stability framework to alternative topic modeling paradigms (for example, non-negative matrix factorization or hierarchical Dirichlet processes) to uncover algorithm-specific limits on semantic granularity and guide model selection.



(a) 20NG: SBERT

(b) WIKI: SBERT

(c) 20NG: Doc2Vec

(d) WIKI: Doc2Vec

(e) 20NG: RepLLaMA

(f) WIKI: RepLLaMA

Figure 7: Trend of ΔS when mapping to the optimal *LDA* model with varying feature dimensions on (left column) 20NG and (right column) WIKI collections over 100 samples by random projection. Rows correspond to embedding models: (1) SBERT, (2) Doc2Vec, and (3) RepLLaMA.

5 Conclusion

Text embeddings have long been studied as a tool for collection description and organisation, but embedding interpretability remains a challenge, particularly at the collection level. There isn't a reliable correspondence measure between text embeddings and human-readable representations. However, recent advances in topic modelling evaluation have provided us with a tool to measure the descriptiveness of topics. To our knowledge, no prior work has considered comparing topic models and text embeddings as distinct representations of document collections or explored how they correspond.

In this paper, we propose exactly such a proxy: a robust, collection-level semantic representation derived from topic-model mappings, coupled with a specificity metric that is both human-interpretable, via topic keywords, and quantitatively tractable. The mapping allows us to conduct a systematic comparison between topical representation and embedding representations of the same collection. The mapping confirms as correspondence between topic modelling and text embedding as collection description tools. We have proposed a numerical indicator of mapping performance, *mean mapping specificity improvement rate*, which captures the change in the mean specificity score over mapped topics to that of the original set of topics. This indicator, with values ranging from $[-1, 1]$, can help us to identify embedding features that correspond to descriptive topics and hence enhance the interpretability of embedding approaches.

This work has three contributions. First, we proposed a mapping method between topic models and text embeddings that contributes to the interpretability of embedding features with an arbitrary number of dimensions. Second, we show that topics and text embedding are not two independent sources of information when used as document representations by using *embedding-based mapping* between topics and embedding features; topics can be used as a proxy to infer conceptual information captured by embedding features. Third, we successfully quantify information loss in embeddings using the dimension reduction technique with the mean mapping specificity improvement rate; the measure can be used for comparison between different embedding techniques for semantic studies. These should allow selection of topics for tasks such as collection or document annotation, and can plausibly provide verification that the generated embeddings and topics are of good quality.

However, this approach also has limitations. The quality and interpretability of the mapping are dependent on the topic model's coherence. Poorly defined topics can lead to misleading correspondences. Moreover, biases inherent in the dataset may propagate through both topic modelling and embedding representations, influencing the observed relationships and specificity scores. Future work should explore techniques to mitigate these effects, such as bias correction and topic refinement strategies.

Overall, by mapping embeddings to topic models, we create a multidimensional interpretability visualisation framework that is a step beyond simple binary judgements on interpretability, scales to longer texts at collection level, and can directly influence downstream semantic tasks. It bridges the gap between task-based explanation and vector-space analysis of existing studies on embedding interpretability, and offers a systematic approach for understanding and tuning embeddings in real-world applications.

References

D. Achlioptas. Database-friendly random projections: Johnson–Lindenstrauss with binary coins. *Jour. of Computer and System Sciences*, 66(4):671–687, 2003. doi: 10.1016/S0022-0000(03)00025-4.

E. Bingham and H. Mannila. Random projection in dimensionality reduction: Applications to image and text data. In *Proc. ACM SIGKDD Int.Conf. on Knowledge Discovery and Data Mining*, pages 245–250, 2001. doi: 10.1145/502512.502546.

D. M. Blei, A. Y. Ng, and M. I. Jordan. Latent Dirichlet allocation. *Jour. of Machine Learning Research*, 3:993–1022, 2003. doi: 10.1162/jmlr.2003.3.4-5.993.

Y. Bu, M. Li, W. Gu, and W. Huang. Topic diversity: A discipline scheme-free diversity measurement for journals. *Jour. of the American Society for Information Science and Technology*, 72(5):523–539, 2021. doi: 10.1002/asi.24433.

X. Cheng, X. Yan, Y. Lan, and J. Guo. Btm: Topic modeling over short texts. *IEEE Trans.on Knowledge and Data Engineering*, 26(12):2928–2941, 2014. doi: 10.1109/TKDE.2014.2313872.

R. Das, M. Zaheer, and C. Dyer. Gaussian LDA for topic models with word embeddings. In *Proc. Annual Meeting of the Association for Computational Linguistics and Int. Joint Conf. on Natural Language Processing*, pages 795–804, 2015. Association for Computational Linguistics. doi: 10.3115/v1/P15-1077.

S. Deerwester, S. T. Dumais, G. W. Furnas, T. K. Landauer, and R. Harshman. Indexing by latent semantic analysis. *Jour. of the American Society for Information Science and Technology*, 41(6):391–407, 1990. doi: 10.1002/(SICI)1097-4571(199009)41:6⟨391::AID-ASI1⟩3.0.CO;2-9.

J. Devlin, M. Chang, K. Lee, and K. Toutanova. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proc. Conf. of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 4171–4186, 2019. Association for Computational Linguistics. doi: 10.18653/v1/N19-1423.

K. Ethayarajh. How contextual are contextualized word representations? Comparing the geometry of BERT, ELMo, and GPT-2 embeddings. In *Proc. Conf. on Empirical Methods in Natural Language Processing*, pages 55–65, 2019. Association for Computational Linguistics. doi: 10.18653/v1/D19-1006.

J. Ganesh, G. Manish, and V. Vasudeva. Doc2sent2vec: A novel two-phase approach for learning document representation. In *Proc. ACM-SIGIR Int. Conf. on Research and Development in Information Retrieval*, page 809–812, 2016. Association for Computing Machinery. doi: 10.1145/2911451.2914717.

D. Ganguly, D. Roy, M. Mitra, and G. J. F. Jones. Word embedding based generalized language model for information retrieval. In *Proc. ACM-SIGIR Int. Conf. on Research and Development in Information Retrieval*, SIGIR '15, page 795–798, 2015. Association for Computing Machinery. doi: 10.1145/2766462.2767780.

N. Gillis and S. A. Vavasis. Fast and robust recursive algorithmsfor separable nonnegative matrix factorization. *IEEE Trans.on Pattern Analysis and Machine Intelligence*, 36(4): 698–714, 2014. doi: 10.1109/TPAMI.2013.226.

M. Grootendorst. Bertopic: Neural topic modeling with a class-based TF-IDF procedure. *ArXiv*, abs/2203.05794, 2022. doi: 10.48550/arXiv.2203.05794.

J. Guo, Y. Cai, Y. Fan, F. Sun, R. Zhang, and X. Cheng. Semantic models for the first-stage retrieval: A comprehensive review. *ACM Trans.on Information Systems*, 40(4), 2022. doi: 10.1145/3486250.

H. Gupta and M. Patel. Method of text summarization using LSA and sentence based topic modelling with BERT. In *Int. Conf. on Artificial Intelligence and Smart Systems*, pages 511–517, 2021. doi: 10.1109/ICAIS50930.2021.9395976.

S. Han, M. Shin, S. Park, C. Jung, and M. Cha. Unified neural topic model via contrastive learning and term weighting. In *Proc. Conf. European Chapter of the Association for Computational Linguistics*, pages 1802–1817, 2023. Association for Computational Linguistics. doi: 10.18653/v1/2023.eacl-main.132.

H. A. M. Hassan, G. Sansonetti, F. Gasparetti, and A. Micarelli. Semantic-based tag recommendation in scientific bookmarking systems. In *Proc. ACM Conf. on Recommender Systems*, page 465–469, 2018. Association for Computing Machinery. doi: 10.1145/3240323.3240409.

F. Incitti, F. Urli, and L. Snidaro. Beyond word embeddings: A survey. *Information Fusion*, 89:418–436, 2023. ISSN 1566-2535. doi: 10.1016/j.inffus.2022.08.024.

A. Jha, V. Rakesh, J. Chandrashekar, A. Samavedhi, and C. K. Reddy. Supervised contrastive learning for interpretable long-form document matching. *ACM Trans.on Knowledge Discovery from Data*, 17(2), 2023. doi: 10.1145/3542822.

Y. Jin, H. Zhao, M. Liu, L. Du, and W. Buntine. Neural attention-aware hierarchical topic model. In *Proc. Conf. on Empirical Methods in Natural Language Processing*, pages 1042–1052, 2021. Association for Computational Linguistics. doi: 10.18653/v1/2021.emnlp-main.80.

W. B. Johnson and J. Lindenstrauss. Extensions of lipschitz mappings into a hilbert space. In *Conf. in Modern Analysis and Probability*, volume 26 of *Contemporary Mathematics*, pages 189–206. American Mathematical Society, 1984. doi: 10.1090/conm/026/737400.

P. Karvelis, D. Gavrilis, G. Georgoulas, and C. Stylios. Topic recommendation using Doc2Vec. In *Int. Joint Conf. on Neural Networks*, pages 1–6, 2018. IEEE. doi: 10.1109/IJCNN.2018.8489513.

H. K. Kim, H. Kim, and S. Cho. Bag-of-concepts: Comprehending document representation through clustering words in distributed representation. *Neurocomputing*, 266:336–352, 2017. doi: 10.1016/j.neucom.2017.05.046.

K. Lang. 20 newsgroups dataset, 1995. URL `http://people.csail.mit.edu/jrennie/20Newsgroups/`.

Q. Le and T. Mikolov. Distributed representations of sentences and documents. In *Proc. Int. Conf. on Machine Learning*, pages 1188–1196, 2014. JMLR.org.

X. Ma, L. Wang, N. Yang, F. Wei, and J. Lin. Fine-tuning llama for multi-stage text retrieval. In *Proc. ACM-SIGIR Int. Conf. on Research and Development in Information Retrieval*, SIGIR '24, page 2421–2425, 2024. Association for Computing Machinery. doi: 10.1145/3626772.3657951.

A. Meddeb and L. B. Romdhane. Using topic modeling and word embedding for topic extraction in Twitter. *Procedia Computer Science*, 207(C):790–799, 2022. ISSN 1877-0509. doi: 10.1016/j.procs.2022.09.134.

Y. Meng, Y. Zhang, J. Huang, Y. Zhang, and J. Han. Topic discovery via latent space clustering of pretrained language model representations. In *Proc. World-Wide Web Conference*, page 3143–3152, 2022. Association for Computing Machinery. doi: 10.1145/3485447.3512034.

T. Mikolov, K. Chen, G. S. Corrado, and J. Dean. Efficient estimation of word representations in vector space. *ArXiv*, abs/1301.3781, 2013. doi: 10.48550/arXiv.1301.3781.

Tim Miller. Explanation in artificial intelligence: Insights from the social sciences. *Artificial Intelligence*, 267:1–38, 2019. doi: 10.1016/j.artint.2018.07.007.

D. Mimno, H. M. Wallach, E. Talley, M. Leenders, and A. McCallum. Optimizing semantic coherence in topic models. In *Proc. Conf. on Empirical Methods in Natural Language Processing*, page 262–272, 2011. Association for Computational Linguistics. doi: 10.5555/2145432.2145462.

F. Morstatter and H. Liu. In search of coherence and consensus: Measuring the interpretability of statistical topics. *Jour. of Machine Learning Research*, 18(1):6177–6208, 2017. ISSN 1532-4435. doi: 10.5555/3122009.3242026.

W. J. Murdoch, C. Singh, K. Kumbier, R. Abbasi-Asl, and B. Yu. Definitions, methods, and applications in interpretable machine learning. *Proc. National Academy of Sciences*, 116(44):22071–22080, 2019. doi: 10.1073/pnas.1900654116.

B. A. H. Murshed, S. Mallappa, J. Abawajy, M. A. N. Saif, H. D. E. Al-ariki, and H. M. Abdulwahab. Short text topic modelling approaches in the context of big data: taxonomy, survey, and analysis. *Artificial Intelligence Review*, 56(6):5133–5260, 2022. ISSN 0269-2821. doi: 10.1007/s10462-022-10254-w.

M. Nauta, J. Trienes, S. Pathak, E. Nguyen, M. Peters, Y. Schmitt, Y. Schlötterer, M. van Keulen, and C. Seifert. From anecdotal evidence to quantitative evaluation methods: A systematic review on evaluating explainable AI. *ACM Computing Surveys*, 55(13s), 2023. ISSN 0360-0300. doi: 10.1145/3583558.

D. Newman, J. H. Lau, K. Grieser, and T. Baldwin. Automatic evaluation of topic coherence. In *Annual Conf. North American chapter of the association for computational linguistics*, pages 100–108, 2010. Association for Computational Linguistics. doi: 10.5555/1857999.1858011.

A. Panigrahi, H. V. Simhadri, and C. Bhattacharyya. Word2Sense: Sparse interpretable word embeddings. In *Proc. Annual Meeting of the Association for Computational Linguistics*, pages 5692–5705, 2019. Association for Computational Linguistics. doi: 10.18653/v1/P19-1570.

J. Pennington, R. Socher, and C. Manning. GloVe: Global vectors for word representation. In *Proc. Conf. on Empirical Methods in Natural Language Processing*, pages 1532–1543, 2014. Association for Computational Linguistics. doi: 10.3115/v1/D14-1162.

M. E. Peters, M. Neumann, M. Iyyer, M. Gardner, C. Clark, K. Lee, and L. Zettlemoyer. Deep contextualized word representations. In *Annual Conf. North American chapter of the association for computational linguistics*, pages 2227–2237, 2018. Association for Computational Linguistics. doi: 10.18653/v1/N18-1202.

T. Prouteau, N. Dugué, N. Camelin, and S. Meignier. Are embedding spaces interpretable? results of an intrusion detection evaluation on a large French corpus. In *Proc. Language Resources and Evaluation Conference*, pages 4414–4419, 2022. European Language Resources Association.

N. Reimers and I. Gurevych. Sentence-BERT: Sentence embeddings using Siamese BERT-networks. In *Proc. Conf. on Empirical Methods in Natural Language Processing and Int. Joint Conf. on Natural Language Processing*, pages 3982–3992, 2019. Association for Computational Linguistics. doi: 10.18653/v1/D19-1410.

S. Seifollahi, M. Piccardi, and A. Jolfaei. An embedding-based topic model for document classification. *ACM Trans.Asian Low-Resources Language Information Processing*, 20(3), 2021. doi: 10.1145/3431728.

L. K. Şenel, F. Şahinuç, V. Yücesoy, H. Schütze, T. Çukur, and A. Koç. Learning interpretable word embeddings via bidirectional alignment of dimensions with semantic concepts. *Information Processing & Management*, 59(3):102925, 2022. doi: 10.1016/j.ipm.2022.102925.

S. Sia, A. Dalmia, and S. J. Mielke. Tired of topic models? clusters of pretrained word embeddings make for fast and good topics too! In *Proc. Conf. on Empirical Methods in Natural Language Processing*, pages 1728–1736, 2020. Association for Computational Linguistics. doi: 10.18653/v1/2020.emnlp-main.135.

C. Singh, A. Askari, R. Caruana, and J. Gao. Augmenting interpretable models with large language models during training. *Nature Communications*, 14(1):7913, 2023. doi: 10.1038/s41467-023-43713-1.

J. Steinberger and K. Ježek. Text summarization and singular value decomposition. In Tatyana Yakhno, editor, *Advances in Information Systems*, pages 245–254, 2005. Springer Berlin Heidelberg.

A. Subramanian, D. Pruthi, H. Jhamtani, T. Berg-Kirkpatrick, and E. H. Hovy. Spine: Sparse interpretable neural embeddings. *ArXiv*, abs/1711.08792, 2017. doi: 10.48550/arXiv.1711.08792.

S. Tanabe, M. Ohta, A. Takasu, and J. Adachi. An approach to estimating cited sentences in academic papers using Doc2Vec. In *Proc. Int. Conf. on Management of Digital EcoSystems*, page 118–125, New York, NY, USA, 2018. Association for Computing Machinery. ISBN 9781450356220. doi: 10.1145/3281375.3281391.

J. Tao, L. Zhou, and K. Hickey. Making sense of the black-boxes: Toward interpretable text classification using deep learning models. *Jour. of the American Society for Information Science and Technology*, 74(6):685–700, 2023. doi: 10.1002/asi.24642.

F. Viegas, S. Canuto, C. Gomes, W. Luiz, T. Rosa, S. Ribas, L. Rocha, and M. A. Gonçalves. Cluwords: exploiting semantic word clustering representation for enhanced topic modeling. In *Proc. ACM Int. Conf. on Web Search and Data Mining*, pages 753–761, 2019. Association for Computing Machinery. doi: 10.1145/3289600.3291032.

E. M. Voorhees and D. K. Harman, editors. *TREC Experiment and Evaluation in Information Retrieval.* The MIT Press, 2005. ISBN 9780262220736.

I. Vulić and M. Moens. Monolingual and cross-lingual information retrieval models based on (bilingual) word embeddings. In *Proc. ACM-SIGIR Int. Conf. on Research and Development in Information Retrieval*, page 363–372, 2015. Association for Computing Machinery. doi: 10.1145/2766462.2767752.

Y. Wang and Y. Zhang. Nonnegative matrix factorization: A comprehensive review. *IEEE Trans.on Knowledge and Data Engineering*, 25(6):1336–1353, 2012. doi: 10.1109/TKDE.2012.51.

C. Wu, E. Kanoulas, and M. Rijke. Learning entity-centric document representations using an entity facet topic model. *Information Processing & Management*, 57(3):102216, 2020. ISSN 0306-4573. doi: 10.1016/j.ipm.2020.102216.

X. Yang, D. Lo, X. Xia, L. Bao, and J. Sun. Combining word embedding with information retrieval to recommend similar bug reports. In *IEEE Int. Symposium on Software Reliability Engineering*, pages 127–137, New York, NY, United States, 2016. IEEE. doi: 10.1109/ISSRE.2016.33.

M. Yuan, P. Lin, L. Rashidi, and J. Zobel. Asessment of the quality of topic models for information retrieval applications. In *Proc. ACM-SIGIR Int. Conf. on Theory of Information Retrieval*, ICTIR '23, 2023. Association for Computing Machinery. doi: 10.1145/3578337.3605118.

J. Zhan, J. Mao, Y. Liu, M. Zhang, and S. Ma. Repbert: Contextualized text embeddings for first-stage retrieval. *ArXiv*, abs/2006.15498, 2020. doi: 10.48550/arXiv.2006.15498.

T. Zheng and M. Wang. Using SVD for topic modeling. *Jour.of the American Statistical Association*, 0(0):1–16, 2022. doi: 10.1080/01621459.2022.2123813.

G. Zuccon, B. Koopman, P. Bruza, and L. Azzopardi. Integrating and evaluating neural word embeddings in information retrieval. In *Proc. Australian Document Computing Conf.*, 2015. Association for Computing Machinery. doi: 10.1145/2838931.2838936.

Y. Zuo, C. Li, H. Lin, and J. Wu. Topic modeling of short texts: A pseudo-document view with word embedding enhancement. *IEEE Trans.on Knowledge & Data Engineering*, 35 (01):972–985, 2023. ISSN 1558-2191. doi: 10.1109/TKDE.2021.3073195.

Appendix A. Qualitative Mapping Performance Tables

In this Appendix we show examples of mapped and unmapped topics to illustrate the quantitative results in the body of the paper. Although strong conclusions cannot be drawn from these illlustrations, intuitively the mapped topics tend to be more interpretable (or semantically consistent) than the unmapped – noting that we have included cases even though they are counterexamples, such as topic 60 in Table 4 and topic 19 in Table 5.

Embedding	TopicID	#Feat.	Keyword Description
Doc2Vec	2	9	think, go, like, know, time, get, want, thing, come, way
	22	7	god, jesus, church, christ, sin, love, christian, bible, lord, say
	0	6	space, earth, planet, moon, launch, solar, orbit, spacecraft, system, mission
Sentence-BERT	52	8	price, sell, new, sale, offer, good, buy, include, interested, like
	62	7	game, play, period, goal, score, shot, lead, pt, espn, win
	22	6	god, jesus, church, christ, sin, love, christian, bible, lord, say
RepLLaMa	2	21	think, go, like, know, time, get, want, thing, come, way
	45	18	armenian, turkish, turk, turkey, russian, armenians, genocide, muslim, population, village
	15	8	window, problem, thank, help, run, work, know, try, use, like
Unmapped	24		orthodox, son, slipper, till, candida, italy, beast, abstract, explore, presentation
	68		keyboard, sub, rgb, virtual, thanx, custom, interior, silicon, plastic, fish
	54		dry, push, playback, evolution, depth, clean, pop, educational, reverse, quit

Table 2: Mapped and unmapped topics for the 20NG collection with LDA. The ‘Feat.’ column is the number of features matched to the corresponding topic. The ‘Unmapped’ row is three unmapped topics.

Embedding	TopicID	#Feat.	Keyword Description
Doc2Vec	8	13	smiley, manhattan, beauchaine say, blow bronx, manhattan sea, bronx away, queen stay, say queen, stay blow, sink manhattan
	4	8	gun, homosexual, people, right, gay, drug, sex, moral, think, man
	2	6	car, price, new, bike, sell, sale, engine, good, appear, buy
Sentence-BERT	8	7	smiley, manhattan, beauchaine say, blow bronx, manhattan sea, bronx away, queen stay, say queen, stay blow, sink manhattan
	2	7	car, price, new, bike, sell, sale, engine, good, appear, buy
	1	6	key, encryption, chip, clipper, government, use, message, algorithm, phone, system
RepLLaMa	0	8	game, team, play, player, win, year, season, hockey, league, hit
	30	7	battery, discharge, concrete, charge, acid, temperature, heat, lead, reaction, dirt
	1	6	key, encryption, chip, clipper, government, use, message, algorithm, phone, system
Unmapped	35		donation, copyright, shareware, john, notice, author, commercial, copy, fee, jan
	29		helmet, pocket, jacket, piece, liner, fit, pant, woman, clothing, foam
	28		mouse, driver, ball, problem, apple, mouse driver, window, roller, load, button

Table 3: Mapped and unmapped topics for the 20NG collection with BERTopic. The ‘Feat.’ column is the number of features matched to the corresponding topic. The ‘Unmapped’ row is three unmapped topics.

Embedding	TopicID	#Feat.	Keyword Description
Doc2Vec	24	8	empire, roman, emperor, war, city, army, king, battle, byzantine, rome
	14	8	philosophy, theory, knowledge, view, philosopher, school, idea, truth, nature, world
	43	6	oil, water, gas, fuel, air, increase, fire, energy, carbon, chemical
Sentence-BERT	12	5	roman, rome, empire, italy, emperor, latin, romans, greek, italian, city
	66	5	university, publish, society, science, laplace, mathematic, research, write, book, professor
	16	5	culture, cultural, people, group, country, right, human, united, social, community
RepLLaMa	45	8	language, italian, dialect, speak, latin, romance, french, vowel, word, italy
	25	6	integral, function, theory, series, theorem, set, problem, calculus, mathematic, mathematical
	24	6	empire, roman, emperor, war, city, army, king, battle, byzantine, rome
Unmapped	60		bridge, arch, build, river, stone, aqueduct, span, construct, roman, welding
	76		madhava, motto, sine, trigonometric, series, school, seal, magister, arc, acta
	48		instrument, curl, cosmetic, musical, rope, perfume, hair, hand, scallop, ensemble

Table 4: Mapped and unmapped topics for the WIKI collection with LDA. The ‘Feat.’ column is the number of features matched to the corresponding topic. The ‘Unmapped’ row is three unmapped topics.

Embedding	TopicID	#Feat.	Keywords
Doc2Vec	0	12	roman, rome, emperor, empire, city, caesar, century, senate, romans, military
	30	5	plants, plant, species, taxonomy, glossary, organisms, flora, animals, biology, botany
	3	4	italian, sardinian, language, dialect, catalan, speak, italy, calabria, occitan, romance
Sentence-BERT	0	14	roman, rome, emperor, empire, city, caesar, century, senate, romans, military
	4	7	greek, the, of, was, and, were, in, empire, greeks, to
	2	5	letter, letters, vowel, alphabet, syllable, used, greek, style, form, character
RepLLaMa	19	10	displaystyle, x2212, x2061, integral, x222b, int, function, x221e, x03c0, integrals
	0	8	roman, rome, emperor, empire, city, caesar, century, senate, romans, military
	47	8	shen, chinese, china, dynasty, confucian, confucius, han, mao, dynasty, ruler
Unmapped	56		milk, feed, livestock, dairy, animal, production, farm, feed, cow, bee
	73		babylon, assyrian, city, assyria, king, makuria, baghdad, bc, antioch, mesopotamia
	55		oil, fuel, bitumen, gasoline, sands, diesel, petroleum, asphalt, crude, refinery

Table 5: Mapped and unmapped topics for the WIKI collection with BERTopic. The ‘Feat.’ column is the number of features matched to the corresponding topic. The ‘Unmapped’ row is three unmapped topics.

Embedding	TopicID	#Feat.	Keywords
Doc2Vec	48	6	business, market, big, industry, time, go, small, cost, analyst, buy
	47	5	bush, party, political, campaign, election, democratic, president, state, candidate, vote
	40	4	like, people, time, know, day, go, come, get, man, old
Sentence-BERT	28	8	president, executive, chief, vice, officer, chairman, name, director, old, board
	24	6	share, stock, common, exchange, cent, trading, close, outstanding, inc, dividend
	46	4	drug, health, medical, care, hospital, research, patient, fda, test, study
RepLLaMa	39	12	sell, agreement, group, inc, unit, corp, sale, agree, acquire, transaction
	24	6	share, stock, common, exchange, cent, trading, close, outstanding, inc, dividend
	28	6	president, executive, chief, vice, officer, chairman, name, director, old, board
Unmapped	13		food, restaurant, franchise, chain, india, fast, spanish, taylor, knight, rice
	33		san, pacific, california, francisco, southern, santa, railroad, georgia, diego, forest
	43		insurance, steel, life, insurer, cellular, premium, policy, national, corp, industry

Table 6: Mapped and unmapped topics for the WSJ collection with LDA. The ‘Feat.’ column is the number of features matched to the corresponding topic. The ‘Unmapped’ row is three unmapped topics.

Embedding	TopicID	#Feat.	Keywords
Doc2Vec	0	6	share, say, company, million, stock, year, market, inc, new, sale
	1	5	bonds, bank, yield, treasury, rate, dollar, issue, million, price, market
	3	4	rise, increase, rate, adjust, month, year, unemployment, economist, consumer, price
Sentence-BERT	0	17	share, say, company, million, stock, year, market, inc, new, sale
	3	6	rise, increase, rate, adjust, month, year, unemployment, economist, consumer, price
	23	4	kodak, polaroid, cannon, shamrock, film, camera, tape, recorder, say, video
RepLLaMa	0	26	share, say, company, million, stock, year, market, inc, new, sale
	46	4	islam, moslem, koran, ayatollah, mosque, iran, khomeini, indonesia, religion, islamic
	1	3	bonds, bank, yield, treasury, rate, dollar, issue, million, price, market
Unmapped	38		constitution, convention, church, woman, constitutional, madison, amendment, people, document, hofmann
	31		pri, salinas, mexico, election, pinochet, party, opposition, mexican, gen, haiti
	28		south, africa, black, african, apartheid, anc, south african, government, white, south africa

Table 7: Mapped and unmapped topics for the WSJ collection with BERTopic. The ‘Feat.’ column is the number of features matched to the corresponding topic. The ‘Unmapped’ row is three unmapped topics.

Submitted 9/25; Published 12/25

Emancipatory Information Retrieval

Bhaskar Mitra BHASKAR.MITRA@ACM.ORG
Tiohtià:ke/Montréal, Canada

Editor: Djoerd Hiemstra

Abstract

Our world today is facing a confluence of several mutually reinforcing crises each of which intersects with concerns of social justice and emancipation. This paper is a provocation for the role of computer-mediated information access in our emancipatory struggles. We define emancipatory information retrieval as the study and development of information access methods that challenge various forms of human oppression, and situates its activities within broader collective emancipatory praxis. The term "emancipatory" here signifies the moral concerns of universal humanization of all peoples and the elimination of oppression to create the conditions under which we can collectively flourish. To develop an emancipatory research agenda for information retrieval (IR), in this paper we speculate about the practices that the community can adopt, enumerate some of the projects that the field should undertake, and discuss provocations to spark new ideas and directions for research. We challenge the field of IR research to embrace humanistic values and commit to universal emancipation and social justice. We also invite scholars from fields such as human-computer interaction, information sciences, media studies, design, science and technology studies, social and political sciences, philosophy, law, environmental sciences, public health, educational sciences, as well as legal and policy experts, civil rights advocates, social justice activists and movement organizers, and artists to join us in realizing this transformation. In this process, we must both imagine post-oppressive worlds, and reimagine the role of IR in that world and in the journey that leads us there.

Keywords: IR and society, Emancipatory praxis, Technology and power

1 Introduction

> "The old world is dying, and the new world struggles to be born: now is the time of monsters."
>
> – Antonio Gramsci

Our world in crises. Our world today is facing a confluence of several mutually reinforcing crises pushing us universally, but notably not uniformly, towards precarity. The recent COVID-19 pandemic did not only result in over seven million deaths (Mathieu et al., 2024) and disrupt everyday lives globally; it also laid bare and further reinscribed existing political, economic, and social hierarchies that are saliently colonial and racialized (Papamichail, 2023), gendered (Carli, 2020; Madgavkar et al., 2020), casteist (Mondal and Karmakar, 2024), and ableist (Lund and Ayers, 2022). Likewise, the dangers of recent and impending climate catastrophes as well as the discourse around mitigative actions are inextricably intertwined with concerns of racial (Sultana, 2024; Williams, 2022; Patnaik et al., 2020; Stefani

WWW: https://irrj.org DOI: 10.54195/irrj.24531

et al., 2021), gender (Gender and Alliance, 2016; Women, 2022), and disability (Cram et al., 2022) justice. Increasing global wealth and income inequality (Chancel et al., 2022) both mirrors historical dimensions of marginalization and is contributing to erosion of democracies worldwide (Rau and Stokes, 2025). Globally, we are still reckoning with the ongoing legacies of settler-colonialism (Wolfe, 2006) and present-day neocolonialism (Nkrumah, 1965), recontesting women's rights to bodily autonomy (Beck et al., 2024), and witnessing growing assaults on transgender people's right to exist (Foundation, 2023; Horbury and Yao, 2020).

The range of crises we face in the twenty-first century requires us to reckon with multiple intersecting and co-constituting processes of hierarchical organization of our society and as a consequence with the systemic dehumanization, marginalization, and oppression of those designated to its bottom rung. Therefore, any attempt at addressing the crises of our times must begin with recommitment to the principles of social justice and universal humanization of all peoples and support for emancipatory struggles against social, economic, political, cultural, religious, sexual, and bodily oppression. This paper is a provocation for the role of computer-mediated information access in these struggles. Our aim is to challenge the field of information retrieval (IR) research to reimagine itself as what it can (or, ought to) be if grounded in humanistic values and committed to universal emancipation and social justice, and to encourage the community to explicitly articulate and make conscious choices about the values and incentives that shapes our field of computing. This is also an invitation to scholars from fields such as human-computer interaction (HCI), information sciences, media studies, design, science and technology studies (STS), social and political sciences, philosophy, law, environmental sciences, public health, educational sciences, as well as legal and policy experts, civil rights advocates, social justice activists and movement organizers, and artists to join us to collectively reimagine radically new sociotechnical futures and realize desired transformations.

The information reaches out. IR, the field of computing that is concerned with the design and study of information access systems, such as web search and recommender systems, has its roots in library science and linguistics (Harman et al., 2019). Early IR research was primarily concerned with the design of effective indexing mechanisms that enable fast pattern-based lookup of text from large corpora. The modern-day successors of these systems, however, go far beyond serving the function of simple lookup. Modern search systems observe and model collective behaviors of how their users interact with retrieved information to infer intent behind user queries, and to predict users' future needs and actions. Some of these systems may proactively surface new information and breaking news, recommend commercial products and services, and even suggest activities and actions for the user to undertake. Conversational systems, such ChatGPT,[1] Microsoft Copilot,[2] and Google Gemini[3], may be entrusted to summarize retrieved information in response to users' search queries, effectively shifting the burden of assessing the relevance and trustworthiness of retrieved information and their contextual interpretation from the user to the statistical models underlying these systems. These systems affect what information and perspectives receive exposure which shapes consumer behavior, political discourse, and culture (Grimmelmann,

[1] `https://chat.openai.com/`

[2] `https://copilot.microsoft.com/`

[3] `https://gemini.google.com/`

2008; Gillespie, 2019; Hallinan and Striphas, 2016), and directly influence material outcomes for people (Singh and Joachims, 2018; Datta et al., 2014).

Today's search and recommender systems are, therefore, much more than passive lookup tools. These systems have become embedded in our online social discourse, both shaping and being shaped by our values, cultural identities, sociopolitical beliefs, and shared understanding of our environment and ourselves. We could choose to see this transformation of information access systems from passive lookup tools to agents of influence as the information reaching out and permeating every aspect of our individual and social lives—akin to how Grudin (1990) described the transition of computers from applications in narrow specialist domains to their present-day ubiquitous presence in our lives as "the computer reaches out".

Importantly, these systems and platforms may privilege certain perspectives and ways of knowing over others as forms of epistemic injustice (Fricker, 2007) that entrench social hierarchies and intensify oppression. Or, alternatively they could serve to raise up marginalized voices and challenge the status quo. Access to information is critical to collective sensemaking of our place and relationships in this world. Therefore, it is unsurprising that throughout history authoritarian forces have tried to control what information is disseminated and how. In both historical and ongoing conflicts, we can observe these struggles for control over both traditional media (*e.g.*, newspapers, radio, and television) and online media (*i.e.*, social media and other online information access platforms). For the oppressed, they serve as medium for building shared understanding and solidarity necessary for social movements and collective action. For the oppressor, they represent sites to apply force to censor information and perspectives and manipulate public opinion. Information access platforms, therefore, represent sites of conflict between exploitation and liberation. *How should we then reflect on and reconcile the implications of IR research and development on our communities and our collective liberation?* In this paper, we assert that the field of IR would benefit from critically reflecting on how our work relates to current sociopolitical orders and our possible sociotechnical futures, and argues for reimagining IR research grounded explicitly in the adoption of humanistic and emancipatory values and aspirations.

2 What is emancipatory IR?

We define *emancipatory IR* as the study and development of information access methods that challenge all forms of human oppression, and situates its activities within broader collective emancipatory praxis. The term "*emancipatory*" here signifies the moral concerns—central to this field of study—of universal humanization of all peoples and the elimination of oppression to create the conditions under which we can collectively flourish. Historically, the term "emancipation" has taken on various meanings in context of diverging intellectual traditions (Susen, 2015). It described the abolitionist struggles against slavery, but later came to signify a broader vision of dismantling all forms of structural oppression. Examples of structural oppression in this context include colonialism, racism, patriarchy, casteism, transphobia, religious persecution, and ableism. Emancipatory thinking in humanities have manifested in many forms, including Marxism, Critical Theory, feminism, and decolonialism. Emancipatory IR aims to be informed by and incorporate these different theories and epistemic frameworks in the conception and design of information access methods and systems.

This framing recognizes information and access to information both as profoundly political, and enacts research and development activities—such as, formalization, theorization, design, experimentation, publishing, open sourcing, deployment, platform governance, and community building—in service of universal struggles against all forms of structural oppression. It rejects the techno-deterministic premise that there is a single pre-determined path forward for information access technology development, and challenges the community to employ reflexivity to uncover the deeply embedded values, incentives, and sociotechnical imaginaries that shape the field of IR. It recognizes the agency of the field and its contributing members in shaping the world, while respecting that the challenges and necessary interventions are both fundamentally sociotechnical in nature. Therefore, the theories and practices of this field must be co-developed with cross-disciplinary collaboration including with scholars from HCI, design, information sciences, media studies, STS, social and political sciences, philosophy, law, environmental sciences, public health, educational sciences, as well as with legal and policy experts, civil rights advocates, social justice activists and movement organizers, and artists, among others.

Emancipatory IR challenges us to employ technological research and development in service of dismantling hierarchies of oppression and challenge power. It discourages non-performative academic gaze, and urges this research to be situated in movement building, and calls for recognizing the role of movement building practices within this research. It encourages us to prioritize praxis—*i.e.*, research activities and reflections directed at structural change—over proxy metrics of success—*e.g.*, state-of-the-art performances and leaderboard rankings that do not translate to scientific or social progress). And above all, it challenges the field to move beyond the restrictive view of IR research that emphasizes algorithmic advances as measurable on shared benchmarks and leaderboards to a more expansive view of IR research situated in a project of affecting social good and universal emancipation.

Emancipation research in other fields. To develop an emancipatory IR research agenda, we can take inspiration from other scholarly fields that have incorporated humanistic values and emancipatory aspirations. Young et al. (2021) surveyed the body of emancipation research in information science, and identified four components of emancipation prevalent in the literature: Agency (*i.e.*, freedom to act), dialogue (*i.e.*, freedom to express), inclusion (*i.e.*, freedom to belong), and rationality (*i.e.*, freedom to think). Wright (2020) defines emancipatory social science as an intellectual enterprise seeking to "generate scientific knowledge relevant to the collective project of challenging various forms of human oppression". Wright outlines a framework that enumerates three specific tasks for the field: (i) Systematic diagnosis and critique of the world as it exists, (ii) envisioning viable alternatives, and (iii) elaborating a theory of social transformation. According to Wright, emancipatory social science is "a theory of a journey from the present to a possible future: the diagnosis and critique of society tells us why we want to leave the world in which we live; the theory of alternatives tells us where we want to go; and the theory of transformation tells us how to get from here to there – how to make viable alternatives, achievable".

In HCI, Bardzell and Bardzell (2016a, 2015) discuss emancipation under the heading of humanistic HCI, which they define as research and practices in the field that deploy humanistic epistemologies and methodologies. They note that emancipatory HCI is "a fundamental goal of virtually all humanistic HCI". Related, several strands of HCI research incorporate epistemologies, theories and practices that are anti-oppressive and emancipatory (Smyth and Dimond, 2014), feminist (Bardzell, 2010; Bardzell and Bardzell, 2011; Bardzell et al., 2011; Bardzell and Bardzell, 2016b; Bardzell, 2018), queer (Light, 2011; Klipphahn-Karge et al., 2024), postcolonial and decolonial (Irani et al., 2010; Dourish and Mainwaring, 2012; Sun, 2013; Akama et al., 2016; Irani and Silberman, 2016), anti-racist (Abebe et al., 2022), anti-casteist (Vaghela et al., 2022a,b), anti-ableist (Williams et al., 2021; Sum et al., 2024), post-capitalistic (Feltwell et al., 2018; Browne and Green, 2022), and anarchist (Keyes et al., 2019; Linehan and Kirman, 2014).

Similarly, in the fields of machine learning (ML) and artificial intelligence (AI) there are several works that call for perspectives and design interventions that are anti-oppressive and emancipatory (Kane et al., 2021; Saxena et al., 2023), anti-fascist (McQuillan, 2022), decolonial (Adams, 2021; Mohamed et al., 2020), anti-casteist (Kalyanakrishnan et al., 2018; Sambasivan et al., 2021), and abolitionist (Benjamin, 2019; Barabas, 2020; Earl, 2021; Williams and Haring, 2023). In data science, D'ignazio and Klein (2020) and Guyan (2022) discuss how feminist and queer epistemologies can inform the field, and Monroe-White (2021) adopted Wright's framework (Wright, 2020) to propose an emancipatory agenda for data science to mitigate the harms to marginalized populations.

Towards emancipatory research in IR. Belkin and Robertson (1976) acknowledged information science's social responsibility nearly half a century ago. But it was not until more recently that this perspective gained serious traction within the IR community, starting with the SWIRL (Culpepper et al., 2018) and the FACTS-IR (Olteanu et al., 2019) workshops. Subsequently, there has been a wide range of IR research on fairness (Ekstrand et al., 2021), explainability (Anand et al., 2022), and addressing misinformation (Zhou and Zafarani, 2020) among other societally-motivated topics. However, the field has been largely reluctant to acknowledge the saliently political nature of this work leaving the underlying colonial, cisheteropatriarchal, and capitalist values that has (and continue to) critically shape the field of IR unchallenged. Even IR conference tracks dedicated to research that tries to affect social good have often side-stepped the deeply sociopolitical question of defining what constitutes social good (Mitra and Heuss, 2025).

This has led to several recent calls to explicate and critically examine the norms and values (Vrijenhoek et al., 2023; Trippas and Culpepper, 2025) shaping the field as well as the sociotechnical futures that the IR community wants to realize through their research (Mitra, 2025; Azzopardi et al., 2024); while also reasserting the interdisciplinary roots of IR (Zangerle et al., 2025). This year, the ECIR IR-for-Good track's call for papers[4] explicitly define IR-for-Good as "*IR research and practices that contribute towards realizing more equitable, emancipatory, and sustainable futures*" (Mitra and Heuss, 2025), and updated the list topics relevant to the track to shift the emphasis from desired attributes and interventions in IR systems (*e.g.*, fairness and accessibility)—that were the norm in previous years[5]—to desired justice-oriented real-world outcomes (*e.g.*, racial, disability, and sexuality justice) that IR

[4] `https://ecir2026.eu/calls/call-for-ir-for-good-papers`

[5] `https://ecir2025.eu/call-for-ir-for-good-papers/`

research should try to realize. They also require papers to include explicit discussion on their theory of change.

Several of these developments—*i.e.*, Trippas and Culpepper (2025); Azzopardi et al. (2024); Mitra and Heuss (2025)—were prompted by conversations based on an earlier preprint of this current manuscript. In this work, we argue that moving forward the IR community should adopt emancipation as an explicit objective for the field.

3 Practices, projects, and provocations

To develop an emancipatory research agenda for IR, we speculate about the practices that the community can adopt, enumerate some of the projects that the field should undertake, and discuss provocations to spark new ideas and directions for research. However, the list of practices, projects, and provocations discussed here are neither immutable nor complete. It is ultimately the responsibility of the community engaged in emancipatory research to regularly reflect on, produce evidence-based critique, challenge, and reshape its agenda to affect tangible progress in our collective struggles against dehumanization, oppression, and marginalization. It may be worth reemphasizing at this stage that due to the saliently sociotechnical nature of the challenges in this area, the practices, projects, and provocations too will rarely be purely technological or reside strictly in the domain of computing, and must be informed by perspectives from outside of the field of computer science and the lived experiences of marginalized peoples. We illustrate the practices, projects, and provocations described in this section in Figure 1.

3.1 Practices of emancipatory IR

We adapt Wright's emancipatory social science framework (Wright, 2020) to develop a framework for the practices of emancipatory IR. Like Wright's, our framework consists of diagnosing and critiquing, imagining viable alternative futures, and elaborating our theories of change.

Diagnose and critique. The starting point for emancipatory IR research is identifying the ways in which existing IR methods and systems, the cost of IR research and development, and the arrangements within the IR community may contribute systemic harms to peoples, act contrary to aspirations of social and political justice, and impede emancipatory struggles of the peoples. These harms may include:

- Suppressing voices of the oppressed and marginalized
- Enabling surveillance and public opinion manipulation
- Imposing representational harms through derogatory portrayal, stereotyping, or erasure in presentation of retrieved results
- Imposing allocative harms to groups through under-exposure of their content in retrieved results or under-exposing groups to socioeconomic opportunities
- Exploitation of ecological resources and labor for IR research and development that mirror racial capitalism and coloniality

- Capture of IR research by the military–industrial complex that pushes the field to disproportionately prioritize economic and military interests over concerns of knowledge production, public health education, and information literacy, and produce technologies that serve as tools of oppression (*e.g.*, for surveillance)

- Concentration of extreme power and wealth among few individuals and institutions in ownership of popular information access platforms

- Lack of representation and diversity within the IR community leading to propping up colonial, cisheteropatriarchal, and capitalist values and erasure of feminist, queer, decolonial, anti-racist, anti-casteist, anti-ableist, and abolitionist perspectives

Diagnose and critique must extend to the policy and regulatory landscape, such as analyzing the European Union's General Data Protection Regulation (GDPR) (Regulation, 2016) and Digital Services Act (Regulation, 2022) in the context of our emancipatory goals which we identify as critical future work.

Imagine viable alternative futures. The second practice of emancipatory IR is to imagine and develop desirable, viable, and achievable alternatives to our IR technologies and ways of organizing our communities with the aim of addressing the harms and injustices identified in the diagnosis and critique stage. Previous work (Mitra, 2025) have argued that the IR community should explicitly articulate the sociotechnical imaginaries (Jasanoff and Kim, 2009, 2015) that influence the design and development of IR technologies and platforms; and enquires "what would IR systems look like if designed for futures informed by feminist, queer, decolonial, anti-racist, anti-casteist, anti-ableist, and abolitionist thoughts, and if the focus of IR research was not to prop up colonial cisheteropatriarchal capitalist structures but to dismantle them?" This is a call for us to imagine post-oppressive worlds, and the role of IR both in that world and in the journey that leads us there.

We intentionally do not speculate about any specific futures in this paper. To satisfy that ask, IR needs spaces and processes for design futuring (Fry, 2009). The goal of futuring here is not just to envision a plurality of alternative utopian and protopian futures, but also to build shared understanding, aspirations, and commitment towards emancipatory outcomes through that participatory process. After all, the act of exercising one's imagination itself can be liberating, or as Benjamin (2024) calls it "an invitation to rid our mental and social structures from the tyranny of dominant imaginaries".

Elaborate theories of change. Our final practice concerns with the development of theories of change (Weiss, 1995; Brest, 2010; Taplin and Clark, 2012; Wikipedia contributors, 2013) that makes explicit how we aim to realize the alternative futures in the face of current realities and how our own work contributes towards those goals. This should be grounded in theories of power and co-developed with cross-disciplinary scholars, legal and policy experts, activists, artists, and others to realize real structural changes. This involves both developing new research agendas as well as experimenting with and realizing new arrangements within the IR community and critically reflecting on our relationships with other institutions (*e.g.*, industry and government). To provide concrete examples, we next discuss potential projects that we believe emancipatory IR should invest in.

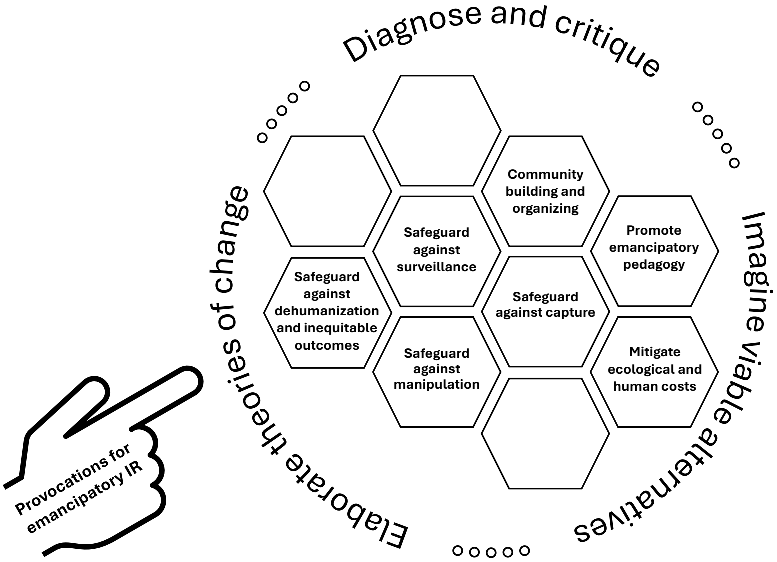

Diagnose and critique
Imagine viable alternatives
Elaborate theories of change
Community building and organizing
Safeguard against surveillance
Promote emancipatory pedagogy
Safeguard against dehumanization and inequitable outcomes
Safeguard against capture
Safeguard against manipulation
Mitigate ecological and human costs
Provocations for emancipatory IR

Figure 1: Illustrating the practices, projects, and provocations of emancipatory IR. The hexagons at the center represent several potential projects. The empty hexagons indicate that the set of mentioned projects are not complete, and the community should over time identify new ones and shape existing ones, as appropriate. Surrounding the projects is the framework of practices of emancipatory IR that are relevant to all projects in this area, consisting of (i) diagnose and critique, (ii) imagine viable alternative futures, and (iii) elaborate theories of change. Provocations to spark new ideas and research directions are depicted as a nudging hand.

3.2 Emancipatory projects in IR

Next, we enumerate some of the potential projects relevant to emancipatory IR. Such examples are useful for readers to get a more concrete sense of the aspirations and type of work relevant to this field, and can also serve as a starting point for developing a more thorough research agenda. Ultimately, the direction and what specific project we invest in must be decided and iterated on over time by those who see themselves as part of this community, which we must ensure includes more than just IR researchers and technologists as we emphasized previously.

Safeguard against capture. A central concern of emancipatory IR should be to ensure that our information access platforms cannot be easily captured by few privileged individuals and institutions or authoritarian regimes, and also to ensure that access to information services cannot easily be blocked by authoritarian regimes in times of protests to suppress dissent. These concerns and initiatives are timely in the face of the 2024 edition of the World Economic Forum's Global Risks Report (Forum, 2024) that calls out "technological power concentration" as one of the top global risks for the upcoming decade, and broader growing concerns about the alignment of tech industry with authoritarianism and fascism (Kang, 2025; Gebru and Torres, 2023; Varoufakis, 2024; LaFrance, 2024; Duran, 2024; Akbari, 2025). In the social media landscape, there have been recent announcements (Mastodon, 2025; Bsk, 2025) of formation of new non-profit foundations for managing key social media ecosystems and platforms, specifically with the goal to safeguard against billionaires' control over our digital public squares. In IR, there has also been recent investments to create open infrastructures for internet search, such as the Open Web Search project.[6]

While decentralization of IR infrastructure and federation may be critical for protecting against capture, it is also important to acknowledge that in a world shaped by white supremacy, colonialism, patriarchy, cisheteronormativity, neuronormativity, ableism, and casteism as organizing sociopolitical forces, the decentralization of the technological infrastructures in themselves do not constitute sufficient conditions for creating safe and welcoming online spaces for historically marginalized peoples nor to ensure these platforms align with our emancipatory aspirations (Hendrix and Flowers, 2022). Instead, the challenges that we face require both decentralized technology infrastructure, as well as social innovations in the formation of new global governance structures for the ownership, management, and moderation of these platforms to ensure that the voices of the oppressed and marginalized are uplifted, not suppressed.

Safeguard against surveillance. IR applications—such as web search, recommender systems, and targeted digital advertising—have been major drivers of surveillance capitalism (Zuboff, 2019). An important commitment IR can make towards anti-oppression is to ensure that the field is neither creating the technologies and infrastructures for surveillance, nor creating additional demands for such infrastructures out of commercial interests. User behavior data have been the "secret" sauce for many successful commercial IR platforms. For example, in search, user interactions with search result pages are continuously logged and modeled at incredible scale to provide better search relevance to customers. It is therefore not surprising that modeling of user clicks (Chuklin et al., 2015) and even cursor movements (Diaz et al., 2013) have historically garnered substantial interest in IR research. However, this creates demands for unscrupulous surveillance of users on these platforms. In response, different regulatory safeguards have been proposed, including notably the principles of data minimization defined in Article 5 of the GDPR (Regulation, 2016).

To safeguard against surveillance, the IR community must imagine alternative ways of building its systems that does not necessitate such large-scale data collection to simply make these platforms usable. We must both push the boundaries of effective retrieval methods in the absence of user behavior data *and* develop new protocols for collection, storage, and application of user data that implements stricter safeguards for user privacy and require stricter and more meaningful user consent for access by the platforms. We need to explore both technological innovations (such as, homomorphic encryption) as well policy safeguards and changing social norms around data stewardship and transparency of how institutions collect, store, and use data, and implement concrete safeguards to ensure that authoritarian governments and their institutions cannot get access to this data which may enable harassment and intimidation of activists, journalists, and marginalized populations. Lastly, we should be wary of potential dual-use of IR technologies—such as the impact of improved

[6] `https://openwebsearch.eu/`

cross-lingual search on surveillance of foreign individuals, institutions, and governments—and be critical of and refuse funding for such research that comes from governments and security agencies.

Safeguard against manipulation. Misinformation and disinformation pose serious threats to democracies worldwide (Lewandowsky et al., 2023; Ecker et al., 2024; Lewandowsky, 2024). The 2024 edition of the World Economic Forum's Global Risks Report (Forum, 2024) ranks misinformation and disinformation as the top global risk for the subsequent two years. A study spanning six years, 26 countries, and several election periods find that political misinformation is particularly salient in current wave of radical right populism and its opposition to liberal democratic values and institutions (Törnberg and Chueri, 2025). Curtailing spread of online misinformation and disinformation that aims to dehumanize immigrants, the colonized, the poor, trans and nonbinary, and racially marginalized peoples, is critical to our social justice efforts and our effort to build solidarity with each other.

Beyond misinformation, we must also pay close attention to new vectors of public manipulation that may become plausible in near future based on our current technological trends. A particularly worrying development in this area is the recent emergence of "persuasive AI" (Burtell and Woodside, 2023; Carroll et al., 2023; Park et al., 2023; El-Sayed et al., 2024). The massive trove of detailed data on user behavior and preferences combined with the capabilities of generative AI to produce persuasive language and visualizations could create tools of mass manipulation and pose serious risks to functioning of global democracies (Mitra et al., 2024). Imagine every time you searched online or accessed information via your digital assistant, the information was presented to you exactly in the form most likely to alter your consumer preferences or political opinions. Or, consider AI-generated digital characters in ads and videos that appropriate marginalized identities to say or act in ways that real members of that community may be strongly opposed to—a new form of "Digital Blackface" (Johnson and Joe, 2023). Emancipatory IR researchers must engage in evidence-based critique of such technology development and push for policies and community norms that push back against research in support of such applications within the field before more such applications of AI materialize and are normalized.

Safeguard against dehumanization and inequitable outcomes. Emancipatory IR should concern itself with ensuring that IR systems do not lead to representational harms (*e.g.*, denigration, stereotyping, and erasure) of historically marginalized peoples and inequitable outcomes which may further reinscribe historical sociopolitical and economic oppression and exploitation. Over the last several years, there has been several works on fairness in IR. These have typically focused on fairness of quality-of-service—*e.g.*, (Mehrotra et al., 2017, 2018; Neophytou et al., 2022; Wu et al., 2024)—and fairness of exposure—*e.g.*, (Biega et al., 2018; Singh and Joachims, 2018, 2019; Diaz et al., 2020; Zehlike and Castillo, 2020; Patro et al., 2020; Wu et al., 2022). However, there is less clarity on how much this body of fairness research has practically influenced material changes in popular IR platforms (Mitra, 2025). Fairness research in the emancipatory context must move beyond just formalization and theorization of fairness; it must recognize and be grounded in the understanding of historical injustices and structural violence, and must ensure that this research opposes dehumanization and material inequities in the real world. This includes also contending with the politics of classification (Crawford, 2021) and critiquing practices in fairness research, such as in context of gender fairness (Pinney et al., 2023), that may themselves perpetuate further harm.

We must also hold institutions that own and operate IR platforms accountable to measurable equitable and humanistic outcomes. If an institution invests in fairness research but does not operationalize them in their products, then that research only serves to launder institutional reputation and is counter to our real emancipatory goals. To hold institutions accountable, our research agenda should also include challenges of auditing IR platforms and also explore other regulatory, sociopolitical, and technological mechanisms of accountability that are grounded in the recognition of the structural power inequities that make it challenging to do so.

Promote emancipatory pedagogy. Emancipatory IR must abandon any false notions of neutrality between the oppressor and the oppressed, and develop frameworks situated in theories of power to uplift marginalized voices and make spaces for emancipatory pedagogy. It must expressly refuse to serve as a soapbox for the powerful and wealthy. To make this further challenging, we must do so without concentrating the power to define oppressor-oppressed relations in the hands of platform owners. Can we, for example, learn from Freire's (Freire, 2020) anti-oppressive pedagogy to imagine IR systems that do not merely retrieve but provide spaces for dialogical interactions between information seekers? Can we build communities of experts who can provide context from critical scholarship to inform search results and aid in our collective sense-making and solidarity-building on topics relevant to our emancipatory struggles? What if we go beyond fixing under-representation of marginalized groups in, say, image search results for occupational roles and reclaim those digital spaces as sites of resistance and emancipatory pedagogy (Mitra, 2025)?

Emancipatory IR must also critically interrogate the relationship between information access platforms and the publishers and news media. Kim et al. (2024) find media indoctrination as one of the strongest predictors of autocratic survival. In light of these dynamics, we must reexamine how platforms can support local newsrooms that often do critical work in reporting on issues relevant to marginalized populations. In this context, support must include both ensuring these publishers get exposure in search results as well as innovations in platform business models that can support local newsrooms financially, such as through revenue sharing.

Visibilize and mitigate ecological and human costs of IR research and development. We must also be concerned of the cost of doing IR research and operating IR platforms. This includes ecological costs (Scells et al., 2022; Zuccon et al., 2023), such as energy and water consumption and harmful emissions, as well as the appropriation of data labor and impact on the wellbeing of workers. Belkhir and Elmeligi (2018) estimate that Information and Communications Technology industry on the whole will account for 14% of global emissions by 2040. In the US, data centers are projected to consume around 6% of the total national electricity by 2026 (Halper, 2024). Appropriation of data labor includes both uncompensated appropriation of works by writer, artists, and programmers (Cohan, 2023; Coldewey, 2023; Vincent, 2023, 2022; Shrivastava, 2023; Burke, 2023, 2024; Vincent and Li, 2021; Vincent, 2020; Appel et al., 2023; Marr, 2023; Chesterman, 2024; Chayka, 2023; Gertner, 2023; Vincent and Li, 2023) and under-paid crowdwork for data labeling (Perrigo, 2023; Williams et al., 2022; Tan and Cabato, 2023; Altenried, 2020; Hao and Hernández, 2022; Xiang, 2023; Hao and Seetharaman, 2023). The latter mirrors historical patterns of global labor exploitation and racial colonial capitalism (Hao, 2022; Birhane, 2020; O'Gorman, 2023; Klein, 2023; Couldry and Mejias, 2019; Muldoon and Wu, 2023; Tacheva and Ramasubramanian, 2023), and have been associated with severe mental health cases among crowdworkers (Booth, 2024; Hao and Seetharaman, 2023). Other concerns include land use for data centers and how that affects communities living on those lands and indigenous sovereignty (Rubayita, 2025; Verma, 2024).

Emancipatory IR must invest in accurate accounting, reporting, and reduction of the ecological and human cost of our research. This includes the cost of training models and experimentation, the cost of human annotations, and the cost of air travel for conference attendance, among others. It includes a commitment towards not pursuing research directions whose cost include accelerating climate change and broader adoption of virtual conferencing to minimize air travel. It also includes the recognition that these costs cannot just be minimized by building more efficient models—*e.g.*, because of Jevons paradox[7]—or automated labeling methods, and that the solution as always is at the intersection of the social, the political, the economic, and the technological.

Community building and organizing. The work of building an emancipatory research agenda must start with building an emancipatory research community. We need to create safe spaces where researchers from the IR community can engage and meaningfully interact with a diverse set of scholars, designers, legal and policy experts, activists, and artists to co-develop an agenda for this field. We must reflect on the current social and organizational arrangements within the field and our relationships with other institutions, such as industry and government. We must take on the difficult challenge of critiquing the funding mechanisms that enable IR research and its role in privileging certain areas of research explorations and technology development that are counter to our emancipatory aspirations and social justice. We must reimagine these arrangements, build counter-structures, and establish new means of funding emancipatory IR research that can sustain itself in the face of attack from the privileged, powerful, and wealthy. We must nurture a culture in our community that recognizes and encourages direct actions to challenge power, such as demonstrating and whistleblowing (Hicks, 2025), and create support mechanisms to protect individuals from reprisal from those whose power and visions we want to challenge. We must ensure that our community is inclusive and representative of marginalized peoples and that we view their lived experiences with structural oppression as critical to informing research in our field. And we should view emancipatory IR as situated in broader movement building against marginalization, dehumanization, and oppression.

3.3 Provocations

Emancipatory IR must leverage instruments of provocation to challenge our thinking and help us reimagine radically alternative futures. Expanding on previous call (Mitra, 2025) to safeguard the IR community from falling victim to crises of imagination (Haiven, 2014), we posit that the inception of emancipatory IR must not be in a vacuum, but must be informed by emancipatory ideologies borne of our historical and ongoing struggles, including feminist, queer, decolonial, anti-racist, anti-casteist, anti-ableist, and abolitionist thoughts.

[7] `https://en.wikipedia.org/wiki/Jevons_paradox`

Our challenge is not only to imagine radically new emancipatory futures, but also to put into practice those desired values and build prefigurative counter-structures both in the form of new technological designs and sociopolitical rearrangements within our communities. The task to change technology is inextricably linked to the task of changing ourselves and our own social arrangements. The practice of reflexivity becomes important in this context to articulate (and critique if necessary) our own values, positionality, and incentives, as well as our relationships with institutions and social structures that we in turn intend to change. It is in this process that we desperately need to engage scholars outside of computing, learn from generations of activists and movement organizers who have been engaged in emancipatory struggles, and let ourselves be influenced by the artists and writers that dare to imagine radically free futures and humanistic social structures.

4 Discussion

4.1 The *emancipatory* in emancipatory IR

For emancipatory IR to be authentically emancipatory, this work must not be simply limited to an intellectual exercise. Our scholarship must be a piece of a broader movement building praxis grounded in ongoing struggles in our world against marginalization, dehumanization, and oppression. Our critiques must be followed by a tangible push for dismantling oppressive structures and building counter-structures in its place. Our theories must be put into practice and tested in the real world, and our success measured in concrete emancipatory outcomes. Access to information is critical to our collective sense-making of our place in this world and our relations with it, and it is exactly why throughout history, from print media to present-day social networks, authoritarian forces have tried to control what and how information is disseminated. Emancipatory IR stands in solidarity with the oppressed and their struggles, and accepts the challenge of building the information infrastructure for resistance.

4.2 The *IR* in emancipatory IR

But what makes emancipatory IR truly *IR*? How are questions of emancipation relevant to this field of computing whose research is typically concerned with questions of ranking, evaluation, efficient system design, language modeling, human information interaction, and other similar topics? To this we respond by challenging the reader to reflect on how each of these important concerns in IR intersect with the exemplar projects we discussed in Section 3.2. What new research questions emerge in context of indexing and ranking when our IR platforms are federated and decentralized? What new research challenges emerge in human information interaction if our goal is to promote emancipatory pedagogy? How do we mitigate the harms from appropriation of data labor that is so very prevalent in our modern-day language modeling approaches? How do we reimagine efficiency and evaluation when our aspirations go beyond fetching information artifacts at lightning speed to encouraging deeper collective understanding and new knowledge production that are key to our social progress? With these questions our aim is to challenge the field to abandon a more restrictive view of what IR research is which may limit itself to algorithmic advances as measurable on shared benchmarks, and adopt a more expansive view of IR that aims to situate information access as a vehicle for social good. These aspirations are also reflected in the recent SWIRL 2025 report (Trippas and Culpepper, 2025) that calls for centering emancipatory values in IR research.

4.3 Where do we begin?

Bootstrapping an emancipatory IR research agenda will be challenging. The process requires existing IR researchers to expand their horizons of concerns and expertise, bring in others from outside of the field experienced in emancipatory struggles, and consider new assemblages of both its technologies and sub-communities. While there may be several potential starting points for this process, we recommend that the work should begin by pulling together a seed community interested in this research direction. Towards that goal, one proposal would be to organize community-building workshops whose structure is informed by the practices, projects, and provocations discussed in Section 3. A sketch of such a workshop may look as follows:

1. A series of talks covering critiques and provocations relevant to emancipatory IR.
2. A design futuring session where participants are encouraged to radically reimagine IR systems and how they should be embedded in our society and our daily lives.
3. Following the futuring exercise, participants sketch out paths to our desired futures and identify concrete challenges and initiatives that the emancipatory IR community should take on. Some of these may align with the projects enumerated in Section 3.2 or may identify the need for additional projects that we have not considered here.
4. Finally, the participants must discuss funding considerations to sustain our research agenda, as well as steps that we can take to protect our community members from reprisal from whose power our research aims to challenge.

The goal of such workshops would be to build trust, solidarity, and a shared understanding of our aspirations and plans within the seed community. It is also our aim with these workshops to challenge the seemingly unassailable wall of colonial capitalist values that have arrested our present visions of society and technologies. We hope such a workshop can be successful in finding the cracks in the wall through which we can collectively peep out and free ourselves to imagine a plurality of social arrangements that are possible in which we can all flourish free from oppression.

5 Conclusion

The field of IR has undergone seismic shifts over the last decade largely instigated by rapid developments in deep learning (Mitra and Craswell, 2018) and more recently generative AI (White and Shah, 2025). Between trying to stay updated in the face of daily firehoses of new AI publications and chasing the coveted badges of “state-of-the-art” it can be easy as an IR researcher to overlook the pressing needs of society and envision only the possibilities as conjured by the latest AI news cycle. If we are to address the multitude of crises facing our world today—spanning the ecological, the social, the political, and the economic spectrums—IR must undergo another seismic shift over the next decade. This piece is a provocation for the IR community to situate itself in the generational emancipatory struggles and stand in resistance to the growing epidemic of authoritarianism and fascism through the practices of making and unmaking of information access technologies. It is also an invitation to scholars from other fields as well as to social movement organizers and artists to join us and be part of this shared struggle. And as IR researchers, we must remember that "it is vital that we approach these spaces with curiosity and humility; in recognition of our own incomplete understanding of the world; open to change and be changed by these encounters" (Mitra, 2025).

A central thesis underlying this work is that the emancipatory struggles in the real world and one that we are encouraging the IR community to acknowledge and center within our research are one and the same. So, the work of emancipatory IR and the work of movement building for social justice and emancipation outside of IR must co-construct the conditions for dismantling oppressive regimes, such as those of colonialism, cisheteropatriarchy, and capitalism. While focused on IR, it is also a challenge to the broader computing and technology community to reimagine our work to not serve the privileged few by further concentrating immense power and wealth, but as a liberatory force that connects us and humanizes us universally. Technology after all *is* a project of world-making. So, what kind of a world are *we* collectively going to build together?

Positionality statement

I, the author of this paper, spent most of my career at a large technology corporation in the global north. However, the perspectives presented in this work is intended to challenge Big Tech and global north's view of technology and our collective futures. Outside of information retrieval research, I participate in social movement spaces and Southasian anti-fascist organizing. I stand in solidarity with the peoples of Palestine, Sudan, Congo, Rohingya, and other groups who are being subjected to unimaginable oppression and ongoing genocide. I also stand in solidarity with the people of Kashmir in their emancipatory struggles against state violence. These ongoing atrocities and cases of systemic oppression significantly informed and influenced my work.

Acknowledgments and Disclosure of Funding

The author gratefully acknowledges feedback from Ida Larsen-Ledet on the various drafts of this paper. No external funding was received in support of this work.

References

Veronica Abebe, Gagik Amaryan, Marina Beshai, Ilene, Ali Ekin Gurgen, Wendy Ho, Naaji R Hylton, Daniel Kim, Christy Lee, Carina Lewandowski, et al. Anti-racist HCI: notes on an emerging critical technical practice. In *CHI Conference on Human Factors in Computing Systems Extended Abstracts*, pages 1–12, 2022.

Rachel Adams. Can artificial intelligence be decolonized? *Interdisciplinary Science Reviews*, 46(1-2):176–197, 2021.

Yoko Akama, Seth Keen, and Peter West. Speculative design and heterogeneity in indigenous nation building. In *Proceedings of the 2016 ACM Conference on Designing Interactive Systems*, pages 895–899, 2016.

Azadeh Akbari. Big tech authoritarianism: Political synergies of an emerging power. *Dialogues on Digital Society*, page 29768640251370968, 2025.

Moritz Altenried. The platform as factory: Crowdwork and the hidden labour behind artificial intelligence. *Capital & Class*, 44(2):145–158, 2020.

Avishek Anand, Lijun Lyu, Maximilian Idahl, Yumeng Wang, Jonas Wallat, and Zijian Zhang. Explainable information retrieval: A survey. *arXiv preprint arXiv:2211.02405*, 2022.

Gil Appel, Juliana Neelbauer, and David A Schweidel. Generative AI has an intellectual property problem. *Harvard Business Review*, 7, 2023.

Leif Azzopardi, Charles LA Clarke, Paul Kantor, Bhaskar Mitra, Johanne R Trippas, Zhaochun Ren, Mohammad Aliannejadi, Negar Arabzadeh, Raman Chandrasekar, Maarten de Rijke, et al. Report on the search futures workshop at ECIR 2024. In *ACM SIGIR Forum*, volume 58, pages 1–41. ACM New York, NY, USA, 2024.

Chelsea Barabas. Beyond bias: Re-imagining the terms of "ethical AI" in criminal law. *Geo. JL & Mod. Critical Race Persp.*, 12:83, 2020.

Jeffrey Bardzell and Shaowen Bardzell. What is humanistic HCI? In *Humanistic HCI*, pages 13–32. Springer, 2015.

Jeffrey Bardzell and Shaowen Bardzell. Humanistic HCI. *Interactions*, 23(2):20–29, 2016a.

Shaowen Bardzell. Feminist HCI: taking stock and outlining an agenda for design. In *Proc. SIGCHI*, pages 1301–1310, 2010.

Shaowen Bardzell. Utopias of participation: Feminism, design, and the futures. *ACM Transactions on Computer-Human Interaction (TOCHI)*, 25(1):1–24, 2018.

Shaowen Bardzell and Jeffrey Bardzell. Towards a feminist HCI methodology: social science, feminism, and HCI. In *Proc. SIGCHI*, pages 675–684, 2011.

Shaowen Bardzell and Jeffrey Bardzell. Feminist design in computing. *The Wiley Blackwell Encyclopedia of Gender and Sexuality Studies*, pages 1–7, 2016b.

Shaowen Bardzell, Elizabeth Churchill, Jeffrey Bardzell, Jodi Forlizzi, Rebecca Grinter, and Deborah Tatar. Feminism and interaction design. In *CHI'11 Extended Abstracts on Human Factors in Computing Systems*, pages 1–4. 2011.

Elizabeth Beck, Kristie Seelman, Moon Charania, Susan M Snyder, and Sophie Saffan. Reproductive justice, bodily autonomy, and state violence. *Affilia* 39(3), pages 554–568, 2024.

Lotfi Belkhir and Ahmed Elmeligi. Assessing ict global emissions footprint: Trends to 2040 & recommendations. *Journal of cleaner production*, 177:448–463, 2018.

NJ Belkin and SE Robertson. Some ethical and political implications of theoretical research in information science. In *Proceedings of the ASIS Annual Meeting*, 1976.

Ruha Benjamin. Race after technology: Abolitionist tools for the new jim code. *Social Forces*, 2019.

Ruha Benjamin. *Imagination: A Manifesto (A Norton Short).* WW Norton & Company, 2024.

Asia J. Biega, Krishna P. Gummadi, and Gerhard Weikum. Equity of attention: Amortizing individual fairness in rankings. In *The 41st International ACM SIGIR Conference on Research &#38; Development in Information Retrieval*, SIGIR '18, pages 405–414, New York, NY, USA, 2018. ACM.

Abeba Birhane. Algorithmic colonization of Africa. *SCRIPTed*, 17:389, 2020.

Robert Booth. More than 140 Kenya Facebook moderators diagnosed with severe PTSD. 2024. URL https://www.theguardian.com/media/2024/dec/18/kenya-facebook-moderators-sue-after-diagnoses-of-severe-ptsd.

Paul Brest. The power of theories of change, 2010.

Jacob Browne and Laurel Green. The future of work is no work: A call to action for designers in the abolition of work. In *CHI Conference on Human Factors in Computing Systems Extended Abstracts*, pages 1–8, 2022.

Kelly Burke. 'Biggest act of copyright theft in history': thousands of australian books allegedly used to train AI model. *The Guardian*, 2023.

Kelly Burke. Generative AI is a marvel. Is it also built on theft? *The Economist*, 2024.

Matthew Burtell and Thomas Woodside. Artificial influence: An analysis of AI-driven persuasion. *arXiv preprint arXiv:2303.08721*, 2023.

Linda L Carli. Women, gender equality and COVID-19. *Gender in management: an International Journal*, 35(7/8):647–655, 2020.

Micah Carroll, Alan Chan, Henry Ashton, and David Krueger. Characterizing manipulation from AI systems. In *Proceedings of the 3rd ACM Conference on Equity and Access in Algorithms, Mechanisms, and Optimization*, pages 1–13, 2023.

Lucas Chancel, Thomas Piketty, Emmanuel Saez, and Gabriel Zucman. *World inequality report 2022.* Harvard University Press, 2022.

Kyle Chayka. Is A.I. art stealing from artists? *The New Yorker*, 2023.

Simon Chesterman. Good models borrow, great models steal: intellectual property rights and generative AI. *Policy and Society*, page puae006, 2024.

Aleksandr Chuklin, Ilya Markov, and Maarten de Rijke. Click models for web search. *Synthesis lectures on information concepts, retrieval, and services*, 7(3):1–115, 2015.

William D. Cohan. AI is learning from stolen intellectual property. It needs to stop. *The Washington Post*, 2023.

Devin Coldewey. Thousands of authors sign letter urging AI makers to stop stealing books. *TechCrunch*, 2023.

Nick Couldry and Ulises A Mejias. Data colonialism: Rethinking big data's relation to the contemporary subject. *Television & New Media*, 20(4):336–349, 2019.

E Cram, Martin P Law, and Phaedra C Pezzullo. Cripping environmental communication: A review of eco-ableism, eco-normativity, and climate justice futurities. *Environmental Communication*, 16(7):851–863, 2022.

Kate Crawford. The atlas of AI: Power, politics, and the planetary costs of artificial intelligence, 2021.

J Shane Culpepper, Fernando Diaz, and Mark D Smucker. Research frontiers in information retrieval: Report from the third strategic workshop on information retrieval in lorne (SWIRL 2018). In *ACM SIGIR Forum*, volume 52, pages 34–90. ACM New York, NY, USA, 2018.

Amit Datta, Michael Carl Tschantz, and Anupam Datta. Automated experiments on ad privacy settings: A tale of opacity, choice, and discrimination. *arXiv preprint arXiv:1408.6491*, 2014.

Fernando Diaz, Ryen White, Georg Buscher, and Dan Liebling. Robust models of mouse movement on dynamic web search results pages. In *Proc. CIKM*, pages 1451–1460, 2013.

Fernando Diaz, Bhaskar Mitra, Michael D Ekstrand, Asia J Biega, and Ben Carterette. Evaluating stochastic rankings with expected exposure. In *Proc. CIKM*, pages 275–284, 2020.

Catherine D'ignazio and Lauren F Klein. *Data feminism*. MIT press, 2020.

Paul Dourish and Scott D Mainwaring. Ubicomp's colonial impulse. In *Proceedings of the 2012 ACM conference on ubiquitous computing*, pages 133–142, 2012.

Gil Duran. The tech baron seeking to "ethnically cleanse" san francisco. *The New Republic*, 2024.

Charles C Earl. Towards an abolitionist AI: the role of historically black colleges and universities. *arXiv preprint arXiv:2101.02011*, 2021.

Ullrich Ecker, Jon Roozenbeek, Sander van der Linden, Li Qian Tay, John Cook, Naomi Oreskes, and Stephan Lewandowsky. Misinformation poses a bigger threat to democracy than you might think. *Nature*, 630(8015):29–32, 2024.

Michael D Ekstrand, Anubrata Das, Robin Burke, and Fernando Diaz. Fairness and discrimination in information access systems. *arXiv preprint arXiv:2105.05779*, 2021.

Seliem El-Sayed, Canfer Akbulut, Amanda McCroskery, Geoff Keeling, Zachary Kenton, Zaria Jalan, Nahema Marchal, Arianna Manzini, Toby Shevlane, Shannon Vallor, et al. A mechanism-based approach to mitigating harms from persuasive generative AI. *arXiv preprint arXiv:2404.15058*, 2024.

Tom Feltwell, Shaun Lawson, Enrique Encinas, Conor Linehan, Ben Kirman, Deborah Maxwell, Tom Jenkins, and Stacey Kuznetsov. "Grand visions" for post-capitalist human-computer interaction. In *Extended Abstracts of the 2018 CHI Conference on Human Factors in Computing Systems*, pages 1–8, 2018.

World Economic Forum. World economic forum global risks report 2024, 2024. URL `https://www.weforum.org/publications/global-risks-report-2024/`.

Human Rights Campaign Foundation. The epidemic of violence against the transgender and gender non-conforming community in the United States. 2023.

Free our feeds: Save social media from billionaire capture, 2025. URL `https://freeourfeeds.com/`.

Paulo Freire. Pedagogy of the oppressed. In *Toward a sociology of education*, pages 374–386. Routledge, 2020.

Miranda Fricker. *Epistemic injustice: Power and the ethics of knowing.* Oxford University Press, 2007.

Tony Fry. Design futuring. *University of New South Wales Press, Sydney*, pages 71–77, 2009.

Timnit Gebru and Émile P Torres. Eugenics and the promise of utopia through artificial general intelligence. *First Monday*, 2023.

Global Gender and Climate Alliance. Gender and climate change: A closer look at existing evidence. 2016.

Jon Gertner. Wikipedia's moment of truth. *The New York Times Magazine*, 2023.

Tarleton Gillespie. Algorithmically recognizable: Santorum's google problem, and Google's santorum problem. In *The Social Power of Algorithms*, pages 63–80. Routledge, 2019.

James Grimmelmann. The Google dilemma. *NYL Sch. L. Rev.*, 53:939, 2008.

Jonathan Grudin. The computer reaches out: The historical continuity of interface design. In *Proc. SIGCHI*, pages 261–268, 1990.

Kevin Guyan. Queer data. 2022.

Max Haiven. *Crises of imagination, crises of power: Capitalism, creativity and the commons.* Bloomsbury Publishing, 2014.

Blake Hallinan and Ted Striphas. Recommended for you: The Netflix prize and the production of algorithmic culture. *New media & society*, 18(1):117–137, 2016.

Evan Halper. Amid explosive demand, America is running out of power. *The Washington Post*, 2024.

Karen Hao. Artificial intelligence is creating a new colonial world order. *MIT Technology Review*, 2022.

Karen Hao and Andrea P Hernández. How the AI industry profits from catastrophe. *MIT Technology Review*, 2022.

Karen Hao and Deepa Seetharaman. Cleaning up ChatGPT takes heavy toll on human workers. *The Wall Street Journal*, 24, 2023.

Donna Harman et al. Information retrieval: the early years. *Foundations and Trends® in Information Retrieval*, 13(5):425–577, 2019.

Justin Hendrix and J Flowers. The whiteness of Mastodon. *Tech Policy Press. November*, 23:2022, 2022.

Mar Hicks. History in the making: Whistleblowers and big tech. *First Monday*, 2025.

Ezra Horbury and Christine "Xine" Yao. Empire and eugenics: Trans studies in the United Kingdom, 2020.

Lilly Irani, Janet Vertesi, Paul Dourish, Kavita Philip, and Rebecca E Grinter. Postcolonial computing: a lens on design and development. In *Proc. SIGCHI*, pages 1311–1320, 2010.

Lilly C Irani and M Six Silberman. Stories we tell about labor: Turkopticon and the trouble with" design". In *Proceedings of the 2016 CHI conference on human factors in computing systems*, pages 4573–4586, 2016.

Sheila Jasanoff and Sang-Hyun Kim. Containing the atom: Sociotechnical imaginaries and nuclear power in the United States and South Korea. *Minerva*, 47:119–146, 2009.

Sheila Jasanoff and Sang-Hyun Kim. *Dreamscapes of modernity: Sociotechnical imaginaries and the fabrication of power.* University of Chicago Press, 2015.

Lauren Johnson and Ryan Joe. Generative AI has a digital blackface problem. *Business Insider*, 2023.

Shivaram Kalyanakrishnan, Rahul Alex Panicker, Sarayu Natarajan, and Shreya Rao. Opportunities and challenges for artificial intelligence in India. In *Proceedings of the 2018 AAAI/ACM conference on AI, Ethics, and Society*, pages 164–170, 2018.

Gerald C Kane, Amber G Young, Ann Majchrzak, and Sam Ransbotham. Avoiding an oppressive future of machine learning: A design theory for emancipatory assistants. *Mis Quarterly*, 45(1):371–396, 2021.

Jay Caspian Kang. The big tech takeover of American politics, 2025. URL https://www.newyorker.com/news/fault-lines/the-big-tech-takeover-of-american-politics.

Os Keyes, Josephine Hoy, and Margaret Drouhard. Human-computer insurrection: Notes on an anarchist HCI. In *Proceedings of the 2019 CHI conference on human factors in computing systems*, pages 1–13, 2019.

Wooseok Kim, Eugenia Nazrullaeva, Anja Neundorf, Ksenia Northmore-Ball, and Katerina Tertytchnaya. Strategies of political control and regime survival in autocracies. *Available at SSRN*, 2024.

Naomi Klein. AI machines aren't 'hallucinating'. but their makers are. *The Guardian*, 2023.

Michael Klipphahn-Karge, Ann-Kathrin Koster, and Sara Morais dos Santos Bruss. Introduction: Queer AI. In *Queer Reflections on AI*, pages 1–19. Routledge, 2024.

Adrienne LaFrance. The rise of techno-authoritarianism. *The Atlantic*, 2024.

Stephan Lewandowsky. Truth and democracy in an era of misinformation, 2024.

Stephan Lewandowsky, Ullrich KH Ecker, John Cook, Sander Van Der Linden, Jon Roozenbeek, and Naomi Oreskes. Misinformation and the epistemic integrity of democracy. *Current opinion in psychology*, page 101711, 2023.

Ann Light. HCI as heterodoxy: Technologies of identity and the queering of interaction with computers. *Interacting with computers*, 23(5):430–438, 2011.

Conor Linehan and Ben Kirman. Never mind the bollocks, I wanna be anarchi: a manifesto for punk HCI. In *CHI'14 Extended Abstracts on Human Factors in Computing Systems*, pages 741–748. 2014.

Emily M Lund and Kara B Ayers. Ever-changing but always constant: "waves" of disability discrimination during the COVID-19 pandemic in the United States. *Disability and Health Journal*, 15(4):101374, 2022.

Anu Madgavkar, Olivia White, Mekala Krishnan, Deepa Mahajan, and Xavier Azcue. COVID-19 and gender equality: Countering the regressive effects. *JGI*, 2020.

Bernard Marr. Is generative AI stealing from artists? *Forbes*, 2023.

Mastodon. The people should own the town square, 2025. URL https://blog.joinmastodon.org/2025/01/the-people-should-own-the-town-square/.

Edouard Mathieu, Hannah Ritchie, Lucas Rodés-Guirao, Cameron Appel, Daniel Gavrilov, Charlie Giattino, Joe Hasell, Bobbie Macdonald, Saloni Dattani, Diana Beltekian, et al. Coronavirus (COVID-19) cases. *Our World in Data*, 2024.

Dan McQuillan. Anti-fascist AI. In *Resisting AI*, pages 135–148. Bristol University Press, 2022.

Rishabh Mehrotra, Ashton Anderson, Fernando Diaz, Amit Sharma, Hanna Wallach, and Emine Yilmaz. Auditing search engines for differential satisfaction across demographics. In *Proc. WWW*, pages 626–633, 2017.

Rishabh Mehrotra, James McInerney, Hugues Bouchard, Mounia Lalmas, and Fernando Diaz. Towards a fair marketplace: Counterfactual evaluation of the trade-off between relevance, fairness & satisfaction in recommendation systems. In *Proceedings of the 27th acm international conference on information and knowledge management*, pages 2243–2251, 2018.

Bhaskar Mitra. Search and society: Reimagining information access for radical futures. *Information Retrieval Research*, 1(1):47–92, 2025.

Bhaskar Mitra and Nick Craswell. An introduction to neural information retrieval. *Foundations and Trends® in Information Retrieval*, 2018.

Bhaskar Mitra and Maria Heuss. What is IR-for-Good? `https://bhaskar-mitra.github.io/posts/2025/09/01/what-is-ir-for-good/`, 2025.

Bhaskar Mitra, Henriette Cramer, and Olya Gurevich. Sociotechnical implications of generative artificial intelligence for information access. In *Information Access in the Era of Generative AI*, pages 161–200. Springer, 2024.

Shakir Mohamed, Marie-Therese Png, and William Isaac. Decolonial AI: Decolonial theory as sociotechnical foresight in artificial intelligence. *Philosophy & Technology*, 33:659–684, 2020.

Sandip Mondal and Ranjan Karmakar. Caste in the time of the COVID-19 pandemic. *Contemporary Voice of Dalit*, 16(1):114–121, 2024.

Thema Monroe-White. Emancipatory data science: a liberatory framework for mitigating data harms and fostering social transformation. In *Proceedings of the 2021 on Computers and people research conference*, pages 23–30, 2021.

James Muldoon and Boxi A Wu. Artificial intelligence in the colonial matrix of power. *Philosophy & Technology*, 36(4):80, 2023.

Nicola Neophytou, Bhaskar Mitra, and Catherine Stinson. Revisiting popularity and demographic biases in recommender evaluation and effectiveness. In *Proc. ECIR*, 2022.

Kwame Nkrumah. *Neo-colonialism.* Nelson, 1965.

Alexandra Olteanu, Jean Garcia-Gathright, Maarten de Rijke, Michael D Ekstrand, Adam Roegiest, Aldo Lipani, Alex Beutel, Alexandra Olteanu, Ana Lucic, Ana-Andreea Stoica, et al. FACTS-IR: fairness, accountability, confidentiality, transparency, and safety in information retrieval. In *ACM SIGIR Forum*, volume 53, pages 20–43. ACM New York, NY, USA, 2019.

Marcel O'Gorman. At the heart of artificial intelligence is racism and colonialism that we must excise. *The Globe and Mail web edition*, pages NA–NA, 2023.

Andreas Papamichail. Reinscribing global hierarchies: COVID-19, racial capitalism and the liberal international order. *International Affairs*, 99(4):1673–1691, 2023.

Peter S Park, Simon Goldstein, Aidan O'Gara, Michael Chen, and Dan Hendrycks. AI deception: A survey of examples, risks, and potential solutions. *arXiv preprint arXiv:2308.14752*, 2023.

Aneesh Patnaik, Jiahn Son, Alice Feng, and Crystal Ade. Racial disparities and climate change. *Princeton Student Climate Initiative*, 2020.

Gourab K. Patro, Arpita Biswas, Niloy Ganguly, Krishna P. Gummadi, and Abhijnan Chakraborty. Fairrec: Two-sided fairness for personalized recommendations in two-sided platforms. In *WWW*, pages 1194–1204. ACM / IW3C2, 2020.

Billy Perrigo. Exclusive: OpenAI used Kenyan workers on less than $2 per hour to make ChatGPT less toxic. *Last accessed*, 19, 2023. URL `https://time.com/6247678/openai-chatgpt-kenya-workers/`.

Christine Pinney, Amifa Raj, Alex Hanna, and Michael D Ekstrand. Much ado about gender: Current practices and future recommendations for appropriate gender-aware information access. In *Proceedings of the 2023 Conference on Human Information Interaction and Retrieval*, pages 269–279, 2023.

Eli G Rau and Susan Stokes. Income inequality and the erosion of democracy in the twenty-first century. *Proceedings of the National Academy of Sciences*, 122(1):e2422543121, 2025.

Parliament and Council EU. Regulation (EU) 2022/2065 of the European Parliament and of the Council of 19 October 2022 on a Single Market For Digital Services and amending Directive 2000/31/EC (Digital Services Act) (Text with EEA relevance). 2022.

Protection Regulation. Regulation (EU) 2016/679 of the European parliament and of the council. *Regulation (EU)*, 679:2016, 2016.

Emilie Rubayita. Alberta first nation voices 'grave concern' over Kevin O'Leary's proposed $70b AI data centre. 2025. URL `https://www.cbc.ca/news/canada/edmonton/alberta-first-nation-voices-grave-concern-over-kevin-o-leary-s-proposed-70b-ai-data-centre-1.7431550`.

Nithya Sambasivan, Erin Arnesen, Ben Hutchinson, Tulsee Doshi, and Vinodkumar Prabhakaran. Re-imagining algorithmic fairness in India and beyond. In *Proc. FAccT*, pages 315–328, 2021.

Deepak Saxena, PJ Wall, and Dave Lewis. Artificial intelligence (AI) ethics: A critical realist emancipatory approach. In *2023 IEEE International Symposium on Technology and Society (ISTAS)*, pages 1–5. IEEE, 2023.

Harrisen Scells, Shengyao Zhuang, and Guido Zuccon. Reduce, reuse, recycle: Green information retrieval research. In *Proc. SIGIR*, pages 2825–2837, 2022.

Rashi Shrivastava. OpenAI and Microsoft sued by nonfiction writers for alleged 'rampant theft' of authors' works. *Forbes*, 2023.

Ashudeep Singh and Thorsten Joachims. Fairness of exposure in rankings. In *KDD*, pages 2219–2228. ACM, 2018.

Ashudeep Singh and Thorsten Joachims. Policy learning for fairness in ranking. In H. Wallach, H. Larochelle, A. Beygelzimer, F. d'Alché-Buc, E. Fox, and R. Garnett, editors, *Proc. NeurIPS*, pages 5427–5437. Curran Associates, Inc., 2019.

Thomas Smyth and Jill Dimond. Anti-oppressive design. *Interactions*, 21(6):68–71, 2014.

G Stefani, J Devine, A Kelly, T Heimer, J Nguyen, H O'Shea, C Reiser, G Rose, J Sherry, J Skene, et al. Indigenous leaders at the frontlines of environmental injustice and solutions. *Natural Resources Defense Council*, 2021.

Farhana Sultana. *Confronting Climate Coloniality: Decolonizing Pathways for Climate Justice.* Taylor & Francis, 2024.

Cella M Sum, Franchesca Spektor, Rahaf Alharbi, Leya Breanna Baltaxe-Admony, Erika Devine, Hazel Anneke Dixon, Jared Duval, Tessa Eagle, Frank Elavsky, Kim Fernandes, et al. Challenging ableism: A critical turn toward disability justice in HCI. *XRDS: Crossroads, The ACM Magazine for Students*, 30(4):50–55, 2024.

Huatong Sun. Critical design sensibility in postcolonial conditions. *AoIR Selected Papers of Internet Research*, 2013.

Simon Susen. Emancipation. 2015.

Jasmina Tacheva and Srividya Ramasubramanian. AI empire: Unraveling the interlocking systems of oppression in generative AI's global order. *Big Data & Society*, 10(2): 20539517231219241, 2023.

Rebecca Tan and Regine Cabato. Behind the AI boom, an army of overseas workers in'digital sweatshops'. *The Washington Post*, pages NA–NA, 2023.

Dana H Taplin and Heléne Clark. Theory of change basics: A primer on theory of change. *New York: Actknowledge*, page 72, 2012.

Petter Törnberg and Juliana Chueri. When do parties lie? misinformation and radical-right populism across 26 countries. *The International Journal of Press/Politics*, page 19401612241311886, 2025.

Johanne R Trippas and J Shane Culpepper. Report from the fourth strategic workshop on information retrieval in lorne (SWIRL 2025). In *ACM SIGIR Forum*, volume 59, page 68, 2025.

Palashi Vaghela, Steven J Jackson, and Phoebe Sengers. Interrupting merit, subverting legibility: Navigating caste in 'casteless' worlds of computing. In *Proc. SIGCHI*, pages 1–20, 2022a.

Palashi Vaghela, Ramaravind Kommiya Mothilal, Daniel Romero, and Joyojeet Pal. Caste capital on twitter: A formal network analysis of caste relations among indian politicians. *Proceedings of the ACM on Human-Computer Interaction*, 6(CSCW1):1–29, 2022b.

Yanis Varoufakis. *Technofeudalism: What killed capitalism.* Melville House, 2024.

Pranshu Verma. In the shadows of arizona's data center boom, thousands live without power. 2024. URL `https://www.washingtonpost.com/technology/2024/12/23/arizona-data-centers-navajo-power-aps-srp/`.

James Vincent. The lawsuit that could rewrite the rules of AI copyright. *The Verge*, 22, 2022.

James Vincent. AI art tools stable diffusion and midjourney targeted with copyright lawsuit. *The Verge*, 2023.

Nicholas Vincent. Don't give OpenAI all the credit for gpt-3: You might have helped create the latest "astonishing" advance in AI too, 2020. URL `https://www.psagroup.org/blogposts/62`.

Nick Vincent and Hanlin Li. Github copilot and the exploitation of "data labor": A wake-up call for the tech industry, 2021. URL `https://www.psagroup.org/blogposts/62`.

Nick Vincent and Hanlin Li. Chatgpt stole your work. so what are you going to do?, 2023.

Sanne Vrijenhoek, Lien Michiels, Johannes Kruse, Alain Starke, Jordi Viader Guerrero, and Nava Tintarev. Report on normalize: The first workshop on the normative design and evaluation of recommender systems. In *CEUR Workshop Proceedings*, volume 3639. CEUR-WS, 2023.

Carol H Weiss. Nothing as practical as good theory: Exploring theory-based evaluation for comprehensive community initiatives for children and families. *New approaches to evaluating community initiatives: Concepts, methods, and contexts*, 1:65–92, 1995.

Ryen W White and Chirag Shah. Information access in the era of generative AI, 2025.

Wikipedia contributors. Theory of change — Wikipedia, the free encyclopedia, 2013. URL `https://en.wikipedia.org/wiki/Theory_of_Change`.

Adrienne Williams, Milagros Miceli, and Timnit Gebru. The exploited labor behind artificial intelligence. *Noema Magazine*, 13, 2022. URL `https://www.noemamag.com/the-exploited-labor-behind-artificial-intelligence/`.

Jeremy Williams. Why climate change is inherently racist. *BBC, accessed January*, 26:2023, 2022.

Rua M Williams, Kathryn Ringland, Amelia Gibson, Mahender Mandala, Arne Maibaum, and Tiago Guerreiro. Articulations toward a crip HCI. *Interactions*, 28(3):28–37, 2021.

Tom Williams and Kerstin Sophie Haring. No justice, no robots: From the dispositions of policing to an abolitionist robotics. In *Proceedings of the 2023 AAAI/ACM Conference on AI, Ethics, and Society*, pages 566–575, 2023.

Patrick Wolfe. Settler colonialism and the elimination of the native. *Journal of genocide research*, 8(4):387–409, 2006.

UN Women. Explainer: How gender inequality and climate change are interconnected. 28, 2022. URL `https://wrd.unwomen.org/explore/insights/explainer-how-gender-inequality-and-climate-change-are-interconnected`.

Erik Olin Wright. *Envisioning real utopias.* Verso Books, 2020.

Haolun Wu, Bhaskar Mitra, Chen Ma, Fernando Diaz, and Xue Liu. Joint multisided exposure fairness for recommendation. In *Proc. SIGIR*, 2022.

Haolun Wu, Bhaskar Mitra, and Nick Craswell. Towards group-aware search success. In *Proc. ICTIR*, pages 123–131, 2024.

Chloe Xiang. OpenAI used Kenyan workers making $2 an hour to filter traumatic content from ChatGPT. *VICE*, 2023.

Amber Young, Yaping Zhu, and Viswanath Venkatesh. Emancipation research in information systems: Integrating agency, dialogue, inclusion, and rationality research. 2021.

Eva Zangerle, Alan Said, and Christine Bauer. Beyond algorithms: Reclaiming the interdisciplinary roots of recommender systems (beyond 2025). In *Proc. RecSys*, pages 1360–1361, 2025.

Meike Zehlike and Carlos Castillo. Reducing disparate exposure in ranking: A learning to rank approach. In *Proceedings of the web conference 2020*, pages 2849–2855, 2020.

Xinyi Zhou and Reza Zafarani. A survey of fake news: Fundamental theories, detection methods, and opportunities. *ACM Computing Surveys (CSUR)*, 53(5):1–40, 2020.

Shoshana Zuboff. *The age of surveillance capitalism.* Profile Books, 2019.

Guido Zuccon, Harrisen Scells, and Shengyao Zhuang. Beyond co2 emissions: The overlooked impact of water consumption of information retrieval models. In *Proc. ICTIR*, pages 283–289, 2023.

Information Retrieval Research 1 (2025) 339-343 Published 12/2025

Editorial

AI Resilience and Information Retrieval

Paul B. Kantor PAUL.KANTOR@RUTGERS.EDU
University of Wisconsin
Madison, Wisconsin, United States of America

> *"Tout est pour le mieux dans le meilleur des mondes"*
> Dr. Pangloss, in Voltaire's Candide

IR provides multiple versions and sources to confirm an assertion. AI provides (essentially statistical) compilation and weighting of all those inputs. And generates a coherent and syntactically appropriate statement. What could go wrong?

With the rise of artificial intelligence, we have all asked what role Information Retrieval (IR) is to play, if any, in this new technological revolution. I want to suggest that there are three dimensions to this, all bearing on the complex relation between the world of documents and the physical world in which we all live. I'll put my concerns under three distinct headings: consistency, confidence, and completeness.

As to consistency: all of us who've engaged in what one of my colleagues has called "the torturing of AI" can provide instances where a generative AI (genAI) has confidently provided contradictory answers to the same question. If the AI were a human assistant, we would only need to point out the contradiction for that human assistant to understand that when there is a contradiction, both answers cannot be correct. At most one of them is.

While instances of genAI will apologize quite politely, they do not seem to understand that as descriptions of the real world (as opposed to, say, the world of spiritual beliefs) it is not acceptable to make contradictory assertions. To this, the friends and relations of genAI respond that since computers will soon be able to discover, and then prove mathematical theorems, they will "of course" be able to recognize and correct contradictory assertions about the real world. I've been wrong about many things before, but I remain intensely skeptical about this hope. Aside from the profound difficulties of mapping the real world to mathematics, Gödel (1931, 1987) has shown that even in mathematics not every true statement must be provable.

As to confidence: for many years we in IR have worked to adorn the results of information retrieval with numbers indicating confidence. More than 50 years ago, Sherman Kent, working at the United States Central Intelligence Agency, noticed that reports and assessments were filled with words reflecting confidence, but no one had sought to normalize the meaning of those terms. In a famous note (Kent, 1964) he proposed that these terms be assigned to ranges of probability. While admirable, this presents the challenge of assigning

WWW: https://irrj.org DOI: 10.54195/irrj.25004

probabilities to decisions that may not yet have been made, and will be unique events. What does it mean to say that "the probability that Madame X will win election in country Y is 75%"? (Menzel, 2025). Even if we resort to the quantum mechanical concept of a many worlds picture (Everett, 1957), we seem to live in only one of them, and so the answer is "yes" or "no." There are Brier scores (Wikipedia, 2025a, "Brier score"), and other ways to compute how well we are doing with such predictions. But the truth is that such scores represent aggregated accuracy, and are only meaningful when they are compiled over a large number of scored projections.

However, what I mean today by confidence is something less esoteric, but more troubling. Across the world, swindlers, politicians, and other kinds of "confidence men" have long known that the key to deceiving us is to look and sound like someone who can be trusted. In the generation of texts, we call this "the look and feel" of being trustworthy. Sadly, by exploiting the relatively low entropy of text and natural language, genAI is precisely able to achieve the correct "look and feel" absolutely without regard to whether what it says is or is not true. In other words, we have finally built a machine that is specifically capable of fooling us about whether it knows what it is talking about.

This leads to the third point: *completeness.* The problem here is what I call, with apologies to a man who is in some ways, both the father and the godfather of the scientific revolution, the "*Aristotelian fallacy.*" Aristotle was at heart a scientist, who gathered as many facts as he could. However, in organizing and understanding those facts, and seeking deeper understanding, he placed what we now believe is too much faith in deduction, and not enough in induction or in abduction, in its original sense (Douven, 2025). Although he used induction (the weighing of evidence on one side or another) Aristotle believed that any such conclusions also needed a deductive justification, such as "the efficacy of this form of government flows from the virtue of the citizens of the state." (This is a notion that was accepted without question at the founding of America, and is under substantial threat as I write). The modern scientific revolution began by rejecting a belief that all explanations are to be found in that "Aristotelian" way (Wikipedia, 2025c, "Roger Bacon").

At any specific moment in the evolution of human knowledge, more than induction and deduction are needed. Abduction supports the *creation of hypotheses* spurred by some consistency or inconsistency in the data about the world. Or, indeed, for the desire for more parsimonious explanations. It is an equal, perhaps even more critical component of scientific and technological advance. For some years, historians of science such as (Galison, 1997), have opposed the so-called Kuhnian notion of science as progressing by "change of [theoretical] paradigm" and have emphasized the importance of technologies in the advancement of science. Indeed, technology and theory fare best when they are united to form science. In 2025 Nobel laureate Joel Mokyr (Mokyr, 2005) was honored for pointing out that technological advance is hugely exponentiated by corresponding theoretical advances, which give us, its human masters, some usable understanding of both its strength and its limitations.

In this context, we must think carefully about the relation(s) between "the real world" and "the literature about that world." There is an axiom in chemistry "Why spend a day in the library when you can learn the same thing by working in the laboratory for a month?" (Corey, 2007). Indeed, the IR researcher Donald Swanson initiated the possibility of finding an unexpected connection between a treatment and an illness, not in the clinic, but in the literature, a concept now called literature-based discovery (Wikipedia, 2025b, "Literature-based discovery"). While science, as the joke reminds us, wants to "look for a lost wallet under the streetlamp," unexpected connections revealed in the literature give us "new kinds of streetlamps." However, *it cannot be the case that whatever is to be discovered tomorrow is hidden in the accumulated knowledge of today.*

The literature of all mankind at any given moment, even if we hoovered up the contents of all nine billion brains, simply cannot contain predictions of all that will be found later. Nothing known or imagined before the invention of the microscope was anywhere near a true picture of what it would reveal. Before graphene was invented, using only scotch tape and pencil lead, there was no description in any text or brain of how it could be produced and of its astonishing physical properties.

So, taken together, these concerns about consistency, confidence, and completeness suggest that Retrieval-Assisted-genAI may slouch toward Bethlehem on three feet of clay. Can the techniques and principles of information retrieval help to remedy these weaknesses?

The danger involved, when machines will read what other machines have written, and the original "facts on the ground" will become ever more remote, was anticipated, in the magnificent dystopian novella called "The Machine Stops" (Forster, 1909). In some future world humans live in identical underground cells, and communicate their ideas without ever leaving those cells. At what we would now recognize as Zoom sessions, they hear from expert "lecturers" who, to become expert, had ventured up to the surface of the earth, to observe it. But the machine that maintains air, food, heat and light for all of the people had decided to eliminate the essential respirators, without which no one dared to venture to that real world. Let Forster, speaking from the grave, tell us how humanity responded:

> *. . . even the lecturers acquiesced when they found that a lecture on the sea was none the less stimulating when compiled out of other lectures that had already been delivered on the same subject. "Beware of first-hand ideas!" exclaimed one of the most advanced of them. "First-hand ideas do not really exist. They are but the physical impressions produced by love and fear, and on this gross foundation who could erect a philosophy? Let your ideas be second-hand, and if possible tenth-hand, for then they will be far removed from that disturbing element — direct observation. Do not learn anything about this subject of mine — the French Revolution. Learn instead what I think that Enicharmon thought Urizen thought Gutch thought Ho-Yung thought Chi-Bo-Sing thought Lafcadio Hearn thought Carlyle thought Mirabeau said about the French Revolution. Through the medium of these ten great minds, the blood that was shed at Paris and the windows that were broken at Versailles will be clarified to an idea which you may employ most profitably in your daily lives."*

Grim stuff. All three of the stated problems: consistency, confidence, and completeness seem, in one way or another, to circle back to these questions: are AI-generated statements true? Can that truth, in any way, be inferred from knowledge about how the statements were generated? Is there any way that our skill in retrieving and organizing what has been known and recorded can help us to know whether it and what we deduce from it is true?

There are some threads of hope related to the ideas of tracking provenance, authentication, and some kind of agreement about what is to be considered a faithful source of knowledge. Perhaps if the World Wide Web had more closely followed the road map of Ted Nelson's Xanadu (Nelson, 1999) we would be in better shape. Perhaps if search, and guideposts to related documents had become distributed rather than centralized (Kantor et al., 2000) we would be in better shape. But today the Internet lets any of us share our opinions. Any centralized robot can weigh and count them. The Internet itself can be locally shaped by nation-states wishing to advance one or another view of the real world. All in all, it is very hard to believe that we can define any scientific process that will tell us what to trust.

There is still room for optimism. My coeditor of the predecessor Information Retrieval Journal, Stephen Robertson, with his colleagues, devised a weight formula (Robertson et al., 1994) whose name, BM25, hints that it was better than perhaps some 24 other "Best Match" methods"[1]. In his Salton award lecture Robertson (2000) Steve argued quite well that IR "cannot have a theory." In this note I am arguing that establishing the "truth" of retrieval-augmented AI, as part of the IR enterprise, cannot become a science. Thus, it may fail to meet Mokyr's criterion for exponential growth – circular funding deals notwithstanding.

The advocates and builders of the new AI do recognize some risks. We surely lack a scientific understanding of its powers, and its limitations. But perhaps some clever *ad hoc* tool, some Truth Finder 25, will help us to hold our ground against the advancing flood of overconfident assertions. As we search for it, I hope that the innovators and inventors of such *Truth Finding* tools will share their ideas with us in the pages of this important new journal.

References

E. J. Corey. Frank H. Westheimer, major figure in 20th century chemistry, dies at 95. *Harvard Gazette*, April 19 2007. URL https://news.harvard.edu/gazette/story/2007/04/frank-h-westheimer-major-figure-in-20th-century-chemistry-dies-at-95/.

Igor Douven. *The Stanford Encyclopedia of Philosophy*, chapter: Abduction. Metaphysics Research Lab, Stanford University, 2025. URL https://plato.stanford.edu/archives/sum2025/entries/abduction/.

Hugh Everett. *'Relative State' Formulation of Quantum Mechanics.* PhD Thesis, Princeton University, 1957. URL https://journals.aps.org/rmp/abstract/10.1103/RevModPhys.29.454.

E. M. Forster. The machine stops. *The Oxford and Cambridge Review*, 1909. URL https://www.cs.ucdavis.edu/~koehl/Teaching/ECS188/PDF_files/Machine_stops.pdf.

[1] I believe there were both fewer and more than 24. The name may refer to two components of the model. As their paper explains, the Okapi team explored many values of several model parameters to find the published model. This remains the basic principle by which Artificial Intelligence models are tuned and trained today.

Peter Galison. Image and logic: A material culture of microphysics. *University of Chicago Press*, 1997.

Kurt Gödel. Über formal unentscheidbare sätze der principia mathematica und verwandter systeme I. *Monatshefte für Mathematik und Physik*, 38:173–198, 1931. doi: 10.1007/BF01700692.

Kurt Gödel. *Collected Works, Volume I: Publications 1929–1936*, chapter: On Formally Undecidable Propositions of Principia Mathematica and Related Systems I. Oxford University Press, 1987.

Paul B. Kantor, Endre Boros, Benjamin Melamed, Vladimir Meñkov, Bracha Shapira, and David J. Neu. Capturing human intelligence in the net. *Communications of the ACM*, 43(8):112–115, 2000.

Sherman Kent. Words of estimative probability. *Studies in Intelligence*, 8(4):49–65, 1964. URL `https://www.cia.gov/resources/csi/static/Words-of-Estimative-Probability.pdf`.

Christopher Menzel. *The Stanford Encyclopedia of Philosophy*, chapter: Possible Worlds. Metaphysics Research Lab, Stanford University, 2025. URL `https://plato.stanford.edu/archives/fall2025/entries/possible-worlds/`.

Joel Mokyr. The intellectual origins of modern economic growth. *Journal of Economic History*, 65(2):285–351, 2005.

Theodor Holm Nelson. The unfinished revolution and Xanadu. *ACM Computing Surveys*, 31(4es), 1999.

Stepen E. Robertson, Steve Walker, Susan Jones, and Michelle Hancock-Beaulieu. Okapi at trec-3. In *Proceedings of the Third Text REtrieval Conference (TREC)*. NIST, 1994. URL `https://trec.nist.gov/pubs/trec3/t3_proceedings.html`.

Stephen E. Robertson. Salton award lecture on theoretical argument in information retrieval. *ACM SIGIR Forum*, 34(1):1–10, 2000.

Wikipedia. Brier score, October 4, 2025a. URL `https://en.wikipedia.org/w/index.php?title=Brier_score&oldid=1315096299`.

Wikipedia. Literature-based discovery, September 30, 2025b. URL `https://en.wikipedia.org/w/index.php?title=Literature-based_discovery&oldid=1314260367`.

Wikipedia. Roger Bacon, September 15, 2025c. URL `https://en.wikipedia.org/w/index.php?title=Roger_Bacon&oldid=1311560035`.