Information Retrieval Research 1 (2025) 221-245 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 dis-

tribution, 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.