AIIR Lab at COLIEE 2025: Exploring Applications of Large Language Models for Legal Text Retrieval and Entailment

Deiby Wu
University of Southern Maine
Portland, Maine, USA
deiby.wu@maine.edu

Sarah Lawrence
University of Southern Maine
Portland, Maine, USA
sarah.lawrence@maine.edu

Behrooz Mansouri
University of Southern Maine
Portland, Maine, USA
behrooz.mansouri@maine.edu

Abstract

This paper presents the approaches and results of the Artificial Intelligence and Information Retrieval (AIIR) Lab's participation in the 2025 Competition on Legal Information Extraction and Entailment (COLIEE). The AIIR Lab engaged in all four main tasks, leveraging large language models (LLMs) such as Mistral-7B and LLaMA-3. For the Legal Case Retrieval task (Task 1), the team employed LLMs for case summarization, followed by ranking using a fine-tuned bi-encoder model. In the Legal Case Entailment task (Task 2), a fine-tuned cross-encoder model was utilized to assess entailment between case paragraphs. The Statute Law Retrieval task (Task 3) involved augmenting existing training data with LLMs and then fine-tuning a bi-encoder model for search. Finally, for the Legal Textual Entailment task (Task 4), LLMs were employed with three prompting techniques, zero-shot, few-shot, and chain-of-thought (COT), with majority voting applied to determine the final answer. This paper provides the details of the proposed methodologies and experimental results for each task.

CCS Concepts

• **Information systems** → *Specialized information retrieval.*

Keywords

Legal Case Retrieval, Legal Entailment, Legal Language Processing

ACM Reference Format:
Deiby Wu, Sarah Lawrence, and Behrooz Mansouri. 2025. AIIR Lab at COLIEE 2025: Exploring Applications of Large Language Models for Legal Text Retrieval and Entailment. In *Proceedings of COLIEE 2025 workshop, June 20, 2025, Chicago, USA.* ACM, New York, NY, USA, 7 pages.

1 Introduction

The application of artificial intelligence, particularly large language models (LLMs), has advanced the field of legal document processing. These models have shown remarkable capabilities in tasks such as legal information retrieval and entailment [15], summarization [5], and question answering [12], enhancing the efficiency and accuracy of legal analyses. To support the development and evaluation of AI models in the legal domain, several specialized datasets have been introduced, including FALQU [13], LexGLUE [3], and Pile of Law [8]. These resources provide the necessary data to train, fine-tune, and evaluate models tailored for legal applications.

The Competition on Legal Information Extraction/Entailment (COLIEE) [18] is dedicated to the automated analysis of legal texts. COLIEE has two retrieval tasks, including Legal Case Retrieval (Task 1) and Statute Law Retrieval (Task 3), and two entailment tasks: Legal Case Entailment (Task 2) and Legal Textual Entailment (Task 4).

Tasks 1 and 2 use the Federal Court of Canada case laws as the corpus. In Task 1, for a given case query, the goal is to find noticed cases in the collection. The query case references a noticed case; however, the references are removed from the query case. Task 2 aims to detect the paragraphs that entail the decision for a given relevant case. Tasks 3 and 4 are based on Japanese Civil Law articles (with English translation available). The goal of Task 3 is to return relevant articles for a query. Articles relevant to a query are those that can answer the query (with Yes/No) entailed from the article. Finally, Task 4 focuses on question answering; given a legal bar question, and a Civil Law article, the task explores if the article can entail the query.

The Artificial Intelligence and Information Retrieval (AIIR) Lab from the University of Southern Maine participated for the first time in the COLIEE, proposing different systems for all four tasks. Our runs rely on two large language models (LLMs): Mistral-7B-Instruct-V0.2 [9] (hereafter referred to as Mistral) and LLaMA-3-8B-Instruct [6] (hereafter referred to as LLaMA). These LLMs are used for case summarization, ranking, and classification for entailment tasks. We also used neural information retrieval models for retrieval tasks.

Our most effective systems for each task are as follows. For Task 1, we used Mistral for case summarization, followed by a fine-tuned Sentence-BERT bi-encoder[19] for ranking. In Task 2, our fine-tuned cross-encoder model provided the best effectiveness. For Task 3, we first augmented the existing training data with Mistral and then fine-tuned a bi-encoder model for search. Finally, for Task 4, we used Mistral with three prompting techniques: zero-shot, few-shot, and chain-of-thoughts (COT), and then applied majority voting to get the final answer.

In this paper, for each task, we first review its objectives, then present our proposed models, and finally discuss the experimental results.

2 Task 1: Legal Case Retrieval

The Legal Case Retrieval task aims to evaluate the effectiveness of legal document retrieval systems in identifying relevant case laws that support a given (unseen) query case. The system must retrieve "noticed cases", which are those referenced by the query case, though explicit references are redacted to assess retrieval

COLIEE 2025, Chicago, USA

Table 1: Average number of words in Task 1 collection cases at each cleaning step.

Data	Base	THUIR	YAKE!	Mistral	LLaMA
Train	4654.07	3943.90	3547.54	347.50	196.80
Test	5125.14	4466.06	4027.06	358.77	288.47

accuracy. The corpus consists of Federal Court of Canada case laws, with training data providing query cases and their corresponding noticed cases, while test data includes only query cases without labels. The evaluation metrics for Task 1 include precision, recall, and F-measure. This task uses micro-averaging, where evaluation metrics are calculated collectively over all queries.

2.1 Proposed Models

In our proposed models, both query and candidate cases were first passed through a similar preprocessing pipeline, shown in Figure 1. We used the THUIR team approach [11] at COLIEE 2023 for the initial case cleaning. This cleaning includes steps such as removing extra spaces, handling punctuation, and removing French text. Despite the cleaning, the legal cases are longer than the maximum sequence length that neural information retrieval and large language models can support.

To overcome this issue, we relied on YAKE! keyword extractor [2], and for each legal case, we kept the paragraphs that contained the top-3 important keywords, based on YAKE! scores. We then used two large language models, Mistral and LLaMA, to summarize each case. Both LLMs were used with a temperature of 0.1 and a maximum sequence length of 2048 for generation. For Mistral, we used the prompt "Provide a concise summary of this legal case." and passed the legal case to be summarized by LLM. For LLaMA, we used a similar prompt as Mistral but also considered a system message (role) as:

> You are a legal expert that summarizes long legal cases into a precise paragraph while keeping the main content.

As shown in Table 1, the initial legal cases (Base) contain a high number of words on average, with the test cases being slightly longer than the training cases. The THUIR cleaning step reduces the word count by removing unnecessary elements, leading to a moderate reduction in length. The YAKE! filtering step further reduces the size by retaining only the most relevant paragraphs based on keyword importance, ensuring that critical content is preserved while eliminating less relevant sections. The most noticeable reductions occur after summarization with Mistral and LLaMA. Mistral produces summaries averaging around 347 words for training cases and 359 words for test cases, indicating a strong compression while maintaining consistency between datasets. LLaMA generates even shorter summaries, reducing cases to an average of 197 words for training and 288 words for test cases.

After summarizing each case, we proceeded to fine-tune a bi-encoder retrieval model using the pre-trained 'all-mpnet-base-v2' [19] model. The combination of summarization and transformer-based models has been previously explored in legal case retrieval to address sequence length limitations [1, 21]. We utilized the provided training data, which comprises 1,678 queries with an average of 4.1 associated cases. The dataset was split into a 90:10 ratio for training and validation sets. For fine-tuning, we used the Multiple Negatives Ranking Loss (MNRL) [7], which is well-suited for scenarios where only positive samples are provided. The model was fine-tuned for 10 epochs using a batch size of 16, and we selected the best-performing model based on the Mean Average Precision at rank 100 (MAP@100) on the validation set.

Table 2: Results of AIIR Lab Runs for Task 1 on the COLIEE 2025 Test Set (400 Queries).

Model	F-Measure	Precision	Recall
AIIRmpMist5	**0.2171**	0.2040	**0.2319**
AIIRmpMist3	0.1872	0.2308	0.1575
AIIRcombMNZ	0.1879	**0.2317**	0.1580

Based on these methods, we devised three experimental runs as follows:

(1) **AIIRmpMist5**: In this run, we use the case summaries generated by Mistral to fine-tune a bi-encoder model. For each case query, the model retrieves the top-5 most relevant results.
(2) **AIIRmpMist3**: Using the same approach as above, this run retrieves only the top-3 results, prioritizing precision. This threshold was chosen based on the average number of noticed cases in the training set.
(3) **AIIRcombMNZ**: While the previous runs relied exclusively on Mistral's summaries, we also generate summaries using LLaMA and fine-tune a separate bi-encoder model with them. Since our experiments on the training data showed that LLaMA's summaries were less effective for legal cases, we combine the outputs from both the LLaMA- and Mistral-based bi-encoder models using the CombMNZ fusion [10], and then selected the top-3 results. CombMNZ multiplies the number of ranks where the document occurs by the sum of the scores obtained with the two systems.

2.2 Experimental Results

Table 2 shows our results on the COLIEE 2025 test set for Task 1. 400 case queries are considered for this year's competition, with an average of 4.3975 noticed cases per query. As can be seen from this table, including LLaMA-based results led to a slight improvement in the precision, and with Mistral summaries, the recall gain at cut@5 was higher, leading to the *AIIRmpMist5* run being the most effective. For our runs at cut@3 (*AIIRmpMist3* and *AIIRcombMNZ*), the precision was significantly higher than the run with cut@5 (*AIIRmpMist5*), using the paired Student's t-test ($p < 0.05$). In contrast, with cut@5, the recall was significantly higher than the other two runs. *AIIRmpMist5* F-measure was also significantly higher than the other runs.

Looking at the results with Mistral fine-tuned data, at cut@5, the precision for 159 out of 400 query cases drops compared to cut@3. For 2 query cases, one regarding Harrington v. Microsoft (with ID 063752) and the other Canadian Union of Postal Workers v. Canada Post (with ID 025612), the precision dropped from 1 to 0.6 when the

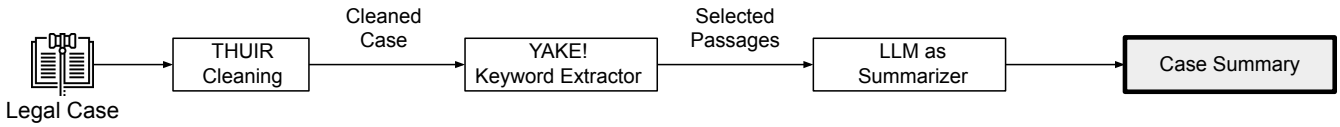

Figure 1: Proposed approach for summarizing legal cases in task 1. After processing each case with the THUIR approach, passages with top-3 keywords are passed to an LLM to summarize the case.

cut was increased from 3 to 5. For case 025612, the retrieval model included two additional cases in the top-5 results due to strong lexical matches with the query terms such as contempt, arbitration, and other procedural phrases, leading the model to identify them as potentially relevant. However, the legal issues in these cases, involving copyright injunctions and maritime contract disputes, differ from the labor and administrative arbitration focus of the query, making them irrelevant despite the initial semantic match. On the other hand, with a cut of 5, the recall increases for 117 queries. In particular, for 7 of these queries, for which there was only one relevant case, the recall increases from 0 to 1 when the depth of retrieved instances is set to 5.

Our models failed to retrieve any relevant cases for 37% of the test queries. One major issue is that the bi-encoder model discards the details of cases. For instance, for case 001083, which is related to the Refugee Protection Division, our fine-tuned bi-encoder model retrieved cases containing common lexical and semantic cues, such as references to the Refugee Appeal Division, judicial review, and subsection 110(4) of the IRPA that matched the query. However, the query concerns the applicant's refugee claim based on sexual orientation and credibility issues, and the model retrieved cases focusing on evidentiary disputes over a high school diploma. Similarly, in a query involving document production under s. 39 of the Canada Evidence Act (with ID 099982), the model retrieved cases that address constitutional challenges to ex parte representations and disputes over personal information disclosures. Although these cases share similarities in statutory references and document production terms, they differ in legal focus from the original query, which deals specifically with document certification procedures and confidentiality issues in a multi-agency litigation context.

3 Task 2: Legal Case Entailment

The Legal Case Entailment task aims to predict the decision of a new case by identifying supporting paragraphs from relevant past cases. Given a query case decision and a noticed case, the system must determine which paragraph in the noticed case entails the decision. Training data consists of triples: a query, a noticed case, and the paragraph number that supports the decision. In the test phase, only queries and noticed cases are provided, without paragraph labels. The process must be fully automated, with no human intervention or system modifications. Similar to Task 1, the evaluation metrics for Task 2 include precision, recall, and F-measure, with equal weighting for precision and recall. Precision measures the proportion of correctly retrieved paragraphs among all retrieved paragraphs, while recall calculates the proportion of correctly retrieved paragraphs among all relevant ones.

3.1 Proposed Models

For Task 2, we proposed three systems. For each query, the top-ranked paragraph according to the model's score is selected. The second-ranked paragraph is included in the result if its score exceeds a threshold. For our first two runs, this threshold is set to 0.5, and for the last run, it is set to less than 10% difference compared to the top-ranked paragraph score. Here are our runs for this task:

(1) **crossAIIRLab.** This approach uses a cross-encoder model, "ms-marco-MiniLM-L-6-v2", with the MiniLM [22] architecture trained for passage re-ranking [16] on the MS MARCO dataset. We fine-tuned this model on Task 2 training data to optimize its ability to rank legal paragraphs according to textual entailment. We used 675 queries from the original 825 queries and their associated paragraphs to fine-tune the model for 30 epochs with a batch size of 8 and a learning rate of 2e-5. The remaining 150 queries were used for testing the model's performance. The model was optimized using binary cross-entropy.
(2) **mT5AIIRLab.** Our second approach fine-tunes a MonoT5 [17] model, "monot5-base-msmarco", which adapts a pre-trained T5 encoder-decoder for passage re-ranking and is also trained on the MS MARCO dataset. For each training example, the input was given as "Query:[query text] Document: [paragraph text] Relevant: ", and the target was either *true* or *false* depending on the relevance judgment given by the model. Similar to the previous model, we fine-tuned the model using 675 queries from the 825 queries for 3 epochs with a batch size of 8 and an AdamW optimizer with loss computed using sequence-to-sequence cross-entropy. At inference, the model computes log-probabilities for both *true* and *false* outputs and uses the difference as the score.
(3) **mergeAIIRLab.** As an ensemble model, this model combines re-ranking scores from four models: BM25 [20], a pre-trained bi-encoder ("all-mpnet-base-v2") [19], and our two previous runs. For each query, scores from all models are min-max normalized and merged using a weighted average with weights set with grid-search on the train set. The weights for BM25 and bi-encoder are set to 0.1, and 0.4 for previous runs.

3.2 Experimental Results

Our results for Task 2 are presented in Table 3. The test set contains 100 queries with an average of 1.81 relevant paragraphs per query. Among all the models evaluated, the fine-tuned cross-encoder (crossAIIRLab) achieved the best performance across all the official metrics.

The *crossAIIRLab* model captured the context of legal reasoning even when different terminologies were used. For example, one

Table 3: Results of AIIR Lab Runs for Task 2 on the COLIEE 2025 Test Set (100 Queries).

Model	F-Measure	Precision	Recall
crossAIIRLab	**0.2368**	**0.2927**	**0.1989**
mergeAIIRLab	0.2229	0.2632	0.1934
mt5AIIRLab	0.1930	0.2050	0.1823

query discussing judicial impartiality and racial dynamics was correctly matched to a paragraph from *R. v. R.D.S.* that examined the reasonable person standard in the context of racial bias. Another query concerning translation errors and credibility was aligned with a paragraph addressing similar Charter protections, even sharing phrases like "differences in nuance between what is said in one language and its translation into another." These examples show the model's ability to not only focus on similarity but also to detect deeper semantic entailment in the legal field.

However, there were instances where the fine-tuned cross-encoder identified topic overlap between the query and a paragraph without achieving logical alignment. For example, a query about the likelihood of confusion in trademark law was matched with a paragraph discussing trademark ownership and first use; reflecting a legal relation but lacking an entailment element. In another case, a query comparing two different standards of judicial review was matched to a paragraph that opposed that very approach. These examples suggest that while our fine-tuned cross-encoder is capable of semantic matching, it can be misled by similar legal vocabulary when the contextual alignment is insufficient.

Similarly, the *mt5AIIRLab* model performs well when the paragraph supports the query's reasoning. For instance, a query asserting the importance of centralized residency in Canada was matched with a paragraph quoting the established legal test for residency, including the phrase "centralizes his ordinary mode of living." In another example, a query regarding the evidentiary weight of interview notes was paired with a paragraph explaining that such notes are insufficient without supporting affidavits. However, the MonoT5 model sometimes relied too heavily on surface-level similarity. In one case, a query on judicial impartiality and racial dynamics (the same query used for the cross-encoder analysis) was matched to a paragraph stating that there was no apprehension of bias, failing to address the query's underlying context and reasoning. In another instance, a query justifying the denial of deferral in light of a pending humanitarian application was paired with a paragraph that argued in favor of deferral under similar circumstances, illustrating the model's difficulty in grasping the logic of the argument.

Overall, both models show strong performance when semantic alignment is unambiguous but struggle when queries and paragraphs present opposing positions using similar terminology. As shown in Table 3, the merged model performed slightly below the cross-encoder, suggesting that weaker components such as BM25 and the bi-encoder may have negatively impacted the overall score by emphasizing lexical or superficial semantic similarity over true entailment. This finding underscores the importance of carefully controlling the contribution of each model in ensembles for entailment tasks. Future improvements include refining ensemble weighting strategies, removing models that detract from overall performance, and training on cases that emphasize differences in legal reasoning.

4 Task 3: Statute Law Retrieval

This task aims to evaluate the effectiveness and reliability of legal document retrieval systems by assessing their performance in retrieving relevant Civil Law articles based on previously unseen queries. The system operates on a static set of Japanese Civil Law articles, provided in both Japanese and English translation, and must automatically identify all relevant articles that contribute to answering a query. An article is considered relevant if its meaning entails a yes/no response to the query.

The evaluation metrics for this task include precision, recall, and F2-measure, with an emphasis on recall since the retrieval process serves as a pre-selection step for entailment. Precision measures the proportion of correctly retrieved articles among all retrieved articles, while recall calculates the proportion of correctly retrieved articles among all relevant articles. The F2-measure prioritizes recall by weighting it more heavily than precision. Additionally, Mean Average Precision (MAP) is used to analyze system performance. The final evaluation score is computed using macro-averaging, where the metric is calculated for each query and then averaged across all queries. For this task, there were 1,206 samples provided for training purposes, and the search collection contained 776 articles with an average of 71.96 words in each case. Compared to Task 1, the cases in this task are cleaner and do not need pre-processing steps.

4.1 Proposed Models

For Task 3, we developed three retrieval models, and in each run, we considered the top-3 results. To increase the number of training samples, we use Mistral for data augmentation with 10 different prompts to rewrite the original sample queries in the training set. Each prompt is designed to provide a different query and increase both the number and diversity of training data. If (Q_i, A_j) are in the training data as positive sample, where Q_i is the original query, and A_j is a relevant article, we generate $\{Q_{i0}, Q_{i1}, ..., Q_{i9}\}$ and consider positive samples as $\{(Q_i, A_j), (Q_{i0}, A_j), (Q_{i1}, A_j), ..., (Q_{i9}, A_j)\}$ for fine-tuning. Our prompts include the following sample that aims for summarization, simplification, and rephrasing:

- Reword the legal query into a **concise summary** that highlights the essential legal elements, using precise legal language.
- Transform the query into a **simplified** yet formally worded version that maintains all critical legal details.
- **Paraphrase** the following legal query using alternative legal terms and sentence structures, while preserving the original legal meaning and nuance.

Table 4 shows how a training query (ID H23-1-4) is rewritten using three augmentation prompts. In the Paraphrased variant, phrases are reworded; for instance, "cannot refused execution of his or her obligation" becomes "cannot withhold performance," while the timing and rescission conditions are retained, showing structural and lexical variation without loss of meaning. The Summarized form highlights the core legal elements: "contract formation through another party's fraud," "the defrauded party cannot refuse

Table 4: Rewriting of training query (H23-1-4) via three augmentation prompts: paraphrase, concise summary, and formal simplification for Task 3 data augmentation.

Query	In cases where the person who makes the manifestation forms a contract based on the fraud of the other person, even if the obligation of the other person is not yet executed based on the contract, the person who makes the manifestation cannot refused execution of his or her obligation for reason of fraud as long as the manifestation of intention is not rescinded.
Paraphrased	When a contract is formed through another party's deceit, prior to their contractual obligation's execution, the deceived party cannot withhold performance based on the fraud. This rule holds as long as the intention to be bound has not been revoked.
Summarized	In instances of contract formation through another party's fraud, the defrauded party cannot refuse performance of their obligations prior to contract execution, provided their manifestation of intent has not been rescinded.
Simplified	In instances where a contract is formed through the deceit of one party, the opposing party's obligation to perform under the contract cannot be refused execution based on the fraudulent inducement, unless and until the manifestation of intent to be bound has been effectively rescinded.

performance prior to contract execution," and the binding "manifestation of intent." Finally, the Simplified version preserves critical details but refines syntax and phrasing, such as rephrasing the obligation's status ("obligation... cannot be refused execution"), to improve readability and flow while retaining formality. We leave further exploration of rewriting and augmentation of legal queries for future work.

After data augmentation, we use the original training samples, along with the augmented data to fine-tune a bi-encoder model, "all-mpnet-v2" for 10 epochs. We split the data in a 90:10 ratio for the training and validation sets. The best model on the validation set is selected based on the highest Spearman correlation score by assessing the similarity of the generated embeddings by comparing them, using cosine similarity, Euclidean, and Manhattan distances, to the gold standard labels. This forms our first run, **mpnetAIIRLab**.

In our second run, **mistAIIRLab**, we re-ranked the top-10 results from the previous run (*mpnetAIIRLab*) for each query using a pairwise approach with Mistral. For this, we used the following prompt and passed the query with two candidates for re-ranking:

> Being a ranking model, your task is to decide for a given legal query in the context of Japanese civil law, which of the two articles is more relevant.

Finally, our third run, **NVAIIRLab**, uses another bi-encoder architecture, using NVIDIA model's 'NV-Embed-v2' [4]. For this approach, we used the pre-trained model as a zero-shot baseline without any further fine-tuning.

4.2 Experimental Results

Table 5 presents our results for Task 3 on the COLIEE 2025 test queries. Among our submissions, the *mpnetAIIRLab* run achieved the best performance across all evaluation metrics. This higher performance was statistically significant (Student's t-test, $p < 0.05$) over *mistAIIRLab* for the MAP metric and over *NVAIIRLab* for Precision.

Looking at the results, the *mpnetAIIRLab* model has a recall of 1 for 82.4% of queries, while it fails to retrieve any relevant documents for 8.1% of queries. However, for 78.3% of the queries, only one of the retrieved articles out of three is considered as relevant, leading to a precision of 0.33 for those instances. On average, each query in the test set has 1.26 relevant articles, with only one query having three relevant articles and the remainder having fewer. Therefore, our approach of always considering top-3 results should be further improved by learning the correct cutoff for the top-k retrieved results.

Table 5: Results of AIIR Lab Runs for Task 3 on the COLIEE 2025 Test Set (74 Queries).

Model	F2	Precision	Recall	MAP
mpnetAIIRLab	**0.6246**	**0.3333**	**0.8291**	**0.7931**
mistAIIRLab	0.5672	0.3034	0.7521	0.6867
NVAIIRLab	0.5554	0.2863	0.7479	0.7412

Exploring *mpnetAIIRLab* retrieval results, consider a topic such as R06-05-O, which is related to a debtor who causes collateral to be lost before the assigned time. Our model, which uses topic similarity, finds two relevant documents; however, it also retrieves a non-relevant article (with ID 706) that addresses an entirely different scenario of early performance of an obligation, making it irrelevant to the query regarding loss of collateral. When re-ranked with Mistral, this irrelevant article was dropped from the result. However, Mistral includes another article (with ID 135) that focuses on the timing for the performance or expiration of a juridical act, rather than addressing the consequences of the debtor's actions. This article describes the general principle of a "time of commencement" for a legal act or obligation.

While *mpnetAIIRLab* was on average more effective than our other two runs, for topics such as R06-15-I concerning ineffective pledge of a claim under a no-pledge clause, our other two runs were able to find the relevant article (with ID 466) as the top result. The bi-encoder approach retrieves statutes by measuring surface-level similarity between the query and statute text embeddings, so it often favors passages that share common keywords ("pledge," "property," "possession") even when those statutes aren't about prohibitions on pledging or their legal effect. In contrast, the language model–based retrieval captures deeper contextual and logical relationships: it recognizes that prohibiting assignment of a claim and the consequences for a third party with knowledge of that prohibition are directly parallel to prohibiting a pledge of the claim.

Overall, our experimental results show that while our fine-tuned encoder model primarily considers topical relevance, there is still room for improvement by better leveraging LLMs' reasoning capabilities. We also need a more effective mechanism for selecting relevant documents, as using a constant top-k for all queries was not effective.

5 Task 4: Legal Textual Entailment

The Legal Textual Entailment task aims to develop yes/no question-answering systems for legal queries by determining whether relevant Civil Law articles entail the query. Given a legal bar exam question, the system evaluates whether the retrieved content entails the question. The training data consists of query-article-answer triples, while the test data includes only queries and relevant articles without answer labels. The evaluation measure is accuracy, based on whether the yes/no question is correctly answered. This task includes 1,206 training and 74 test queries.

5.1 Proposed Models

For this task, we submitted two runs using LLaMA and Mistral LLMs with a similar approach. With each model, we considered three prompting techniques to decide if a legal article entails the legal question:

(1) **Zero-shot:** The LLM directly predicts Yes/No without any examples, based solely on the candidate legal article and the query.
(2) **Few-shot:** The model is provided with one positive (answer: "Yes") and one negative (answer: "No") example before answering the test input.
(3) **Zero-shot COT:** Similar to the zero-shot approach but incorporating "Let's think step-by-step" to encourage reasoning before answering.

For LLaMA, we included the following system prompt to guide the model's responses:

> You are an expert Japanese lawyer who will decide if a given legal query can be entailed by a given legal article. You will answer with Yes or No. Then, you will provide a brief explanation.

To enhance robustness, we aggregated the model outputs using majority voting, inspired by the approach proposed by Nguyen et al. [14]. Our models were named **AIIRLLaMA** (based on LLaMA) and **AIIRMistral** (based on Mistral).

5.2 Experimental Results

Table 6 shows our results on the COLIEE 2025 test set for Task 4. Our results indicate that *AIIRLLaMA* outperforms *AIIRMistral*, achieving an accuracy of 60.81% compared to 56.76%. This suggests that LLaMA's entailment reasoning capabilities were more effective in this legal domain task. For 30 out of 74 queries, almost 90% or more participating runs had the correct answers, which may indicate these queries were less challenging. Among these 30 queries, both models had only two wrong predictions on the same instances. On the other hand, for 7 queries, less than 10 participating runs were able to predict entailment correctly. While Mistral failed to predict any of these instances correctly, LLaMA predictions for two queries (with Ids R06-20-O and R06-22-A) were correct. LLaMA successfully identified the contextual cues linking the shared obligations and reimbursement rights among guarantors, whereas Mistral failed to do so.

Table 6: Accuracy of AIIR Lab Runs for Task 4 on COLIEE 2025 and previous years Test Sets.

Model	R06 (2025)	R02	R01	H30
AIIRLLaMA	**60.81**	**65.43**	36.04	51.43
AIIRMistral	56.76	64.20	**37.84**	**61.43**

While our proposed models used majority voting, the effectiveness of each prompting technique varied. With LLaMA, both zero-shot and few-shot prompting resulted in an accuracy of 58.11%, while chain-of-thought prompting increased the accuracy to 63.51%. The pattern observed for Mistral differed; with zero-shot prompting, the accuracy was 56.76%, which improved to 59.46% using chain-of-thought prompting. Notably, when few-shot prompting was used, the accuracy further increased to 68.92%. These results suggest that the chosen prompting technique influences the performance of large language models in tasks such as legal entailment. Moreover, exploring alternative ensembling techniques could potentially enhance our majority voting approach. Finally, the varying accuracies observed across different query sets from previous labs indicate that the two language models exhibit distinct strengths depending on the query type. This indicates the need for further investigation into query characteristics and more refined strategies for selecting and fine-tuning LLMs for legal entailment.

6 Conclusion

In conclusion, this paper presents the AIIR Lab's exploration of large language models (LLMs) to enhance legal information retrieval and entailment tasks in the 2025 Competition on Legal Information Extraction and Entailment (COLIEE). Our participation spanned four distinct tasks, each tackling unique challenges and opportunities within legal text processing. We found that models such as Mistral and LLaMA could be effectively leveraged for case summarization and relevance ranking. However, accurately capturing the nuanced context of legal language remains a challenge. As this marks our first participation in COLIEE, several avenues remain unexplored. Future work will focus on optimizing the integration of neural rankers and classification and ranking approaches with LLMs, exploring ensembling techniques, and refining strategies for better semantic alignment.

References

[1] Arian Askari, Suzan Verberne, O Alonso, S Marchesin, M Najork, and G Silvello. 2021. Combining Lexical and Neural Retrieval with Longformer-based Summarization for Effective Case Law Retrieval.. In *DESIRES*. 162–170.
[2] Ricardo Campos, Vítor Mangaravite, Arian Pasquali, Alípio Jorge, Célia Nunes, and Adam Jatowt. 2020. YAKE! Keyword extraction from single documents using multiple local features. *Information Sciences* 509 (2020), 257–289. doi:10.1016/j.ins.2019.09.013
[3] Ilias Chalkidis, Abhik Jana, Dirk Hartung, Michael Bommarito, Ion Androutsopoulos, Daniel Katz, and Nikolaos Aletras. 2022. LexGLUE: A Benchmark Dataset for Legal Language Understanding in English. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, Smaranda Muresan, Preslav Nakov, and Aline Villavicencio

(Eds.). Association for Computational Linguistics, Dublin, Ireland, 4310–4330. doi:10.18653/v1/2022.acl-long.297

[4] Gabriel de Souza P. Moreira, Radek Osmulski, Mengyao Xu, Ronay Ak, Benedikt Schifferer, and Even Oldridge. 2025. NV-Retriever: Improving text embedding models with effective hard-negative mining. arXiv:2407.15831 [cs.IR] https://arxiv.org/abs/2407.15831

[5] Aniket Deroy, Kripabandhu Ghosh, and Saptarshi Ghosh. 2024. Applicability of large language models and generative models for legal case judgement summarization. *Artificial Intelligence and Law* (2024), 1–44.

[6] Aaron Grattafiori et al. 2024. The Llama 3 Herd of Models. arXiv:2407.21783 [cs.AI] https://arxiv.org/abs/2407.21783

[7] Matthew Henderson, Rami Al-Rfou, Brian Strope, Yun hsuan Sung, Laszlo Lukacs, Ruiqi Guo, Sanjiv Kumar, Balint Miklos, and Ray Kurzweil. 2017. Efficient Natural Language Response Suggestion for Smart Reply. arXiv:1705.00652 [cs.CL] https://arxiv.org/abs/1705.00652

[8] Peter Henderson, Mark Krass, Lucia Zheng, Neel Guha, Christopher D Manning, Dan Jurafsky, and Daniel Ho. 2022. Pile of law: Learning responsible data filtering from the law and a 256gb open-source legal dataset. *Advances in Neural Information Processing Systems* 35 (2022), 29217–29234.

[9] Albert Q. Jiang, Alexandre Sablayrolles, Arthur Mensch, Chris Bamford, Devendra Singh Chaplot, Diego de las Casas, Florian Bressand, Gianna Lengyel, Guillaume Lample, Lucile Saulnier, Lélio Renard Lavaud, Marie-Anne Lachaux, Pierre Stock, Teven Le Scao, Thibaut Lavril, Thomas Wang, Timothée Lacroix, and William El Sayed. 2023. Mistral 7B. arXiv:2310.06825 [cs.CL] https://arxiv.org/abs/2310.06825

[10] Joon Ho Lee. 1997. Analyses of multiple evidence combination. In *Proceedings of the 20th annual international ACM SIGIR conference on Research and development in information retrieval*. 267–276.

[11] Haitao Li, Changyue Wang, Weihang Su, Yueyue Wu, Qingyao Ai, and Yiqun Liu. 2023. THUIR@COLIEE 2023: More Parameters and Legal Knowledge for Legal Case Entailment. arXiv:2305.06817 [cs.CL] https://arxiv.org/abs/2305.06817

[12] Antoine Louis, Gijs van Dijck, and Gerasimos Spanakis. 2024. Interpretable long-form legal question answering with retrieval-augmented large language models. In *Proceedings of the AAAI Conference on Artificial Intelligence*, Vol. 38. 22266–22275.

[13] Behrooz Mansouri and Ricardo Campos. 2023. Falqu: Finding answers to legal questions. *arXiv preprint arXiv:2304.05611* (2023).

[14] Chau Nguyen, Thanh Tran, Khang Le, Hien Nguyen, Truong Do, Trang Pham, Son T. Luu, Trung Vo, and Le-Minh Nguyen. 2024. Pushing the Boundaries of Legal Information Processing with Integration of Large Language Models. In *New Frontiers in Artificial Intelligence*, Toyotaro Suzumura and Mayumi Bono (Eds.). Springer Nature Singapore, Singapore, 167–182.

[15] Phuong Nguyen, Cong Nguyen, Hiep Nguyen, Minh Nguyen, An Trieu, Dat Nguyen, and Le-Minh Nguyen. 2024. CAPTAIN at COLIEE 2024: large language model for legal text retrieval and entailment. In *JSAI International Symposium on Artificial Intelligence*. Springer, 125–139.

[16] Rodrigo Nogueira and Kyunghyun Cho. 2020. Passage Re-ranking with BERT. arXiv:1901.04085 [cs.IR] https://arxiv.org/abs/1901.04085

[17] Rodrigo Nogueira, Zhiying Jiang, and Jimmy Lin. 2020. Document Ranking with a Pretrained Sequence-to-Sequence Model. arXiv:2003.06713 [cs.IR] https://arxiv.org/abs/2003.06713

[18] Juliano Rabelo, Randy Goebel, Mi-Young Kim, Yoshinobu Kano, Masaharu Yoshioka, and Ken Satoh. 2022. Overview and Discussion of the Competition on Legal Information Extraction/Entailment (COLIEE) 2021. *The Review of Socionetwork Strategies* 16, 1 (01 Apr 2022), 111–133. doi:10.1007/s12626-022-00105-z

[19] Nils Reimers and Iryna Gurevych. 2019. Sentence-BERT: Sentence Embeddings using Siamese BERT-Networks. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, Kentaro Inui, Jing Jiang, Vincent Ng, and Xiaojun Wan (Eds.). Association for Computational Linguistics, Hong Kong, China, 3982–3992. doi:10.18653/v1/D19-1410

[20] Stephen Robertson and Hugo Zaragoza. 2009. The probabilistic relevance framework: BM25 and beyond. *Foundations and Trends in Information Retrieval* 3, 4 (2009), 333–389.

[21] Julien Rossi and Evangelos Kanoulas. 2019. Legal search in case law and statute law. In *Legal Knowledge and Information Systems*. IOS Press, 83–92.

[22] Wenhui Wang, Furu Wei, Li Dong, Hangbo Bao, Nan Yang, and Ming Zhou. 2020. MiniLM: Deep Self-Attention Distillation for Task-Agnostic Compression of Pre-Trained Transformers. arXiv:2002.10957 [cs.CL] https://arxiv.org/abs/2002.10957