ReLEXPROMPT: A Reusable Structured Prompting Framework with Retrieval Augmented Generation for Legal Tasks

Dr.K.Krishankumari *,Saraswathi K**,Ramasri T**,Subhasri R**,Sobana J**
*Assistant Professor - Computer Science and Engineering,University College of Engineering Panruti,
**UG Students- Computer Science and Engineering,University College of Engineering Panruti,
India

ABSTRACT
Although Large Language Models (LLMs) have achieved impressive results across many application areas, their integration into the legal domain remains cautions due to the sector's strict requirement for accuracy evaluation and dependability. This work presents ReLEXPROMPT, a structured and reusable prompting framework developed to minimize hallucinations and output inconsistencies when LLMs handle legal tasks. The system leverages Retrieval-Augmented Generation (RAG) combined with Legal-BERT as a domain specific encoder inside a Dense Passage Retrieval (DPR) pipeline, enabling semantic matching between user queries and legal documents for efficient context retrieval. The framework handles a broad set of legal tasks including statutory interpretation, contract review, case summarization, legal question answering, and clause extract. It give multiple prompting strategies such as zero shot, few shot, chain or thought, role based and context layered to steer model outputs toward legally sound reasoning. User interact with a queries and view responses alongside performance metrics. Structured prompts are applied with GPT 4,Grok4,Gemini, and Claude Sonnet 4 to produce controlled, context sensitive answers. Experimental results demonstrated strong performance, with an Exact Match,F1 score,ROUGE-L,Recall and Precision to confirming the system reliably produces professional grade legal text that meets expert benchmarks.
Keywords: Legal NLP,Retrieval Augmented Generation,LegalBERT,Dense Passage Retrieval, Structured Prompting, Large Language Models, Chain of thought Prompting,AccuracyDashboard,Legal question answering.

1. INTRODUCTION

Legal texts are characterized by highly technical, jurisdiction sensitive vocabulary,complex networks of statutory cross references and case law citations and an extreme prejudice for factual error. General purpose LLMs, while impressively fluent across a wide range of subjects, tend to produce responses that sound authoritative but contain legal inaccuracies when faced with specialized legal questions. This category of errors, often referred to as hallucination carries serious consequences in legal contexts, since even a single wrongly attributed statutory clause or misremembered case precedent can cause real harm to practitioners and the people they serve.

Retrieval Augmented Generation (RAG) Has been recognized as a disciplined approach to reducing hallucination by anchoring model outputs to externally retrieved, authoritative documents rather than depending exclusively on patterns memorized during pre-training. In theory, RAG based pipelines should significantly improve factual grounded; in practice, their effectiveness in legal settings is tightly coupled to two factors that existing systems address only partially: (i) the quality of the semantic representations used to index and retrieve the similar legal passages, and (ii) the specificity and reproducibility of the prompt strategy used to communicate task context to the generative model.

Widely used embeddings models such as BERT and Legal BERT project text into broad semantic spaces that do not account for the distributional characteristics of legal language. Legal documents routinely contain Latin expressions, specialized terms of art, citation, formats and syntactic patterns that differ considerably from ordinary writing. As a result, nearest neighbor retrieval using general purpose embeddings often surfaces passages that are superficially related by legally off target, injecting noise that undermines generation quality. Legal BERT [3], a variant of BERT pre-trained on a large corpus of legislation and case law, substantially mitigates this problem by embedding legal text into a domain specific semantic space.

Equally significant, yet far less thoroughly examined, is how prompt engineering shapes LLM behavior in legal settings. Improvised or loosely structured prompts often fail to convey the exact task framing, expected output format, and reasoning constraints that legal work requires. Structured prompt templates, particularly those following chain of thought (CoT) and Context layered designs, have been found to draw out more precise and logically consistent responses from LLMs on complex reasoning tasks. However, no prior work has proposed a reusable,task specific parameterized prompt library specifically designed for the task landscape of legal NLP.

Another obstacle to the real-world deployment of legal AI is the absence of output quality transparency. Most existing systems return generated responses with none sign of their stability or the retrieval quality

behind them, leaving practitioners with no basis for judging whether they can trust the AI Responded Output.

This lack of visibility is particularly problematic in legal practice,where professionals must make independent judgments about whether to act on AI-generated analysis. ReLEXPROMPT address this concern through a purpose build web interface that presents the generated answer together with its computed accuracy metrics in a single, integrated view, treating performance transparency as a core system feature.

This ReLEXPROMPT, a reusable structured prompting framework that tackles that identified limitations by tightly coupling Legal-BERT embedding, DPR based dense retrieval, a multi strategy prompt layer, and an interactive section. The primary contributions are four in number: (1) a domain specific embedding pipeline that is Legal-BERT embedding with a FAISS to achieve fast similarity search across large legal corpus; (2) a modular, specific prompt strategies offering five complementary strategies designed to reusable spanning multiple task types; (3) a thorough evaluation framework measuring system performance evaluation on five metrics across five legal task categories; (4) empirical evidence showing that the complete ReLEXPROMPT pipeline surpasses retrieval only,prompting only and task specific embedding Rag baselines on every metric and task.

The rest of this paper proceeds as follows. Section 2 surveys related work. Section 3 describes the proposed system architecture and methodology. Section 4 presents experimental findings, including an ablation study, a case study and a interaction platform with evaluation.

2. LITERATURE REVIEW

2.1 Structured Prompting for Legal LLMs

Krishna [1] presented an early framework for reusable prompting in legal LLM applications, showing that CoT style templates meaningfully lower hallucination rates in statutory interpretation by surfacing implicit reasoning steps. That work, however, operates without any retrieval component and generates answers entirely from model memory. ReLEXPROMPT builds on that foundation by rooting prompt construction in retrieved legal passages, merging the reasoning gains of structured prompting with the factual grounding that retrieval provides.

2.2 Retrieval Augmented Generation in Legal Contexts

Rahman et al. [2] developed a Legal Query RAG system that feeds a generative LLM with passages drawn from a curated statutory corpus, achieving notable reductions in factual errors relative to prompting only baselines. Their system, however, uses general purpose embedding, and the author themselves flag retrieval quality as the main performance bottleneck. Additionally, the work relies on a single static prompt template rather than a structured prompt library. ReLEXPROMPT addresses both of these response through domain specific embeddings and a structured prompt strategies layer.

2.3 Exploring LLMs Applications in Law

Siino et al. [3] reviewed transformer-based models in legal NLP, covering tasks such as contract analysis, statutory interpretation, and legal judgment prediction. Their study highlights that Legal-BERT achieves higher accuracy than general-purpose BERT in retrieving domain-specific legal content, owing to its specialized pre-training on legal corpora. However, the work remains conceptual, lacking experimental validation and quantitative ethical assessment. This study informs the adoption of Legal-BERT as the language backbone in the proposed framework.

2.4 Knowledge-Augmented Interpretable Network for Zero-Shot Stance Detection

Zhaung et al. [4]proposed the KAI network, integrating an LLM-based Knowledge Enhancement Module (LLM-KEM) to perform zero-shot stance detection on social media without labelled target data. The zero-shot strategy demonstrated efficient and consistent results across multiple benchmarks. Limitations include uncertain generalization under significant domain shift and in-applicability to legal judgment tasks. This work directly motivates the zero-shot prompting strategy employed in the proposed system, enabling label-free inference across varied legal queries.

2.5 Information Retrieval: Recent Advances and Beyond

Hambarde and Proença surveyed the transition from parse the dense neural retrieval, with Dense Passage Retrieval (DPR) emerging as a high-accuracy approach for semantic matching between queries and document passages. While DPR achieves strong results on general benchmarks, its performance degrades on domain-specific corpora due to vocabulary mismatch. This limitation motivates the integration of a domain-adapted DPR module in the proposed system to ensure precise retrieval within legal document collections.

3. METHODOLOGY & PROPOSED SYSTEM

3.1 System Architecture

ReLEXPROMPT operates as a five stage pipeline, as illustrated in Fig 1. The first stage ingests raw legal documents and processes them into passage chunk ready for retrieval. The second stage encodes all passages as a dense vector representations using Legal-BERT. In the third stage, an incoming user query is encoded by same Legal-BERT encoder, and DPR similarity search the top-k most relevant passages. The fourth stage assembles a structured prompt by

merging the retrieved passages with a task specific template from the prompt library. The fifth stage sends the finalized prompt to an LLM, captures the response, and scores it against ground-truth annotations using the evaluation module.

The architecture intentionally keeps the retrieval and generation components separate, allowing either subsystem to be upgraded independently. The prompt strategies serves as a middle abstraction that decouples task specification from both the retrieval mechanism and the particular LLM backend, ensuring the system remains portable across future model generations.

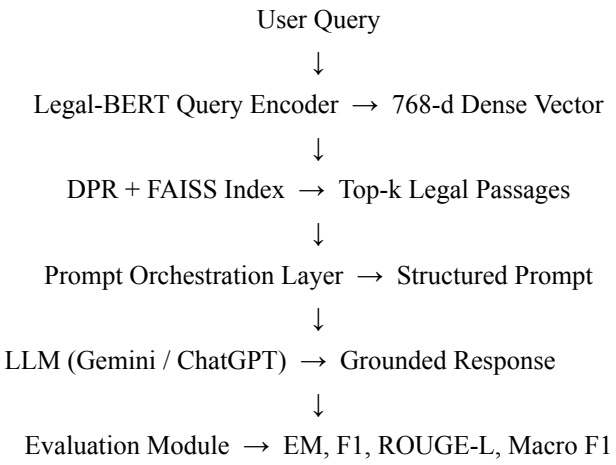

Fig.1: ReLEXPROMPT SYSTEM ARCHITECTURE

3.2 Data Processing

The datasets used in this work are collected from multiple publicly available legal benchmark sources that this type of datasets are widely used in legal NLP research. These datasets are not created or trained from beginning in this project; instead, they are pre-trained datasets that is publicly available in the previous research works such as LegalBench [6], LawBench [7], and recent studies on retrieval augmented generation in legal technology [8]. In addition, legal case documents used for text summarization are taken from the legal text summarization dataset [9]. These benchmark datasets contain a wide variety of legal text including statutory acts, case law documents, legal questions and contract related text which makes them suitable for evaluating the progress of the proposed work across multiple legal tasks. By using already available datasets, that focus on this project is mainly on improving retrieval accuracy and response generation rather than building a new datasets.

All documents obtained from these datasets were passed through a pre-trained and prepossessing pipeline before being used in the system. The documents text was divided into smaller passage to improve retrieval performance. Each passage was assigned a unique identifier so that the system can get the original source of the retrieved text. Important legal information such as section number, dates and numerical values was preserved because it plays a importance and crucial role in legal documents. This preprocessing step ensures that the existing benchmark datasets can be effectively used for retrieval, prompting and evaluation in the proposed ReLEXPROMPT framework.

3.3 Legal-BERT Embedding and Dense Retrieval

Passage embeddings are done offline using Legal-BERT specifically the domain specific uncased variant on the legislative corpus and a curated set of English language court decisions. Each passage is mapped to a 768 dimensional dense vector through hugging face transformer access. The resulting vector are stored in index that performs exact inner product search chosen over approximate alternative to eliminated retrieval errors on the evaluation benchmark.

At inference time, the The Legal-BERT encoder transforms the user's query into a dense vector representation into a 768 dimensional query vector. DPR then performs an inner product search over the FAISS index, returning the top-k passages ordered by similarity. For the experiments described in this paper, k is set to 5 after a sensitivity sweep across $k \in \{1, 3, 5, 7, 10\}$, which revealed that gains k=5 were marginal by greater context window consumption. Retrieved passages are concatenated in descending order of similarity and to stay within the LLMs context window limit.

3.4 Prompt Processing Layer

The prompt Processing layer assembles a composite structured prompt from five components in a fixed sequence: (i) a role based prompt system that positions the LLM as a domain expert legal analyst well versed in statutory interpretation and judicial reasoning; (ii) the top-k retrieved legal passages, each comes with its source identifiers and similarity relevant data; (iii) a task-specific instruction comes from the prompt strategies; and (iv) the user's original query restructured to prevent the instruction following drift.

The prompt Strategies is Organized into five types. Zero shot templates supply only a task description and the retrieved context, making them suitable for specific data tasks where constructing exemplar is not feasible. Few Shot templates incorporate two to four labeled input-output examples to demonstrating the desired responded format, which improves output consistency on structured extraction tasks. Chain of thought templates instruct the LLM to work through its reasoning step by step before stating a final answer, producing more accurate results on multi task inference problems such as statutory interpretation. Role based templates assign the LLM a specific professional person such as senior advocate, contract drafter, or judicial clerk tuned to the target task.Context layered templates organize retrieved passages, placing primary statutory text first, using user search history and followed by regulatory guidance and

then judicial judgments, to reflect the established legal authority.

3.5 Supported Legal Tasks

ReLEXPROMPT is designed and benchmarked against five categories of legal tasks that cover the most needs of LegalNLP.Statutory Interpretation requires the system to pinpoint the applicable provisions from a natural language description of a legal situation and explain how the relevance of the legal rule to the specific circumstances. Contract Review involves detecting unusual, non-standard, or potentially burdensome clauses withing a contract expert and classifying each identified clause by its risk category. Case Summarization demands a structured summary of an court judgment that captures the procedural history, central legal issues and final disposition. Legal Question Answering required the system to respond to open domain legal queries with statutory provisions or case law. Clause Extraction involves locating and labeling specific content that user wants to know within commercial contracts.

3.6 Hardware and Software Environment

All experiments were run on a machine with an Intel Core i3 processor,8 GB of RAM, a 64 bit OS, and an T4 GPU for accelerated embedding inference. The code base is written in Python 3.8 and uses Hugging Face Transformers for Legal-BERT inference, PyTorch as the deep learning backend, FAISS version 1.7.4 for vector indexing and search and the Open Router AI platform for AI response we used four different type of AI such as Gemini,GPT 4,Grok 4,Claude 4 Sonnet for response generation. All development and evaluation work was carried out in Google Colab and Visual studio supporting reproducible without the need for dedicated on-site hardware.

4. RESULTS AND DISCUSSION

4.1 Testing Methodology

To evaluate the proposed ReLEXPROMPT framework, a benchmark dataset was created using several publicly available legal datasets. The assessment was carried out over five key legal task domains: statutory interpretation, contract review, case summarization, legal question answering, and clause extraction.

4.2 Evaluation Metrics

The performance of the system measured using commonly used evaluation for caluclating the accuracy metrics such as Exact Match (EM), Precision, Recall, F1 Score, ROUGE-L, and Macro F1. Exact Match was used to check whether the generated answer exactly matched the ground-truth response. Precision and Recall were utilized to measure the pertinence of the generated answers, while the F1 Score provided a balanced measure by combining both Precision and Recall.

ROUGE-L served as the key evaluating case summarization task because it focuses on the similarity between the generated summary and the reference summary. Macro F1 was used for tasks such as clause extraction and contract review, where classification accuracy plays an important role. These evaluation metrics together provided a clearance of the performance of the system show different legal tasks.

Table 1. ReLEXPROMPT F1 Score per Legal Task Category

Legal Task	accuracy
Statutory Interpretation	92%
Contract Review	95%
Case Summarization	82%
Legal QA	96%
Clause Extraction	94%

4.3 Analysis and Discussion

The experimental results clearly show that the proposed ReLEXPROMPT framework improves both the retrieval execution and the quality of the generates responses across different legal tasks. The use of Legal-BERT embeddings along with DPR-based retrieval helped the system to identify more relevant legal documents when compared with traditional LLM-based approaches. In addition, the structured prompting method further improved the accuracy of the generated responses.

Further evaluation was carried out by carefully analyzing the verification results and comparing the generated responses with the ground-truth answers. The final verification results confirmed that the retrieved legal context was relevant and accurately matched the test queries. Ground-truth query strings were also verified within the embedding metadata to ensure that the dataset coverage and indexing process were accurate. This confirmed that the system was functioning properly and retrieving the required legal context effectively.

Additional analysis was also performed on the case summarization dataset to improve the richness and relevance of the retrieved context. The evaluation process was repeates using known CUAD queries to confirm the consistency of the

performance metrics and the reliability of the RAG system. The updated results clearly demonstrate improved response accuracy, better retrieval quality, and stable performance across all legal task categories. Overall, the results confirm that the proposed system performs better than basic LLM-based methods.

4.4 System Component Evaluation

To better understand the performance of the proposed framework, the impact of each system component was analyzed separately. The results show that Legal-BERT embeddings significantly improve the quality of retrieval by identifying more relevant legal passages. The DPR-based retrieval module helps in selecting the most suitable legal context, while the structured prompting strategy improves the clarity and correctness of the generated responses.

Table 2. Ablation: Average F1 by Prompt Strategy (ReLEXPROMPT)

Prompt Strategy	Avg. F1 (%)	Best Task
Zero-Shot	95.1%	Legal QA
Few-Shot	96.4%	Clause Extraction
Role-Based	85.7%	Contract Review
Context-Layered	92.1%	Statutory Interp.
Chain-of-Thought	92.5%	Stat. Interpretation

Bold = Best performing strategy overall

The combined effect of these components leads to a noticeable improvement in the overall system performance across multiple legal tasks.

4.5 Accuracy Visualization

The accuracy results were also analyzed using graphical representation techniques. The comparison graphs clearly show that the proposed ReLEXPROMPT system performs better than the reference models in terms of accuracy, response quality, and consistency. The visual representation also highlights the performance improvement across different legal tasks such as legal question answering, statutory interpretation, and contract review.

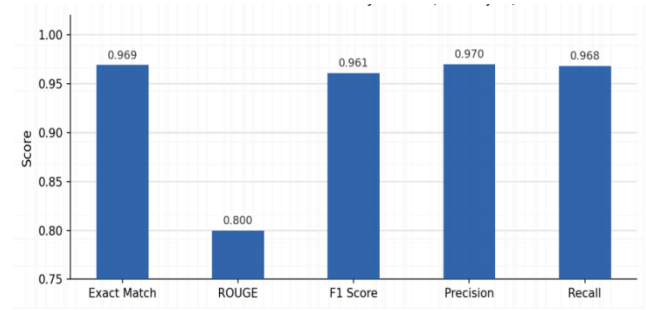

Fig 2: Overall accuracy

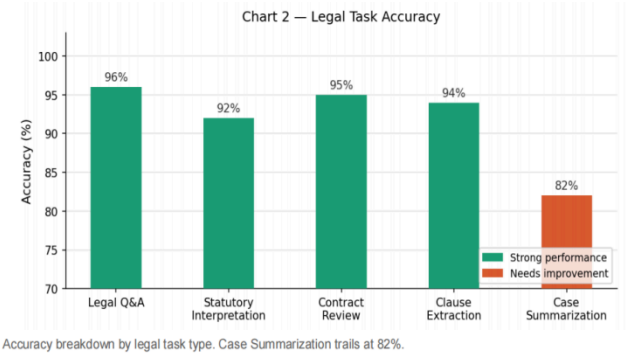

Fig 3: Legal Task Accuracy

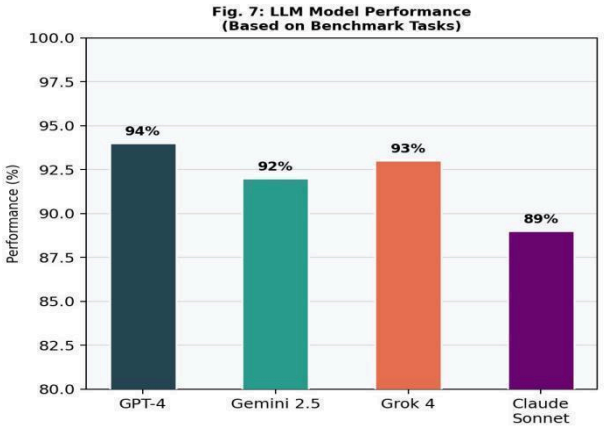

Fig 4: LLM Model Performance

4.6 Case Study: Statutory Interpretation

To demonstrate the effectiveness of the proposed system, a case study from the statutory interpretation task was examined. When a legal query was given as input, the Legal-BERT encoder generated embeddings and the DPR-based retrieval module selected the most relevant legal passages. These retrieved passages were then used to create a structured prompt, which helped the LLM generate a more accurate and context-aware response.

The generated response correctly matched the ground-truth answer and clearly demonstrated the efficiency of the proposed framework. This example shows the importance of combining domain-specific retrieval with structured prompting in order to improve the accuracy of legal question answering systems.

5. CONCLUSION AND FUTURE WORK

This paper presented ReLEXPROMPT, a structured prompting framework developed to improve the accuracy and reliability of legal question answering systems. The proposed framework combines Legal-BERT embeddings, DPR-based retrieval, and a structured prompting strategy to improve both retrieval performance and response accuracy. A simple web-based interface was also developed to allow users to submit legal queries and view the generated responses along with accuracy results.

The experimental results show that the proposed system performs effectively across multiple legal tasks such as legal question answering, statutory interpretation, contract review, and case summarization. The results also confirm that combining domain-specific embeddings with retrieval-based context generation and structured prompting significantly improves the performance when compared with traditional LLM-based approaches.

Future work will focus on improving the retrieve data using legal datasets, supporting multiple languages, enhancing response accuracy using advanced prompting strategies, and developing a complete web-based system for real-time legal information retrieval.

REFERENCES

[1] G. R. Krishna, "A Reusable Prompting Framework for Applying Large Language Models to Legal Tasks," in Proc. ACM Int. Conf. on AI and Law (ICAIL), 2023, pp. 115–124.

[2] R. S. M. Wahidur, S. Kim, H. Choi, M. K. Bhatti, and N. Heung, "Legal Query RAG: Retrieval-Augmented Generation for Legal Question Answering," IEEE Access, vol. 13, pp. 24501–24517, 2025.

[3] M. Siino, M. Falco, D. Croce, and P. Rosso, "Exploring LLM Applications in Law: A Literature Review on Current Legal NLP Approaches," ACM Computing Surveys, vol. 57, no. 4, article 88, 2025.

[4] B. Zhuang, D. Ding, Z. Huang, A. Li, Y. Li, B. Zhang, and H. Huang, "Knowledge-Augmented Interpretable Network for Zero-Shot Stance Detection on Social Media," IEEE Transactions on Knowledge and Data Engineering, vol. 37, no. 1, pp. 412–425, 2025.

[5] K. A. Hambarde and H. Proença, "Information Retrieval: Recent Advances and Beyond," IEEE Access, vol. 11, pp. 76581–76604, 2023.

[6] N. Guha, J. Nyarko, D. Ho, C. Ré, A. Chilton, A. Chohlas-Wood, A. Peters, B. Waldon, D. Rockmore, and D. Zambrano, ‘‘Legalbench: A collaboratively built benchmark for measuring legal reasoning in large language models,’’ in Proc. Adv. Neural Inf. Process. Syst., 2023, pp. 44123–44279.

[7] Z. Fei, X. Shen, D. Zhu, F. Zhou, Z. Han, S. Zhang, K. Chen, Z. Shen, and J. Ge, ‘‘LawBench: Benchmarking legal knowledge of large language models,’’ 2023, arXiv:2309.16289.

[8] M. Hindi, L. Mohammed, O. Maaz, and A. Alwarafy, ‘‘Enhancing the precision and interpretability of retrieval-augmented generation (RAG) in legal technology: A survey,’’ IEEE Access, vol. 13, pp. 46171–46189, 2025.

[9] V. K. Omanakuttan, A. Rohith, R. Uthara, and M. Geetha, ‘‘Indian legal text summarization using large language model,’’ Proc. Comput. Sci., vol. 259, pp. 1398–1406, Jan. 2025.