Data Science and Engineering (DSE) Record, Volume 6, issue 1.

Development of a Retrieval-Augmented Generation System for Legal Data in Thai Language

Pimchanok Promwang [1] and Pruet Boonma [2]

[1] Master's Degree Program in Data Science, Chiang Mai University, Chiang Mai, Thailand
[2] Department of Computer Engineering, Faculty of Engineering, Chiang Mai University, Chiang Mai, Thailand
pimchanok_promwang@cmu.ac.th

Abstract. This study presents a Retrieval-Augmented Generation (RAG) framework tai-lored for Thai legal question answering. The system integrates sparse retrieval (BM25), dense retrieval (SentenceTransformer), and a hybrid approach combin-ing both methods with dynamic weighting. To enhance contextual relevance, a BGE-based re-ranking model was employed. Experiments were conducted on a Thai legal dataset (WangchanX-Legal-ThaiCCL-RAG), and performance was evaluated using Recall@K, Precision@K, MAP, and ROUGE-L. Results showed that while dense retrieval outperformed sparse retrieval in most metrics, the hybrid method—augmented by re-ranking—yielded the highest retrieval accuracy at low K values, with Recall@1 reaching 73.3%. Although this approach introduced additional processing time, the system remained near real-time in response. In the answer generation phase, the model achieved an average ROUGE-L score of 0.4742 (0.6067 when excluding zero-score cases), indicating moderate alignment between generated and reference answers. The findings suggest that hybrid retrieval with reranking improves legal information access in Thai, providing a reproducible baseline for future research in legal question answering for low-resource languages.

Keywords: Retrieval-Augmented Generation, Legal Information Retrieval, Hybrid Retrieval, Thai Legal NLP, Natural Language Processing

1 Introduction

Large language models (LLMs) have transformed natural language processing (NLP), delivering remarkable advances in tasks like summarization, question answering, and dialogue generation. While these models demonstrate impressive fluency, they continue to face significant challenges, including hallucination, opacity in their reasoning processes, and restricted access to current or verifiable external knowledge sources.

To tackle these limitations, Lewis et al. (2020) introduced the Retrieval-Augmented Generation (RAG) framework [1], which integrates external retrieval mechanisms into the generation pipeline. RAG works by retrieving relevant documents from a knowledge base and providing them to the model along with the input query, effectively anchoring the generated responses in factual information. This architectural approach

has demonstrated considerable success in enhancing accuracy and minimizing hallucinated content across diverse knowledge-intensive applications.

The RAG methodology has increasingly attracted interest within specialized domains, particularly in scientific question answering, medical summarization, and legal information retrieval—fields where source verification and traceability are paramount. Nevertheless, its deployment in low-resource languages like Thai faces substantial constraints. The unique linguistic characteristics of Thai, including the absence of explicit word boundaries and the intricacy of domain-specific terminology, create additional hurdles for both document retrieval and text generation processes.

This research presents a Retrieval-Augmented Generation approach specifically designed for Thai legal question answering. We examine three distinct retrieval methodologies—sparse retrieval using BM25, dense retrieval through embedding techniques, and a hybrid combination of both approaches—complemented by a re-rankng component utilizing the BGE CrossEncoder. Our evaluation employs a Thai legal dataset (WangchanX-Legal-ThaiCCL-RAG) and incorporates both retrieval and generation evaluation metrics, with the goal of establishing a robust and reproducible foundation for advancing Thai legal NLP research.

2 Related Works

2.1 Retrieval-Augmented Generation in Legal and Multilingual Contexts

Following the introduction of Retrieval-Augmented Generation (RAG), numerous studies have explored ways to tailor and optimize this framework for legal question answering (QA). Recognizing the limitations of generic retrieval strategies in the legal domain, Pipitone and Houir Alami (2024) [2] proposed LegalBench-RAG, the first benchmark explicitly designed to evaluate retrieval accuracy rather than generation quality. By aligning queries with highly relevant text spans in large legal corpora, this benchmark emphasizes precision over recall, highlighting the importance of snippet-level retrieval in mitigating hallucination and latency issues.

Extending the architecture further, Peng and Chen (2024) [3] introduced Athena, a framework for legal judgment prediction that combines semantic retrieval with prompt-based reasoning. Their system uses query rewriting, a knowledge base of accusations, and dense retrieval to guide the LLM in structured judgment prediction. Athena achieved state-of-the-art accuracy on the CAIL2018 dataset, showing that retrieval-aware prompting improves performance in classification-style legal tasks.

In the context of low-resource languages, Nguyen et al. (2024) [4] developed a Vietnamese legal QA system that integrates BM25, dense retrieval, and an RRF-style re-ranking mechanism. They proposed a method called Active Retrieval, which improves the ordering of retrieved documents before generation. Their experiments demonstrated improved reliability and user satisfaction, even without the need to fine-tune LLMs.

In Chinese legal counseling, Xie et al. (2024) [5] presented DeliLaw, an end-to-end QA system combining two-stage retrieval: BGE-based dense retrievers for statutes, and ElasticSearch for precedent cases. DeliLaw employs fine-tuned domain-specific embeddings, achieving 71.1% recall and 61.6% MRR, while reducing hallucination by providing grounded legal texts from an up-to-date law database.

Lastly, Wiratunga et al. (2024) [6] proposed CBR-RAG, which integrates Case-Based Reasoning (CBR) into the RAG pipeline. Rather than relying solely on dense similarity, CBR-RAG retrieves past legal QA pairs from a curated casebase, encoding cases with dual embeddings (intra- and inter-representations). Their evaluations on the Australian Legal QA dataset showed that this structured retrieval improved factual grounding and interpretability in generated responses.

Collectively, these works illustrate a shift from conventional sparse retrieval to-ward hybrid, semantically enriched, and task-specific retrieval strategies. However, such systems have been primarily evaluated in high-resource languages. Their applicability to under-resourced languages like Thai, which pose unique linguistic and retrieval challenges, remains largely underexplored.

2.2 Thai Legal Question Answering and Retrieval Challenges

While multilingual adaptations of RAG have progressed, Thai legal QA remains a relatively underexplored area due to linguistic complexity and limited benchmarks. Challenges include the absence of whitespace segmentation in Thai, the intricate hierarchical structure of legislation, and frequent inter-section references that compli-cate retrieval and chunking strategies.

One notable contribution is Sommai, developed by VISTEC AI (2024) [7], which leverages a dense retriever based on BGE-M3 embeddings and reranks results using the BGE CrossEncoder. The system uses a Thai legal LLM (LLaMa3.1–8B-Legal-ThaiCCL) finetuned on WangchanX data. Experimental results showed improved performance in Recall@5 and MRR@5 after applying both embedding tuning and reranking. However, the system does not incorporate sparse or hybrid retrieval strategies, and its evaluation remains focused on retrieval effectiveness rather than full end-to-end QA performance.

A more comprehensive study is presented in NitiBench by Akarajaradwong et al. (2025) [8], which provides a benchmark suite covering general financial laws (CCL) and complex tax law cases. The framework introduces domain-aware enhancements such as hierarchical-aware chunking, NitiLink for reference expansion, and a custom evaluation protocol involving multi-label retrieval metrics and LLM-as-a-judge for end-to-end output scoring. The results demonstrate that section-based chunking sig-nificantly improves performance, but current retrievers still struggle with complex queries, especially in the tax dataset. Importantly, RAG-based setups outperformed long-context LLMs, confirming that RAG remains more suitable for Thai legal QA under current model capabilities.

Despite these advances, there remains no prior work has systematically compared sparse, dense, and hybrid retrieval under a unified architecture for Thai legal question answering. Latency, precision, and stability in a near real-time context also remain underreported. This study aims to fill that gap through systematic evaluation and reproducible design

3 Data and Methodology

This chapter describes the overall methodological approach used in this study, including the datasets, research framework, retrieval method, evaluation metrics, and tools and environment.

3.1 Dataset

This study employs the WangchanX-Legal-ThaiCCL-RAG dataset, a Thai legal question answering corpus tailored for developing and evaluating Retrieval-Augmented Generation (RAG) systems. Rich in structured legal content, it is well-suited for both retrieval and generation tasks in the legal domain.

The dataset comprises 8,210 training and 3,740 test entries. Training data supports index construction, while the test set is reserved for evaluation. Each entry includes

- **question**: a legal inquiry expressed in natural Thai language
- **positive_contexts**: one or more legal text passages that provide supporting information
- **hard_negative_contexts**: Irrelevant sections from BGE-M3 retrieval
- **positive_answer**: the correct answer derived from the relevant context
- **hard_negative_answer**: Initial uncorrected answer before expert validation

In particular, a general data preparation process was applied prior to separating the data into sparse and dense retrieval pipelines. This step includes whitespace normalization, punctuation cleanup, and extraction of legal text from the structured positive_contexts field, which is originally stored as a list of dictionaries. To ensure consistency, multiple context passages were concatenated into a single string per entry for indexing and retrieval.

As shown in Table 1, the raw format contains both legal content and metadata, whereas the preprocessed version retains only the normalized legal text. This cleaned version is used as input for both BM25 and dense embedding models in later stages.

positive_contexts (raw)	positive_contexts (after preprocessing)
{'context': 'พระราชบัญญัติธุรกิจสถาบันการเงิน พ.ศ. 2551 มาตรา 8 ให้รัฐมนตรีว่าการกระทรวงการคลังรักษาการตามพระราชบัญญัตินี้ และให้มีอำนาจออกประกาศเพื่อปฏิบัติการตามพระราชบัญญัตินี้\nประกาศตามวรรคหนึ่งเมื่อได้ประกาศในราชกิจจานุเบกษาแล้วให้ใช้บังคับได้\nรัฐมนตรีอาจกำหนดให้ธนาคารแห่งประเทศไทยยื่นรายงานข้อมูลที่ได้รับจากการดำเนินการตามรายการที่รัฐมนตรีกำหนด ทั้งนี้ จะให้ยื่นตามระยะเวลาหรือเป็นครั้งคราวและจะให้ทำคำชี้แจงข้อความเพื่ออธิบายหรือขยายความแห่งรายงานนั้นก็ได้', **'metadata': {'law_code': 'ธ0012-1B-0001', 'law_title': 'พระราชบัญญัติธุรกิจสถาบันการเงิน พ.ศ. 2551', 'section': '8'}, 'unique_key': 'ธ0012-1B-0001-8'}**	พระราชบัญญัติธุรกิจสถาบันการเงิน พ.ศ. 2551 มาตรา 8 ให้รัฐมนตรีว่าการกระทรวงการคลังรักษาการตามพระราชบัญญัตินี้ และให้มีอำนาจออกประกาศเพื่อปฏิบัติการตามพระราชบัญญัตินี้ ประกาศตามวรรคหนึ่งเมื่อได้ประกาศในราชกิจจานุเบกษาแล้วให้ใช้บังคับได้ รัฐมนตรีอาจกำหนดให้ธนาคารแห่งประเทศไทยยื่นรายงานข้อมูลที่ได้รับจากการดำเนินการตามรายการที่รัฐมนตรีกำหนด ทั้งนี้ จะให้ยื่นตามระยะเวลาหรือเป็นครั้งคราวและจะให้ทำคำชี้แจงข้อความเพื่ออธิบายหรือขยายความแห่งรายงานนั้นก็ได้

Table 1: Example of preprocessed positive_contexts field

3.2 Research Framework

This study focuses on the development of an experimental Retrieval-Augmented Generation (RAG) framework for Thai legal question answering. The goal is to evaluate and optimize retrieval strategies—namely sparse, dense, and hybrid retrieval methods—and examine their influence on the quality of generated answers. The overall workflow of the system is illustrated in Figure 1.

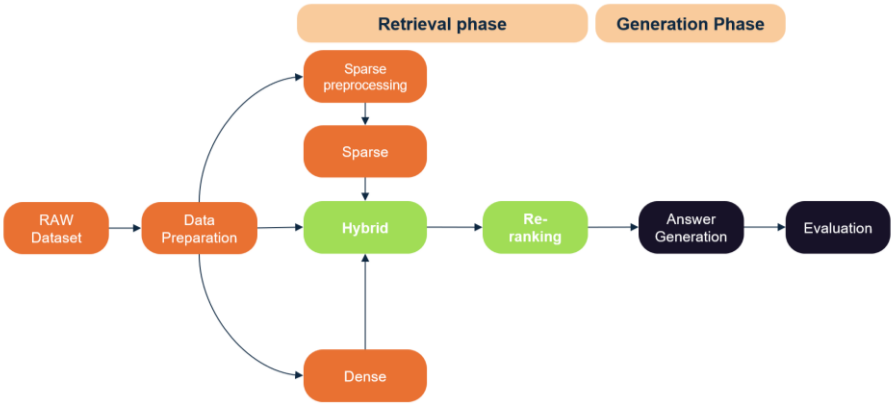

Figure 1: Overview of the experimental workflow.

The process begins with a curated Thai legal QA dataset, which is first preprocessed to extract structured inputs such as questions, contexts, and answers. Based on this data, two retrieval pipelines are constructed in parallel. The first applies sparse retrieval using BM25, while the second performs dense retrieval using SentenceTransformer embeddings and FAISS-based semantic search.

Following this, both retrieval outputs are merged using score-level fusion in a hybrid retrieval stage to leverage the strengths of each approach. To further improve relevance, the top-k contexts are re-ranked using BGE Re-ranker (bge-reranker-v2-m3), a cross-encoder model that refines the ranking based on question-passage relevance.

Finally, the re-ranked contexts are passed to a legal-domain language model to generate responses. The system is evaluated using both retrieval metrics (Recall@K, Precision, MAP) and a generation metric (ROUGE-L). This integrated framework supports a structured and consistent evaluation of retrieval strategies within the Thai legal domain.

3.3 Retrieval Method

3.3.1 Sparse Retrieval Method

The sparse retrieval method in this study was implemented using the BM25 algorithm, a well-established technique for lexical matching. BM25 was selected due to its strong performance in prior legal-domain benchmarks such as LegalBench-RAG (Pipitone & Houir Alami, 2024) and NitiBench (Akarajaradwong et al., 2025), where it served as a competitive baseline for retrieving statutory texts.

To enhance performance in Thai legal contexts, each passage was preprocessed using Thai word segmentation (newmm from PyThaiNLP) followed by stopword removal. The remaining tokens were rejoined into whitespace-separated strings and stored in a new field used for indexing. This preprocessing significantly reduced noise and improved retrieval granularity.

All training passages were indexed using the Rank-BM25 implementation. During inference, questions were processed through the same pipeline, and BM25 scores were computed based on lexical overlap. For each query, the top 5 matching passages were retrieved to balance relevance and computational efficiency. Duplicate results were allowed to capture overlapping legal clauses.

positive_contexts	positive_contexts_tokenized
พระราชบัญญัติทะเบียนพาณิชย์ พ.ศ. 2499 มาตรา 15 เมื่อได้จดทะเบียนพาณิชย์แล้ว ให้ผู้ประกอบพาณิชยกิจจัดให้มีป้ายชื่อที่ใช้ในการประกอบพาณิชยกิจไว้ที่หน้าสำนักงานแห่งใหญ่และสำนักงานสาขาโดยเปิดเผย ภายในสามสิบวันนับแต่วันที่ได้จดทะเบียนป้ายชื่อนี้ให้เขียนเป็นอักษรไทย อ่านได้ง่ายและชัดเจน และจะมีอักษรต่างประเทศด้วยก็ได้ ทั้งนี้ ไม่ว่าจะกระทำบนแผ่นไม้ แผ่นโลหะ แผ่นกระจก กำแพง หรือผนัง ชื่อในป้ายก็ดี ในเอกสาร ใด ๆ ก็ดี ต้องใช้ให้ตรงกับชื่อที่จดทะเบียนไว้ และถ้าเป็นสำนักงานสาขา ต้องมีคำว่า “สาขา” ไว้ด้วย (498 Characters)	พระราชบัญญัติ ทะเบียน พาณิชย์ พ.ศ. 2499 มาตรา 15 จดทะเบียน พาณิชย์ พาณิชย กิจ ป้ายชื่อ พาณิชย กิจ หน้า สำนักงาน สำนักงานสาขา สามสิบ วันที่ จดทะเบียน ป้ายชื่อ อักษร ไทย อ่าน ชัดเจน อักษร ต่างประเทศ แผ่น ไม้ แผ่น โลหะ แผ่น กระจก กำแพง ผนัง ชื่อ ป้าย เอกสาร ใด ชื่อ จด ทะเบียน สำนักงานสาขา (43 Characters)

Table 2: Example of tokenization and stopword removal applied to `positive_contexts `

After preprocessing, all training passages were indexed using the Rank-BM25 implementation. Each query was tokenized and cleaned using the same pipeline, and BM2 5 scores were calculated based on token overlap between the query and each passage. The system retrieved the top 5 scoring passages per query, balancing retrieval quality and efficiency, and allowing duplicates to preserve overlapping legal clauses that might contain the answer.

```
    คำถามที่ใช้ค้นหา:
การไถ่ถอนจำนองต้องทำเป็นหนังสือหรือไม่

    ข้อความที่ควรเป็นคำตอบ (Ground Truth):
ประมวลกฎหมายแพ่งและพาณิชย์ มาตรา 737
ผู้รับโอนจะไถ่ถอนจำนองเมื่อใดก็ได้ แต่ถ้าผู้รับจำนองได้บอกกล่าวว่าจะบังคับจำนอง ผู้รับโอนต้องไถ่ถอนจำนองภายในหกสิบวันนับแต่วันรับคำบอกกล่าว

================================================================================
    เอกสารที่ BM25 ค้นคืนมา:

    อันดับ 1
    ข้อความที่ BM25 พบ:
ประมวลกฎหมายแพ่งและพาณิชย์ มาตรา 737
ผู้รับโอนจะไถ่ถอนจำนองเมื่อใดก็ได้ แต่ถ้าผู้รับจำนองได้บอกกล่าวว่าจะบังคับจำนอง ผู้รับโอนต้องไถ่ถอนจำนองภายในหกสิบวันนับแต่วันรับคำบอกกล่าว

    BM25 Score: 18.4392
    ผลการตรวจสอบ:      ตรงกับเฉลย
--------------------------------------------------------------------------------

    อันดับ 2
    ข้อความที่ BM25 พบ:
ประมวลกฎหมายแพ่งและพาณิชย์ มาตรา 737
ผู้รับโอนจะไถ่ถอนจำนองเมื่อใดก็ได้ แต่ถ้าผู้รับจำนองได้บอกกล่าวว่าจะบังคับจำนอง ผู้รับโอนต้องไถ่ถอนจำนองภายในหกสิบวันนับแต่วันรับคำบอกกล่าว

    BM25 Score: 18.4392
    ผลการตรวจสอบ:      ตรงกับเฉลย
--------------------------------------------------------------------------------

    อันดับ 3
    ข้อความที่ BM25 พบ:
ประมวลกฎหมายแพ่งและพาณิชย์ มาตรา 714
อันสัญญาจำนองนั้น ท่านว่าต้องทำเป็นหนังสือและจดทะเบียนต่อพนักงานเจ้าหน้าที่

    BM25 Score: 17.0719
    ผลการตรวจสอบ:      ไม่ตรงกับเฉลย
```

Figure 2: Example of retrieved passages ranked by BM25 score, showing exact match evaluation for top-ranked results.

3.3.2 Dense Retrieval Method

In addition to lexical retrieval, this study employed a dense retrieval method using the WangchanX-Legal-ThaiCCL-Retriever model. This model is based on the SentenceTransformer architecture and was fine-tuned from the BGE-M3 base on the WangchanX-Legal-ThaiCCL-RAG dataset. It encodes both questions and legal contexts into dense vector representations with 1,024 dimensions, allowing semantic similarity to be measured more effectively than exact lexical overlap.

To build the index, each legal context passage was first preprocessed and then transformed into dense embeddings. These embeddings were stored using FAISS, a library optimized for fast vector search. Specifically, the IndexFlatIP method was used to enable inner product (cosine similarity) comparisons.

At query time, each question was likewise encoded into a dense vector. Cosine similarity scores were then computed between the query vector and all vectors in the FAISS index. The top-K passages with the highest similarity scores were retrieved as candidate contexts for answer generation.

	Dim 0	Dim 1	Dim 2	Dim 3	Dim 4
Vector Index					
0	0.010149	0.005079	-0.033197	0.028278	-0.003744
1	0.026021	0.017284	-0.019525	0.005075	-0.009952
2	0.027700	0.005105	-0.005459	0.009456	-0.037527
3	0.023131	0.037432	-0.018286	0.027331	-0.014963
4	0.018601	-0.017001	-0.018697	0.015934	-0.017880

Figure 3: Sample dense vector representations stored in the FAISS index (showing 5 of 1,024 dimensions).

3.3.3 Hybrid Retrieval Method

To combine the strengths of sparse and dense retrieval approaches, this study adopted a hybrid retrieval pipeline with re-ranking. For each query, the retrieval process followed the same preprocessing and encoding procedures described in Sections 3.3.1 and 3.3.2. The top-5 results from both BM25 and the WangchanX-Legal-ThaiCCL-Retriever model were merged to form a candidate pool, with duplicates retained to preserve overlapping legal content.

A hybrid score was calculated for each candidate passage by combining the normalized scores from BM25 and the dense retriever. Dynamic weighting was applied to determine the contribution of each method based on two factors: the relative confidence

(retrieval score) of each model and the retrieval cutoff k. Specifically, at k = 1, a higher weight was assigned to the dense retriever to prioritize semantic precision. For larger k values, a softmax-based weighting was used to balance the lexical signals from BM25 and the semantic understanding from the dense retriever. This approach enabled the hybrid mechanism to adapt to different retrieval needs while maintaining both precision and coverage.

To further refine retrieval accuracy, the candidate list was re-ranked using the BAAI/bge-reranker-v2-m3 model. This CrossEncoder-based technique was chosen in line with prior legal-domain studies reviewed in Chapter 2, which highlighted the effectiveness of re-ranking for contextual relevance. Furthermore, the re-ranker shares the same architecture family as the dense retriever, ensuring semantic consistency throughout the pipeline.

After re-ranking, the top-1 passage was selected for answer generation. Retrieval effectiveness was evaluated at multiple top-k thresholds (e.g., @1, @3, @7) using Recall, Precision, and Mean Reciprocal Rank (MAP).

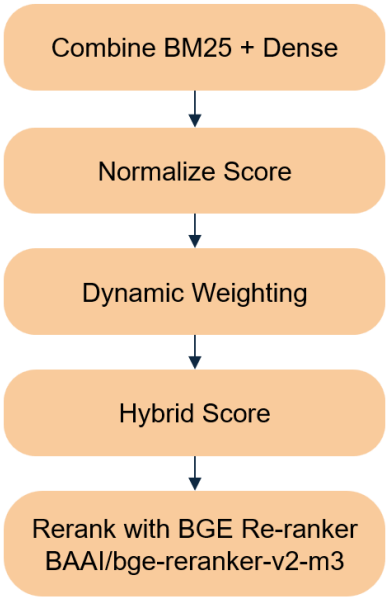

Figure 4: Hybrid Retrieval Pipeline

3.4 Answer Generation Using LLM

This stage focuses on generating answers using a large language model (LLM) conditioned on the retrieved legal context. The goal is to produce coherent and legally grounded responses to user questions, based on top-ranked passages obtained from the hybrid retrieval module described earlier.

For each question, a prompt was constructed by combining the legal context and the original question into an instruction-like format. The structure of the prompt was as follows:

- Context: {positive_contexts}
- Question: {question}
- Answer:

This format was applied uniformly across all samples in the test set. The {positive_contexts} component was drawn from the top-ranked passage following the BGE re-ranking step in the hybrid retrieval pipeline.

To generate the answer, the airesearch/LLaMa3.1-8B-Legal-ThaiCCL-Combine model was used. This 8-billion-parameter LLM was pre-trained and fine-tuned on Thai legal corpora and accessed via the Hugging Face Transformers library. The model was used in zero-shot mode without additional fine-tuning or adaptation.

Text generation was performed using the Hugging Face text-generation pipeline, with key parameters including max_new_tokens=200 and do_sample=False to ensure consistent, deterministic outputs. After generation, the output text was post-processed to isolate the answer portion, discarding prompt echoes or formatting artifacts.

The generated answer for each instance was then stored alongside its corresponding reference answer (positive_answer) for evaluation, which is described in the following section. No further heuristics or formatting rules were applied in the answer generation process.

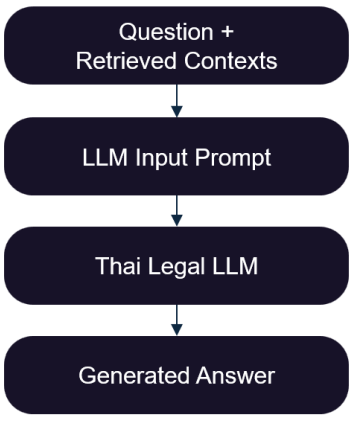

Figure 5: Answer Generation Pipeline

Component	Content
Question	When can a company promoter offer shares to the public?
Reference Answer	After the registration of the memorandum of association, according to Section 23 of the Public Limited Companies Act B.E. 2535 (1992), once the registrar has accepted the documents, the company promoter may offer shares to the public or any individuals.
Generated Answer	According to Section 23 of the Public Limited Companies Act B.E. 2535 (1992), a company promoter can offer shares to the public after the registrar has accepted the registration documents under Section 24. The offering can proceed without additional conditions, as long as the law is complied with.

Table 3: Example of a legal question, reference answer, and generated response.
Note: This table shows an English translation of sample content from the original Thai legal QA dataset.

3.5 Evaluation Metrics

The assessment was divided into two main components: (1) retrieval evaluation, which focuses on how effectively relevant legal contexts are retrieved, and (2) generation evaluation, which examines the quality of answers generated by the language model.

3.5.1 Retrieval Evaluation

The retrieval performance was assessed using three standard metrics—Recall@K, Precision@K, and Mean Average Precision (MAP)—which collectively capture both coverage and ranking quality of retrieved results.

• **Recall@K** measures how effectively the approach retrieves relevant legal information by quantifying the proportion of truly relevant items appearing in the retrieved list. The general formula is:

$$Recall = \frac{TP}{TP + FN}$$

In this study's context, the formula becomes:

$$Recall@K = \frac{Number\ of\ relavant\ items\ retrieved\ \ witin\ top-K}{Total\ number\ of\ relevant\ items}$$

• **Precision@K** indicates the proportion of relevant passages within the top-K retrieved results:

$$Precission = \frac{TP}{TP + FP}$$

For this study:

$$Precission@K = \frac{Number\ of\ relavant\ items\ retrieved\ witin\ top-K}{K}$$

• **Mean Average Precision (MAP)** measures the average precision across multiple relevant documents within the top-K retrieved results for each query:

$$MAP@K = \frac{1}{N}\sum_{i=1}^{N} AP\ @K_i$$

Where N represents the total number of questions and denotes the average of precision values at ranks where relevant documents appear within the top-K results.

Each retrieval method was evaluated across multiple cut-off thresholds (K = 1, 3, and 7) using these metrics before proceeding to answer generation. Additionally, this study measured each method's latency to assess computational efficiency, defined as the time required to complete retrieval for a given legal question. Average latency was recorded and compared across sparse, dense, and hybrid approaches, with results analyzed in Chapter 4.

3.5.2 Answer Generation Evaluation

Answer quality was evaluated using the ROUGE-L metric, which measures similarity between generated and reference answers based on the Longest Common Subsequence (LCS). This metric is particularly suitable for legal texts, as it tolerates minor phrasing variations while capturing the same legal intent.

Due to computational constraints, evaluation was conducted on randomly selected subsets rather than the full 3,740-question dataset. Two subsets of 300 and 600 samples were used in separate runs to assess consistency while maintaining computational efficiency, balancing evaluation reliability with resource limitations.

Each generated answer was compared against the reference answer from the dataset's positive_answer field without post-editing or formatting adjustments. All answers were generated using the top-1 re-ranked passage retrieved by the hybrid retriever.

The ROUGE-L score employs F-measure between LCS-based precision and recall, computed as:

$$ROUGE\text{-}L = F_\beta = \frac{(1+\beta^2)\cdot Precision \cdot Recall}{Recall + \beta^2 \cdot Precision}$$

Where:

$$Precision = \frac{LCS(X,Y)}{length\ of\ X}$$

and:

$$Recall = \frac{LCS(X,Y)}{length\ of\ Y}$$

- X is the generated answer
- Y is the reference answer
- $LCS(X,Y)$ is the length of the longest common subsequence
- β is typically set to 1

By focusing on the longest common subsequence, ROUGE-L effectively captures content similarity while accommodating linguistic variation, making it well-suited for evaluating legal answer generation.

3.6 Tools and Environment

The experiments in this study were conducted using Python version 3 on Google Colab Pro, a Software as a Service (SaaS) platform that enables Python-based development and execution directly in a web browser. The platform supports GPU acceleration, and this study specifically utilized an NVIDIA A100 GPU to improve performance during model retrieval, re-ranking, and answer generation processes.

4 Result

This chapter presents the experimental results in two parts: retrieval performance and answer generation quality. It compares the outcomes across methods and concludes with key findings.

4.1 Retrieval Performance

Retrieval performance was evaluated using Precision@K, Recall@K, and Mean Average Precision (MAP), with K set to 1, 3, and 7. These metrics assess both the relevance and ranking effectiveness of each method. Three retrieval strategies were compared:

- Sparse retrieval using BM25
- Dense retrieval with the WangchanX-Legal-ThaiCCL-Retriever
- Hybrid retrieval combining BM25 and dense scores through dynamic weighting, followed by BGE CrossEncoder re-ranking.

Table 4 presents the average scores across 3,740 legal questions. At K = 1, the hybrid method achieved the highest scores (0.7338) on all metrics, outperforming both BM25 and dense retrieval. As K increased to 3 and 7, Recall improved for all methods, while Precision and MAP declined. Notably, sparse retrieval showed the steepest drop in precision, whereas the hybrid method maintained stable performance and achieved the highest MAP at K = 3. Overall, the hybrid approach consistently delivered the best results across most metrics.

Metric	BM25 K=1	DENSE K=1	Hybrid K=1	BM25 K=3	DENSE K=3	Hybrid K=3	BM25 K=7	DENSE K=7	Hybrid K=7
Precision@K	0.5409	0.7312	0.7338 ✅	0.2464	0.3088	0.2703 ✅	0.1200	0.1416	0.1244 ✅
Recall@K	0.5409	0.7312	0.7338 ✅	0.6943	0.8656	0.7606 ✅	0.7859	0.9238	0.8151 ✅
MAP@K	0.5409	0.7312	0.7338 ✅	0.2464	0.3088	0.2703 ✅	0.1200	0.1416	0.1244 ✅

Table 4: Retrieval performance comparison

In addition to accuracy, retrieval latency was measured to evaluate computational efficiency. Table 5 shows the average time per query across all test cases. BM25 and dense retrieval required 0.0229 and 0.0236 seconds per query, respectively. In contrast, the hybrid method averaged 0.0496 seconds due to added processing from score fusion and re-ranking.

Retrieval Method	Average Time per Query (s)
Sparse Retrieval (BM25)	0.0229
Dense Retrieval	0.0236
Hybrid Retrieval + Re-rank	0.0496

Table 5: Average Retrieval Latency (in seconds) by Method

4.2 Answer Generation Performance

As outlined in Chapter 3, answer generation was evaluated on random subsets of 300 and 600 samples from the full set of 3,740 legal questions due to computational constraints. In both cases, the top-1 passage retrieved by the hybrid method featuring dynamic weighting and BGE re-ranking—was used as context for the LLaMa3.1-8B-Legal-ThaiCCL-Combine model. Performance was measured using the ROUGE-L metric, which compares generated and reference answers.

To establish a baseline, the 300-sample subset was first evaluated, yielding an average ROUGE-L score of 0.4775 (all samples) and 0.5993 (excluding zero-score cases), as shown in Figure 6.

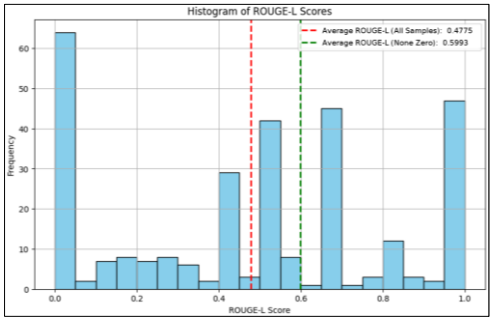

Figure 6: Histogram of ROUGE-L scores for 300 samples.

To assess scalability and consistency, a larger 600-sample subset was then tested, producing similar results: 0.4742 overall and 0.6067 after excluding zero scores (Figure 7).

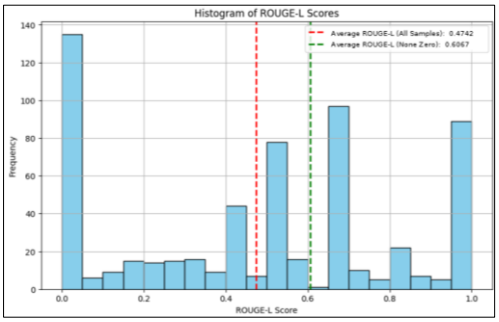

Figure 7: Histogram of ROUGE-L scores for 600 samples.

Overall, the consistent performance across both subsets indicates reliable scalability of the evaluation. Although some answers had limited overlap with the ground truth, many exhibited moderate to high similarity—particularly when the retrieved context was relevant.

4.3 Summary of Key Findings

• Sparse retrieval (BM25) provided fast retrieval times but underperformed in both Recall and MAP metrics, especially at higher K values. Its reliance on exact lexical matching limited its effectiveness in semantically complex queries.

• Dense retrieval outperformed BM25 across all retrieval metrics, particularly in Recall@K, indicating better semantic alignment with legal queries.

• Hybrid retrieval, which combined BM25 and dense scores using dynamic weighting and BGE-based re-ranking, achieved the best overall performance. The improvement was most evident at K = 1, where it reached 0.7338 across Precision, Recall, and MAP.

• Retrieval latency for the hybrid method was approximately double that of individual retrieval strategies, reflecting the additional computational steps required for scoring and re-ranking.

• Answer generation, evaluated using ROUGE-L, yielded an average score of 0.4742 on a 600-sample evaluation. When excluding zero-overlap cases, the score increased to 0.6067.

• The score distribution showed a bimodal pattern, suggesting that the system produced highly relevant answers in some cases, but failed to align in others—likely due to retrieval noise or generation limitations.

These findings highlight the trade-offs between retrieval accuracy and system efficiency, while confirming the benefit of hybrid strategies in improving legal QA performance under low-resource conditions.

5 Conculsion and Discussion

5.1 Conculsion

This study explored a retrieval-augmented generation (RAG) framework for Thai legal question answering, emphasizing the evaluation of various retrieval strategies in a low-resource setting. To enhance retrieval quality, a hybrid approach was developed by integrating sparse and dense retrieval through dynamic weighting, followed by re-ranking with a BGE cross-encoder.

Experimental results indicated that the hybrid method slightly outperformed standalone BM25 and dense retrieval in Recall, Precision, and MAP—particularly at lower K values. Although consistent, these gains were modest. Additionally, the hybrid method nearly doubled latency due to the added scoring and re-ranking steps.

For answer generation, a Thai legal language model generated responses based on the top-1 retrieved context. ROUGE-L evaluation on randomly selected test samples yielded average scores around 0.47 across all cases, and above 0.60 when excluding

zero-overlap outputs. While some responses showed limited overlap with reference answers, many demonstrated strong alignment when relevant contexts were provided.

In summary, the hybrid retrieval pipeline improved legal information access in Thai, though gains were incremental. The observed trade-offs between accuracy and latency, along with moderate generation scores, underscore the need for continued refinement of both retrieval and generation modules in future research.

5.2 Discussion

The findings of this study offer key insights into the effectiveness of retrieval strategies in a Retrieval-Augmented Generation (RAG) framework for Thai legal question answering. As expected, dense retrieval significantly outperformed sparse retrieval (BM25) across all metrics—Recall\@K, Precision\@K, and MAP—confirming trends observed in prior research such as NitiBench (Akarajaradwong et al., 2025) and Sommai (VISTEC AI, 2024), which highlighted the strength of semantic embeddings in capturing legal-specific phrasing and implicit meaning.

The hybrid retrieval method—combining sparse and dense scores via dynamic weighting and re-ranking with the BGE model—achieved the highest overall performance. This improvement was most evident at lower K values (e.g., K = 1), where precision is critical, aligning with results from Vietnamese Legal QA (Nguyen et al., 2024), which also demonstrated early-stage recall gains using hybrid scoring and re-ranking. However, the performance advantage over dense retrieval alone was modest, suggesting that dense embeddings already retrieve most relevant contexts effectively.

This improvement in accuracy came with a trade-off in latency. The hybrid approach nearly doubled the average retrieval time per query compared to standalone methods. This echoes findings from DeliLaw (Xie et al., 2024), where dual-module retrieval increased system complexity. Consequently, for real-time applications—such as legal chatbots or public search tools—latency remains a critical limitation to balance with accuracy.

In the generation phase, the average ROUGE-L score across 600 samples was 0.4742, increasing to 0.6067 when zero-score outputs were excluded. These results indicate moderate lexical alignment with reference answers but fall short of scores achieved using golden passages. For instance, Sommai (VISTEC AI, 2024) reported a ROUGE-L of 0.715 using ideal contexts with dense retrieval and BGE reranking. This contrast highlights the challenge of fully automated RAG pipelines, where retrieval quality inherently limits generation accuracy. Unlike Sommai, this study used natural, unstructured legal questions without curated context inputs.

In terms of retrieval effectiveness, this study's hybrid method reached Recall\@1 of 73.2%, a competitive figure under realistic conditions. However, it trails behind systems like Athena (Peng & Chen, 2024), which achieved Recall\@10 of 89.1% and nDCG of 86.4% on the CAIL2018 dataset using dense retrieval with query rewriting. While CBR-RAG (Wiratunga et al., 2024) did not report quantitative results, it

demonstrated qualitative gains via semantic similarity in case-based retrieval. Compared to Sommai's Recall\@5 of 89.8%, the hybrid model in this study performed reasonably well, particularly considering it was not fine-tuned and operated in an open-domain setting.

Overall, the results support the value of hybrid retrieval with reranking for legal applications. The architecture introduced here offers a practical foundation for advancing Thai legal QA systems. Although its performance lags behind pipelines using golden contexts, the system's realism and reproducibility are strengths. Future directions may include query rewriting, multi-step legal reasoning, and human-in-the-loop evaluation to enhance both interpretability and legal reasoning in retrieval and generation.

6 Future Work

Building on the findings of this study, future work can be organized into five key areas as outlined below.

Improving Retrieval Effectiveness

To enhance retrieval quality, future work may incorporate query rewriting or reformulation techniques. Transforming user queries into clearer, more structured forms—via templates or language models—can reduce ambiguity and improve matching, especially for underspecified legal questions. Additionally, analyzing the impact of high-frequency legal terms and applying techniques like term weighting or stopword adjustment could help refine retrieval precision in domain-specific corpora.

Enhancing Evaluation Methodologies

While this study relied on random subsets, evaluating performance on the full test set would reduce sampling bias and enable more granular error analysis across diverse question types. Moreover, beyond automated metrics like ROUGE-L, integrating human or expert-based evaluations can provide deeper insights into semantic correctness, legal appropriateness, and reasoning consistency—factors not captured by surface-level similarity scores.

Extending Reasoning and Context Scope

To handle questions requiring broader legal context, future systems should support multi-passage retrieval and longer reasoning chains. Such capabilities would enable step-by-step interpretation across related statutes—critical in legal scenarios yet currently underexplored in low-resource RAG setups. Further, incorporating graph-based retrieval—modeling legal provisions as nodes and their relationships (e.g., citations, themes) as edges—could improve multi-provision question answering by retrieving coherent statute clusters.

Supporting Real-World Legal Applications

Moving toward deployment, RAG systems should be integrated into interactive or user-facing tools, such as legal research platforms or chat-based advisors. This would allow real-time feedback, enhance usability, and make the system more applicable in

practical legal workflows. Although this study focused on backend performance, frontend design and user experience are crucial for broader adoption.

Expanding Dataset Representation and Structure

Future research may explore expanding datasets to explicitly encode inter-provision relationships, thematic clusters, and citation networks. Such structured data would better support advanced retrieval algorithms, including graph-based and multi-hop approaches. In doing so, RAG systems could be better positioned to handle complex, interconnected legal queries that require reasoning across multiple documents.

References

1. Lewis, P., Perez, E., Piktus, A., Petroni, F., Karpukhin, V., Goyal, N., ... & Riedel, S. (2020). Retrieval-augmented generation for knowledge-intensive NLP tasks. *Advances in Neural Information Processing Systems, 33,* 9459–9474. https://proceedings.neurips.cc/paper/2020/hash/6b493230205f780e1bc26945df7481e5-Abstract.html
2. Pipitone, T., & Houir Alami, S. (2024). *A benchmark for retrieval-augmented generation in the legal domain*. arXiv preprint arXiv:2402.08149. https://arxiv.org/abs/2402.08149
3. Peng, Y., & Chen, B. (2024). *Athena: Retrieval-augmented legal judgment prediction with large language models*. In Proceedings of the Chinese Conference on Artificial Intelligence and Law (CAIL 2024), 55–66.
4. Nguyen, L. T., Tran, H. M., & Pham, K. (2024). *Vietnamese legal information retrieval in question-answering system*. In Proceedings of the 2024 International Conference on AI and Law in Southeast Asia (AIL-SEA 2024), 22–33.
5. Xie, Z., Li, M., & Zhang, Q. (2024). *DeliLaw: A Chinese legal counselling system based on a large language model*. In Proceedings of the 2024 Chinese Conference on Legal AI Applications, 88–97.
6. Wiratunga, N., Massie, S., & Moffat, D. (2024). *CBR-RAG: Case-based reasoning for retrieval-augmented generation in LLMs for legal question answering*. In Proceedings of the 2024 International Workshop on Case-Based Reasoning for Legal NLP, 45–54.
7. Thitiwat Nopparatbundit. (2024, November 30). *สมหมาย: แชทบอทด้านกฎหมาย Technical Blog*. Medium. https://medium.com/@thitiwat/สมหมาย-แชทบอทด้านกฎหมาย
8. Akarajaradwong, P., Pothavorn, P., Chaksangchaichot, C., Tasawong, P., Nopparatbundit, T., & Nutanong, S. (2025). *NitiBench: A Comprehensive Studies of LLM Frameworks Capabilities for Thai Legal Question Answering*. arXiv preprint arXiv:2502.10868. https://arxiv.org/abs/2502.10868