ThaiLegal: Benchmarking LLM Frameworks on Thai Legal Question Answering Capabilities

Anonymous ACL submission

Abstract

Large language models (LLMs) show promise in legal question answering (QA), yet Thai legal QA systems face challenges due to limited data and complex legal structures. We introduce ThaiLegal, a novel benchmark featuring two datasets: (1) ThaiLegal-CCL , covering Thai financial laws, and (2) ThaiLegal-Tax , containing Thailand's official tax rulings. Our benchmark also consists of specialized evaluation metrics suited for Thai legal QA. We evaluate retrieval-augmented generation (RAG) and long-context LLM (LCLM) approaches across three key dimensions: (1) the benefits of domain-specific techniques like hierarchy-aware chunking and cross-referencing, (2) comparative performance of RAG components, e.g., retrievers and LLMs, and (3) the potential of long-context LLMs to replace traditional RAG systems. Our results reveal that domain-specific components slightly improve over naive methods. At the same time, existing retrieval models still struggle with complex legal queries, and long-context LLMs have limitations in consistent legal reasoning. Our study highlights current limitations in Thai legal NLP and lays a foundation for future research in this emerging domain.

1 Introduction

Large language models (LLMs) offer significant potential for transforming the legal domain, particularly in information retrieval and question-answering. LLM-powered tools are emerging to assist legal professionals with research, including conversational search, summarization, and citation analysis (LexisNexis, 2023; Strumberger, 2023; Takyar). Legal question-answering (QA) systems can provide accessible legal information to the public, often using retrieval-augmented generation (RAG) frameworks (ailawyer; asklegal.bot). These systems must understand complex queries, retrieve relevant legal documents, apply the information, and generate accurate, contextually appropriate responses.

A major challenge in Thai legal QA is ensuring that systems can (i) accurately retrieve the correct legal texts relevant to a given query and (ii) apply those retrieved texts to generate well-grounded responses. Existing Thai legal QA systems, such as Thanoy[1] (Viriyayudhakorn, 2024) exhibit various errors, such as incorrect legislation retrieval and hallucination. Given such limitations in Thai legal QA, we address a major bottleneck in evaluating Thai legal QA systems: the lack of standardized evaluation processes, which can be attributed to the scarcity of Thai legal QA corpora and inadequate evaluation metrics for multi-document retrieval and end-to-end (E2E) performance. We introduce a novel benchmark, ThaiLegal, consisting of two datasets and metrics designed specifically for Thai Legal QA systems.

We further use our benchmark to examine limitations in today's LLM frameworks, such as RAG and Long Context Language Models (LCLMs). Our results reveal limitations in existing retrievers and LLMs for complex legal reasoning, particularly with the ThaiLegal-Tax dataset. Our benchmark and findings aim to facilitate systematic progress in Thai legal NLP.

Our key contributions include:

- **Two Thai QA Dataset for Legal QA**: *ThaiLegal-CCL Dataset* covers general financial law, while the *ThaiLegal-Tax Dataset* specifically focuses on complex tax cases. Each query includes a question, answer, and relevant documents for detailed retrieval and E2E evaluation. We named our benchmark, which consists of two datasets and proposed metrics (shown in §3.2), as **ThaiLegal**.
- **Tailored Metrics for Thai Legal QA**: We propose multi-label retrieval metrics and E2E

[1] https://iapp.co.th/thanoy

metrics that assess accuracy, consistency, and legal citation quality.

- **Comprehensive Analysis**: Using ThaiLegal, we evaluate industry-standard RAG components, addressing three key research questions: **(RQ1)** How can chunking strategies that are tailored to the hierarchical nature of the Thai legal system and a section[2] referencing component improve performance? **(RQ2)** How do retriever and LLM choices impact RAG performance? **(RQ3)** How do long-context LLM (LCLM) based Thai legal QA systems perform compared to RAG-based approaches?

2 Related Works

Legal QA Benchmarks. Benchmarking legal QA systems is crucial for standardized evaluation. Existing English benchmarks such as LexGlue (Chalkidis et al., 2022) and LegalBench (Guha et al., 2023) address various subtasks (e.g., court opinion classification, contract NLI), but often fall short in evaluating end-to-end open-question-answering performance of RAG systems. Recent works (Dahl et al., 2024; Magesh et al., 2024; Es et al., 2023) introduce multiple aspects for evaluating open-domain QA tasks in retrieval-augmented generation (RAG), with a strong emphasis on faithfulness, groundedness, and relevance of the generated answers. As for the retrieval evaluation, to the best of our knowledge, no prior work has developed multi-label variants of traditional retrieval metrics (such as hit rate, MRR, and recall), which are inadequate for capturing the inherent multi-label nature of the legal reasoning process.

RAG in Legal Practice. RAG approaches enhance LLM outputs by incorporating relevant legal texts (Lewis et al., 2021; Wiratunga et al., 2024). Despite promising applications in commercial systems like Lexis+ AI (LexisNexis, 2023), Westlaw (Strumberger, 2023), and Thanoy (Viriyayudhakorn, 2024), hallucination and retrieval accuracy remain problematic (Magesh et al., 2024).

RAG vs Long-Context LLMs. An alternative, Long-Context LLMs (LCLMs), can process extended texts without separate retrieval (Laban et al., 2024; Lee et al., 2024b; Reid et al., 2024). However, while LCLMs offer advantages in context length, studies have found them less effective than RAG for tasks requiring precise citation and comprehensive coverage (Kamradt, 2023; Bai et al., 2024; An et al., 2023; Lee et al., 2024b; Li et al., 2024; Phan et al., 2024)—especially in the legal domain. Our work directly compares RAG and LCLM approaches for Thai legal QA, addressing this important gap.

3 Methodology

In §3, we outline **ThaiLegal** comprising two datasets: **ThaiLegal-CCL** and **ThaiLegal-Tax** . We also cover the evaluation framework of ThaiLegal for Thai legal QA systems, addressing retrieval and end-to-end (E2E) performance.

Formally, given the set of sections L extracted from ThaiLegal-CCL , both formats can be represented as $\mathcal{D} = \{(\mathbf{x}_1, y_1), ..., (\mathbf{x}_N, y_N)\}$, where $\mathbf{x}_i = (q_i, T_i \subset L)$ - q_i denotes query or question, T_i is a set of positive documents (sections) corresponded to q_i. The label y_i is the free-form text answer to question q_i given the context T_i.

3.1 Datasets

3.1.1 ThaiLegal-CCL Dataset

ThaiLegal-CCL (**C**orporate and **C**ommercial **L**aw) is a general Thai financial law QA dataset with 35 pieces of legislation, including a test set for evaluation. ThaiLegal-CCL was derived from XYZ-ThaiLegal-CCL's test set with an additional postprocessing step where we utilize an LLM to extract only the essential answers without the accompanying rationale. The test set only contains a subset of 21 out of 35 pieces of legislation. These legislation are then parsed into sections, resulting in L.

For training data, we use original XYZ-ThaiLegal-CCL training set which contains multiple positives (See Appendix A for more details on XYZ-ThaiLegal-CCLdata curation). Note that the test set contains only single positives. Further details on ThaiLegal-CCL data curation can be found in Appendix B

3.1.2 ThaiLegal-Tax Dataset

ThaiLegal-Tax is a specialized dataset for Thai tax rulings. It includes 50 cases from 2021-2024, with questions, answers, and referenced sections scraped from the Revenue Department of Thailand's website[3]. This dataset only contains a test

[2]In this paper, "section" refers to a component in legislation, while we use "§" to denote a section, subsection, or subsubsection in this document. For more information on Thai legal terminology, see Appendix E.

[3]https://www.rd.go.th

set and is multi-labeled ($|T_i| \geq 1$). We also filtered any relevant section to ensure that the law cited in this dataset matches the set L used in ThaiLegal-CCL as well. For additional information on the ThaiLegal-Tax data curation process, please refer to Appendix C

3.2 Metrics

3.2.1 Retriever Metrics

We adapt traditional retrieval metrics for multi-label scenarios suitable for multi-label setup in our benchmark. Formally, let N be the number of samples in a dataset, k denote the number of top retrieved documents being evaluated, T_i represent the set of positive relevant documents, and R_i^k denote the top-k ranked retrieved documents.

HitRate@k. Measures if any relevant document is retrieved can be defined as:

$$\text{HitRate@k} = \frac{1}{N} \sum_{i=1}^{N} \mathbb{I}(R_i^k \subseteq T_i) \tag{1}$$

Multi-HitRate@k. Requires all relevant documents to be retrieved and is defined as:

$$\text{Multi-HitRate@k} = \frac{1}{N} \sum_{i=1}^{N} \mathbb{I}(T_i \subseteq R_i^k) \tag{2}$$

Recall@k. Evaluates the proportion of relevant documents retrieved defined as:

$$\text{Recall@k} = \frac{1}{N} \frac{\sum_{i=1}^{N} |T_i \cap R_i^k|}{\sum_{i=1}^{N} |T_i|} \tag{3}$$

MRR@k. Assess ranking quality defined by:

$$\text{MRR@k} = \frac{1}{N} \sum_{i=1}^{N} \frac{1}{\text{argmax}(T_i \cap R_i^k)} \tag{4}$$

where $\text{argmax}(T_i \cap R_i^k)$ represents the highest rank number of correctly retrieved documents. The metric is zero if $|T_i \cap R_i^k| = 0$ (retrieved document contains no positive).

MultiMRR@k. Traditional MRR is calculated under the assumption that any of the documents in the ground truth set T is considered a positive label (Zhan et al., 2020; Khattab and Zaharia, 2020). However, this assumption is not true, especially in a legal domain where, sometimes, all relevant laws must be retrieved for the system to be able to answer the question. Therefore, the equation 4 is augmented to MultiMRR as follows:

$$\begin{aligned} \text{MultiMRR@k} = \frac{1}{N} \sum_{i=1}^{N} \Bigg[\frac{\text{Recall@k}_i}{|T_i \cap R_i^k|} \\ \times \sum_{j=1}^{|T_i \cap R_i^k|} \frac{1}{\text{rank}(d_j) - j + 1} \Bigg]. \end{aligned} \tag{5}$$

3.2.2 End-to-End Metrics

We design three complementary metrics to assess end-to-end answer quality and legal grounding:

Coverage. Following (Kamradt, 2023); the coverage score measures the semantic alignment between generated and ground truth answers via a 3-point scale:

- 100: Full coverage (all key points in ground truth addressed)
- 50: Partial coverage (≥1 key point missing)
- 0: No meaningful overlap

Citation. Evaluating precision, recall, and F1 for cited sections following (Kamradt, 2023).

Contradiction. Quantifying hallucination by comparing generated answers to ground truth as a binary (1=contradiction, 0=consistent).

Both citation and contradiction scores are computed using LLM-as-a-judge, where we use `gpt-4o-2024-08-06` (Hurst et al., 2024) as a judge model with a temperature of 0.3. We also tune our prompt to ensure that the judge LLM achieves a high agreement with humans. The details on judge LLM performance are outlined in Appendix D.

4 Experimental Setups

In §4, we outline our experimental setup using our proposed benchmark to address three key research questions.

The LLM prompts are provided with 3-shot examples randomly sampled from the training data. All experiments were conducted on a single DGX A100 node (40GB, 4 GPUs) for both retriever fine-tuning and LLM inference.

4.1 (RQ1) Impact of Tailored Components

For this research question, we aim to address the impact of injecting domain knowledge towards two components in RAG: text chunking and prompt augmenting. We investigate the impact of modifying these two components to better suit domain knowledge and evaluate their effectiveness.

Hierarchy-aware Chunking. We propose a chunking strategy that preserves components in legislation as a hierarchical data structure via extensive regular expression and custom rule-based. We select only section-level nodes for experiments, as suggested in Appendix E. We compared our proposed Hierarchy-aware Chunking with a naive chunking strategy (see Appendix F on how we obtain naive chunking setups).

Since the naive chunking strategy has no awareness of section boundaries, the chunked text might either contain multiple sections (if the section is shorter than the chunk size) or be incomplete (if the section is longer than the chunk size). This makes it hard to justify whether a retrieved incomplete chunk (partially containing section content) is considered a correctly retrieved document. Therefore, to simply retrieve and enable a fair comparison of top-k retrieval across strategies, **chunks that do not fully cover at least one section are discarded**. We also remove sections from the ierarchy-aware chunks that are not covered by the naive chunking strategy.

After filtering out sections that are not contained in the naive chunks, only 19 ThaiLegal-Tax entries and 2,625 ThaiLegal-CCL entries were left. Given the limited size of the ThaiLegal-Tax subset, we perform evaluations solely on ThaiLegal-CCL .

For this setup, we use a three-headed, Human-Finetuned BGE-M3 as a retriever (see § 4.2.1) and `gpt-4o` as the LLM.

LegalRef. To handle inter-section references, we introduce **LegalRef**, a framework that recursively fetches referenced sections and incorporates them into the LLM context. We adopt a depth-first referencing strategy where the referenced section will be placed next to the referencing section. For example, if Section A references Section B, LegalRef retrieves Section B and places it at the next rank after Section A. We evaluate its impact on retrieval and E2E performance using ierarchy-aware Chunking, Human-Finetuned BGE-M3 (see §4.2.1), and GPT-4o. We compare the performance of the RAG with and without LegalRef component using our proposed benchmark. We use a maximum reference depth of 1 due to a significant inference budget required since more reference depth increases prompt length dramatically.

4.2 (RQ2) Impact of Retriever and LLM

This research question aims to investigate the performance of two main components in the RAG system: Retrieval model and LLM. For each component, we conduct an experiment to compare the performance of the baseline ("naive RAG"), our "proposed RAG framework", and RAG with golden context which acts as an upper bound performance.

4.2.1 Retriever Models

Conventionally, BGE-M3 (Chen et al., 2024) was a popular choice for text embeddings due to its superior performance across languages and models. However, in some cases, BGE-M3 was also finetuned towards domain-specific data to improve the performance. Therefore, for this experiment, using our benchmark, we evaluate the effectiveness of the following four retrievers: [4]: (1) **BM25** (Robertson and Zaragoza, 2009): This serves as our baseline for the retrieval model performance. (2) **BGE-M3** (Chen et al., 2024): A retrieval model that shows a strong performance in many languages and domains. (3) **Human-Finetuned BGE-M3 (HF BGE-M3)**: A BGE-M3 model finetuned on ThaiLegal-CCL dataset. (4) **Auto-Finetuned BGE-M3 (AF BGE-M3)**: A finetuned BGE-M3 model on augmented ThaiLegal-CCL where we use `bge-reranker-v2-m3`[5] to rerank documents instead of legal experts.

The goal is to quantify the effectiveness between using a default BGE-M3, finetuned BGE-M3 on human-curated data, and finetuned BGE-M3 using an automatic reranking model. For all BGE-M3 variants, we use all three heads, and we weigh dense, multi-vector, and sparse scores at 0.4, 0.4, and 0.2, respectively.

4.2.2 LLM Choices

Once we identified the best retriever from the previous experiment, we fixed the retriever as HF BGE-M3 and evaluated the following LLMs: (1) GPT-4o[6] (Hurst et al., 2024), (2) Claude 3.5 Sonnet[7] (Anthropic, 2024b), (3) Gemini 1.5 Pro[8] (Reid et al., 2024), (4) Typhoon V2 70b (Pipatanakul et al., 2024) Our goal is to identify the performance of each LLM and select what LLM will be used for E2E evaluation (§ 4.2.3).

All LLMs use 3-shot examples randomly sampled from the training data, a temperature of 0.5, and a max output token limit of 2048.

4.2.3 E2E Evaluations

Building upon previous observations from §4.1 and §4.2, we defined our best setups for a RAG framework and compared each approach using ThaiLegal.

[4]We also conduct these experiments on more retrieval models. The results are outlined in Appendix G

[5]https://huggingface.co/BAAI/bge-reranker-v2-m3

[6]gpt-4o-2024-08-06

[7]claude-3-5-sonnet-20240620

[8]gemini-1.5-pro-002

Specifically, we compare four systems: (1) Parametric Knowledge: LLM-only baseline, (2) Naive RAG: Traditional RAG with naive chunking, (3) Proposed RAG: Enhanced with Hierarchy-aware Chunking and LegalRef, (4) RAG with Golden Context: Upper bound with ground truth context. For "Naive RAG," "Proposed RAG," and "Golden Context," we use Human-finetuned BGE-M3 as the retriever and Claude 3.5 Sonnet as the LLM. Unlike the Hierarchy-aware Chunking Experiment, the benchmark datasets for Naive RAG and Proposed RAG are not filtered to include only queries with relevant laws available in naive chunks. Additionally, in the Proposed RAG system, chunks are used as-is, without discarding those that contain sections absent from the naive chunks.

4.3 (RQ3) Performance of Long-Context LLMs

LCLMs like Gemini 1.5 Pro, which has a context window of over 2M tokens, can ingest all legislation in L into their prompt, potentially replacing the need for a retrieval model. We aim to explore Gemini's capabilities in Thai legal QA, where we use all legislation as a context for the prompt. We evaluate LCLM in two settings: (1) LCLM as Generator: Gemini 1.5 Pro processes all laws as context, answering queries directly without any retrieval model, (2) LCLM as Retriever: Gemini 1.5 Pro retrieves top-k relevant documents, replacing traditional retrievers. We want to explore if Gemini 1.5 Pro can retrieve better documents under complex reasoning setups. Due to budget constraints, experiments are conducted on a 20% stratified subset of ThaiLegal-CCL and the full ThaiLegal-Tax dataset.

5 Results and Discussion

5.1 (RQ1) Impact of Tailored Components

5.1.1 Hierarchy-aware Chunking

Hierarchy-aware chunking achieves a slight but consistent advantage over the naive chunking strategy. From Table 1, the naive chunking strategy performs worse than hierarchy-aware chunking in terms of retrieval performance. This discrepancy likely arises because naive chunks often contain content from multiple sections, introducing "noise" that can negatively impact the retrieval model's ranking of relevant documents.

However, in terms of end-to-end (E2E) performance, the system using Hierarchy-aware chunking only slightly outperforms the one using naive chunking. We suspect that this is because the LLM can effectively filter out the "noise" in the retrieved sections during answer generation. As a result, the coverage and contradiction scores are not significantly different between the two systems. Nevertheless, there remains a discrepancy in the E2E citation score.

Setting	Retriever Multi MRR (↑)	Retriever Recall (↑)	Coverage (↑)	Contradiction (↓)	E2E Recall (↑)	E2E Precision (↑)	E2E F1 (↑)
Naïve Chunking	0.786	0.935	86.6	**0.050**	0.882	0.613	0.722
Hierarchy-aware Chunking	**0.834**	**0.942**	**86.7**	0.054	**0.894**	**0.630**	**0.739**

Table 1: Effect of Chunking Configuration on E2E Performance on the ThaiLegal-CCL dataset.

LegalRef. The results from Table 2 show that there is no clear significant advantage when employing LegalRefin a RAG system.

Metric	ThaiLegal-CCL		ThaiLegal-Tax	
	Ref Depth 1	No Ref	Ref Depth 1	No Ref
Retriever Metrics				
Multi MRR (↑)	0.809	0.809	0.333	0.333
Recall (↑)	0.938	0.938	0.499	0.499
Referencer Metrics				
Multi MRR (↑)	0.800	**0.809**	**0.345**	0.333
Recall (↑)	**0.940**	0.938	**0.602**	0.499
Coverage (↑)	**86.3**	85.2	45.0	**50.0**
Contradiction (↓)	**0.051**	0.055	0.520	**0.460**
E2E Recall (↑)	**0.885**	0.880	**0.354**	0.333
E2E Precision (↑)	0.579	**0.601**	0.630	**0.64**
E2E F1 (↑)	0.700	**0.714**	**0.453**	0.438

Table 2: Effect of augmenter configuration on E2E performance, with separate grouping for Retriever and Referencer metrics.

In a complex legal query, LegalRef improves retriever recall, but the additional correct sections are usually ranked at the bottom. According to the result, we can clearly see that the recall was improved by 10%, yet MRR and MultiMRR were only marginally improved. This suggested that LegalRefdoes provide additional correct sections to the retrieved documents while the document that cited more positives by LegalRefis still ranked at the bottom of the retrieved documents.

LLM	Referencer	Retriever Recall (↑)	E2E Recall (↑)	E2E Precision (↑)	E2E F1 (↑)	Coverage (↑)	Contradiction
ThaiLegal-CCL Dataset							
gpt-4o-2024-08-06	Ref Depth 1	0.938	0.885	0.579	0.700	86.3	0.051
	No Ref		0.880	**0.601**	**0.714**	85.2	0.055
gemini-1.5-pro-002	Ref Depth 1	0.938	0.895	0.491	0.634	87.3	0.042
	No Ref		0.892	0.512	0.651	86.5	0.048
claude-3-5-sonnet-20240620	Ref Depth 1	0.938	0.894	0.443	0.592	89.5	0.044
	No Ref		**0.901**	0.444	0.595	**89.7**	**0.040**
typhoon-v2-70b-instruct	Ref Depth 1	0.938	0.845	0.573	0.683	79.9	0.080
	No Ref		0.862	0.537	0.662	81.2	0.076
ThaiLegal-Tax Dataset							
gpt-4o-2024-08-06	Ref Depth 1	0.499	0.354	0.630	0.453	45.0	0.52
	No Ref		0.333	0.64	0.438	50.0	0.46
gemini-1.5-pro-002	Ref Depth 1	0.499	0.354	0.347	0.351	45.0	0.48
	No Ref		0.361	0.308	0.332	44.0	0.48
claude-3-5-sonnet-20240620	Ref Depth 1	0.499	**0.417**	0.577	**0.484**	49.0	0.56
	No Ref		0.389	0.554	0.457	51.0	**0.44**
typhoon-v2-70b-instruct	Ref Depth 1	0.499	0.333	0.453	0.384	**54.0**	0.46
	No Ref		0.326	**0.662**	0.437	42.0	0.58

Table 3: Effect of LLM configuration on end-to-end performance on ThaiLegal-CCL and ThaiLegal-Tax Datasets. For Retriver Recall, we show only the recall without taking into account of the referenced section for Ref Depth 1.

Improvement in retriever recall from LegalRef doesn't always translate to improvement in generation performance. In the ThaiLegal-Tax dataset, despite recall having a substantial improvement, E2E metrics declined. We hypothesized that the complexity of the ThaiLegal-Tax dataset demands advanced reasoning capabilities that the LLM, even with the correct documents, struggles to provide. Another potential reason that might affect the performance decline is the longer context that the LLM needs to process due to the higher amount of content added by LegalRef. We also further conduct more analysis on increasing reference depth in Appendix H.

5.2 (RQ2) Impact of Retriever and LLM

5.2.1 Retriever Models

Table 4 showed the performance of different retrieval models on both ThaiLegal-CCL and ThaiLegal-Tax . HF BGE-M3 achieved the best performance in ThaiLegal-CCL , as expected, since this is considered an "in-domain" data for the retriever. However, surprisingly, AF BGE-M3 achieves a very close performance compared to HF BGE-M3 (< 1%). This suggested that **for a simple legal query like ThaiLegal-CCL , `bge-reranker-v2-m3` is suitable to approximate the legal experts for annotating retrieval data.**

The ThaiLegal-Tax dataset, on the other hand, showed mixed results. HF BGE-M3 achieves the highest Hit rate but only marginally compared to the base BGE-M3. Interestingly, the base BGE-M3 model achieves a higher Multi MRR compared to both HF and AF BGE-M3. From the result, we can interpret that **finetuning a retrieval model on a simple case, despite improved retrieval performance on generic legal QA, still can't generalize towards a complex legal reasoning query.** Additionally, based on the following results, we opted to use **HF BGE-M3** as a retriever for E2E experiments due to their superior performance in both datasets.

ThaiLegal-CCL

Model	HR/Recall	MRR
BM25	.658	.548
BGE-M3	.880	.773
HF BGE-M3	**.906**	**.805**
AF BGE-M3	.900	.800

ThaiLegal-Tax

Model	HR	Multi HR	Recall	MRR	Multi MRR
BM25	.480	.120	.254	.318	.171
BGE-M3	.720	**.240**	**.435**	.580	**.337**
HF BGE-M3	**.740**	.220	.411	.565	.320
AF BGE-M3	.700	.200	.382	**.587**	.329

Table 4: Retrieval Evaluation Results for BM25 and BGE-M3 Variants (Top-K = 5).

5.2.2 LLM Choices

The benchmark results of varying LLM are shown in Table 3. We also added the configuration of including and not including LegalRefin this experiment as well since the result in §5.1.1 showed no clear conclusion.

Claude 3.5 Sonnet performs best generally for Thai Legal QA. For ThaiLegal-CCL , Claude 3.5 Sonnet outperforms other proprietary LLMs for E2E recall and coverage. Nevertheless, `gpt-4o-2024-08-06`, despite having a lower coverage score, yields a surprisingly high E2E F1 score, highlighting a dominant performance in selecting the relevant section to be cited in the generated answer.

Effective of incorporating LegalRef is still inconclusive. On the ThaiLegal-Tax dataset, most models struggle to reason over the relevant documents based on the performance difference compared to the ThaiLegal-CCL dataset. Claude 3.5 Sonnet clearly outperforms `gpt-4o-2024-08-06` and `gemini-1.5-pro-002` in most E2E metrics. However, `typhoon-v2-70b-instruct`, an open-sourced model, unexpectedly became the only model that incorporated LegalRefand obtained an improved Coverage and Contradiction score.

5.2.3 E2E Evaluations

Given the previous experiments, we have verified the effectiveness of using HF BGE-M3 as a retriever and Claude 3.5 Sonnet as an LLM for RAG. Since the results for incorporating LegalRef were inconclusive, we removed the use of LegalRef for this experiment since it significantly reduced prompt length. We presented the results of a full RAG pipeline in Table 5.

From the results, we use Claude 3.5 Sonnet as the main LLM for the E2E experiment since it yields the most consistent performance across all metrics. Additionally, the proposed RAG with Hierarchy-aware chunking provides the best coverage and contradiction score for both ThaiLegal-CCL and ThaiLegal-Tax . On the other hand, all setups, including golden context, which is the upper bound, still struggle on ThaiLegal-Tax . This indicates that *utilizing RAG alone is insufficient to solve sophisticated legal QA queries, especially when legal reasoning is required.*

We also see a surprising pattern in the parametric knowledge setup where Claude 3.5 Sonnet yields an astonishingly high E2E F1 score. To further investigate this, we inspect the cited section that was generated by LLM. Surprisingly, out of 105 sections cited from LLM parametric knowledge, 58 of them *were not* even retrieved by the best retriever.

Setting	Coverage (↑)	Contradiction (↓)	E2E Recall (↑)	E2E Precision (↑)	E2E F1 (↑)
ThaiLegal-CCL Dataset					
Parametric	60.3	0.199	0.188	0.141	0.161
Naïve RAG	77.3	0.097	0.745	0.370	0.495
Proposed RAG	**89.7**	**0.040**	**0.901**	**0.444**	**0.595**
Golden Context	93.4	0.034	0.999	1.000	1.000
ThaiLegal-Tax Dataset					
Parametric	46.0	0.480	**0.458**	**0.629**	**0.530**
Naïve RAG	50.0	0.460	0.306	0.463	0.368
Proposed RAG	**51.0**	**0.440**	0.389	0.554	0.457
Golden Context	52.0	0.460	0.694	1.000	0.820

Table 5: E2E evaluation results on ThaiLegal-CCL and ThaiLegal-Tax. **Parametric** represents naive few-shot prompts without additional context. **Naive RAG** is a conventional RAG with naive chunking strategy. **Proposed RAG** utilized a hierarchy-aware chunking strategy. **Golden Context** remove retrieval component in RAG, augmented the prompt with groud truth positives.

Among those 58 cited documents, 26 of those were correct. In contrast, only 5 of 101 sections cited by the proposed RAG system are *not* retrieved. This indicates that retriever performance significantly constrains RAG systems, especially with complex queries like those in ThaiLegal-Tax .

5.3 (RQ3) LCLM Performance

LCLM still underperforms RAG on Thai Legal QA both in simple and complex datasets. In Table 6, we can see that LCLM performance for both coverage and contradiction is still below our proposed RAG. This performance gap may stem

Setting	Coverage (↑)	Contradiction (↓)	E2E Recall (↑)	E2E Precision (↑)	E2E F1 (↑)
ThaiLegal-CCL Dataset					
Parametric	60.6	0.198	0.197	0.147	0.169
Naïve RAG	77.7	0.092	0.740	0.379	0.501
Proposed RAG	**90.1**	**0.028**	**0.920**	0.453	0.607
LCLM	83.2	0.063	0.765	**0.514**	**0.615**
Golden Context	94.2	0.025	0.999	1.0	0.999
ThaiLegal-Tax Dataset					
Parametric	46.0	0.480	**0.458**	**0.629**	**0.530**
Naïve RAG	50.0	0.460	0.306	0.463	0.368
Proposed RAG	**51.0**	**0.440**	0.389	0.554	0.457
LCLM	36.0	0.620	0.410	0.484	0.444
Golden Context	52.0	0.460	0.694	1.000	0.820

Table 6: E2E results including LCLM on a 20% stratified subset of the test data on ThaiLegal-CCL dataset and full ThaiLegal-Tax dataset. We use `gemini-1.5-pro-002` for LCLM.

from degradation when processing extremely long contexts (1.2 million tokens). The results suggest that while an LCLM-based Thai legal QA system is feasible, its performance remains significantly behind RAG-based counterparts, highlighting areas for further improvement.

LCLM-as-a-retriever was feasible technically but still unfeasible economically. Table 7 showed the performance of LCLM-as-a-retriever. On a simple query dataset, ThaiLegal-CCL , the performance is still subpar to that of BGE-M3 and its variants. We suspect this might be due to too much distractor in a longer context document, resulting in a lower performance. However, on a complex retrieval dataset, ThaiLegal-Tax , LCLM-as-a-retriever outperforms all retrieval models in all metrics. This indicates the feasibility of using LCLM-as-a-retriever. Nevertheless, performance compared to the cost and latency introduced makes this approach worse tradeoffs than using a conventional embedding model.

ThaiLegal-CCL Dataset

Model	HR/Recall	MRR
BM25	.663	.549
BGE-M3	.888	.779
HF BGE-M3	**.909**	**.819**
AF BGE-M3	.909	.807
LCLM	.776	.667

ThaiLegal-Tax Dataset

Model	HR	Multi HR	Recall	MRR	Multi MRR
BM25	.480	.120	.254	.318	.171
BGE-M3	.720	.240	.435	.580	.337
HF BGE-M3	.740	.220	.411	.565	.320
AF BGE-M3	.700	.200	.382	**.587**	.329
LCLM	**.760**	**.320**	**.515**	**.587**	**.370**

Table 7: Retrieval Evaluation Results (Top-K = 5) for BM25, BGE-M3 variants, and LCLM-as-a-retriever on the ThaiLegal-CCL and ThaiLegal-Tax datasets. We conduct this experiment on a 20% stratified subset of test set due to budget constraint.

We further discuss the effect of the relevant section position in the context of the E2E performance in Appendix I.

5.4 Effectiveness of Multi-label Metrics

To further validate the effectiveness of our proposed multi-label metrics, we compute the correlation between conventional retrieval metrics (Hit Rate and MRR) compared to its multi-label variant. We use eight retriever model performances (see Appendix G) to measure the correlation between retrieval and the E2E metric. The result was presented in Table 8.

According to the result, we can see that our Multi-MRR and Multi-Hit Rate have a higher correlation compared to conventional MRR and hit rate. These results emphasize the importance of using multi-label metrics in legal QA setups.

	Coverage (↑)	Contradiction (↓)	E2E F1 (↑)
Hit Rate	0.741	-0.672	0.780
Multi Hit Rate	**0.989**	**-0.986**	0.984
MRR	0.906	-0.859	0.933
Multi MRR	**0.989**	-0.973	**0.991**

Table 8: Correlation between conventional and multi-retrieval metrics with evaluation measures using data from 8 retrievers (Appendix G)

Conclusion

This work introduces ThaiLegal, a comprehensive benchmark for evaluating Thai legal QA systems, addressing the critical challenges posed by limited-resource languages and the complexity of legal reasoning. Through two novel datasets and tailored evaluation metrics, the study offers a robust framework for assessing both retrieval-augmented generation (RAG) systems and long-context language models (LCLMs). While domain-specific enhancements, such as Hierarchy-aware Chunking and LegalRef, yielded incremental improvements, the persistent challenges in retrieval accuracy and consistent legal reasoning underscore the limitations of existing models. Notably, LCLMs demonstrated potential but struggled with performance degradation in processing extensive contexts. These findings highlight the need for continued innovation in retrieval strategies, model fine-tuning, and domain-specific methodologies.

By establishing ThaiLegal, this study provides a foundational step toward systematic progress in Thai legal NLP, fostering the development of more reliable, efficient, and legally grounded QA systems. Future research can build on these insights to address the nuanced demands of legal reasoning, explore hybrid approaches, and expand the scope of legal AI applications in underrepresented languages like Thai.

Limitations

Despite being the first E2E benchmark for Thai legal QA, both of our datasets still have several limitations.

XYZ-ThaiLegal-CCL and ThaiLegal-CCL Limitations. The XYZ-ThaiLegal-CCL training split was constructed in a semi-synthetic approach with human quality control for the training set and a fully human-annotated process for the test set (ThaiLegal-CCL). While this design effectively manages costs, it presents several issues.

First, let us discuss the ambiguity of queries in the test set caused by single-section sampling. Annotators create questions based solely on a single sampled section from one of the 21 available laws, often leading to queries that are too general and overlap with multiple related sections. This lack of specificity can confuse language models, which incorporate multiple sections even when the query targets just one. This also applies to training data where the answer was first generated by LLM, given only one law section to the prompt.

Second, the absence of truly multi-label queries in both the training and test sets. While annotators in the training set select multiple relevant sections from retrieved documents, the questions themselves originate from single sections, restricting their multi-label nature. This limits the dataset's ability to evaluate reasoning across multiple legal provisions. Although ThaiLegal-Tax partially addresses this gap by including queries requiring multi-label reasoning, this issue persists across the broader dataset.

Finally, the dataset's queries lack natural phrasing and fail to reflect how real users would pose questions in a Thai legal QA system. Current queries are often overly formal or influenced by the dataset construction process, making them less representative of typical user input.

These challenges—ambiguity in queries, the absence of multi-label scenarios, and unnatural phrasing—highlight areas for improvement to enhance both XYZ-ThaiLegal-CCL and ThaiLegal-CCL dataset's relevance and effectiveness for Thai legal QA systems.

Reliability of Multi-label Metrics. Our proposed Multi-HitRate and Multi-MRR, although shown in §5.4 to correlate more strongly with the E2E metrics, were calculated using only eight retrievers. This limited data point is primarily due to the substantial cost associated with inferencing a larger pool of retrievers, coupled with the scarcity of available retriever models specifically tailored for the Thai legal domain. Consequently, while our initial findings are promising, the restricted number of retrievers may impact the generalizability of these metrics. Future work should explore expanding the set of retrievers and consider additional domain-specific datasets to further validate and potentially refine the robustness of our multi-label evaluation framework.

Legal Reasoning Evaluation. Beyond Coverage, Contradiction, and Citation scores, legal reasoning is crucial for Legal QA. It differs from general reasoning by operating within a structured legal framework, demanding strict adherence to legal principles and precise interpretation of authoritative sources. Evaluating legal reasoning, where the process matters as much as the answer, enhances the performance assessment. This work, although highlighting how to evaluate the final answer, still lacks the measurement of LLM legal reasoning and focuses specifically on the final generated response. Existing studies explore reasoning evaluation in LLMs using metrics for semantic alignment, logical inference, and language coherence (Golovneva et al., 2023) and qualities like correctness and informativeness (Prasad et al., 2023). LLM Reasoner (Hao et al., 2024) automate error categorization using LLMs. However, reasoning evaluation for LLMs, especially in the Thai legal domain, remains challenging. Obstacles include defining "good" legal reasoning and acquiring datasets that require complex legal reasoning beyond simple lookups.

References

ailawyer. AI Lawyer | Your personal AI legal assistant. ailawyer.pro. [Accessed 14-01-2025].

Chenxin An, Shansan Gong, Ming Zhong, Xingjian Zhao, Mukai Li, Jun Zhang, Lingpeng Kong, and Xipeng Qiu. 2023. L-eval: Instituting standardized evaluation for long context language models. *Preprint*, arXiv:2307.11088.

Anthropic. 2024a. The claude 3 model family: Opus, sonnet, haiku. Technical report.

Anthropic. 2024b. Claude 3.5 sonnet model card addendum.

asklegal.bot. AskLegal.bot - AI Legal Help. asklegal.bot. [Accessed 14-01-2025].

Yushi Bai, Xin Lv, Jiajie Zhang, Hongchang Lyu, Jiankai Tang, Zhidian Huang, Zhengxiao Du, Xiao Liu, Aohan Zeng, Lei Hou, Yuxiao Dong, Jie Tang, and Juanzi Li. 2024. LongBench: A bilingual, multitask benchmark for long context understanding. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 3119–3137, Bangkok, Thailand. Association for Computational Linguistics.

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)*, pages 4310–4330, Dublin, Ireland. Association for Computational Linguistics.

Jianlv Chen, Shitao Xiao, Peitian Zhang, Kun Luo, Defu Lian, and Zheng Liu. 2024. Bge m3-embedding: Multi-lingual, multi-functionality, multi-granularity text embeddings through self-knowledge distillation. *Preprint*, arXiv:2402.03216.

Somyot Chuathai. 2023. *Introduction to Law*, 30th edition. Winyuchon.

Matthew Dahl, Varun Magesh, Mirac Suzgun, and Daniel E Ho. 2024. Large legal fictions: Profiling legal hallucinations in large language models. *Journal of Legal Analysis*, 16(1):64–93.

Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Amy Yang, and Angela Fan et al. 2024. The llama 3 herd of models. *Preprint*, arXiv:2407.21783.

Shahul Es, Jithin James, Luis Espinosa-Anke, and Steven Schockaert. 2023. Ragas: Automated evaluation of retrieval augmented generation. *Preprint*, arXiv:2309.15217.

Olga Golovneva, Moya Chen, Spencer Poff, Martin Corredor, Luke Zettlemoyer, Maryam Fazel-Zarandi, and Asli Celikyilmaz. 2023. Roscoe: A suite of metrics for scoring step-by-step reasoning. *Preprint*, arXiv:2212.07919.

Neel Guha, Julian Nyarko, Daniel E. Ho, Christopher Ré, Adam Chilton, Aditya Narayana, Alex Chohlas-Wood, Austin Peters, Brandon Waldon, Daniel N. Rockmore, Diego Zambrano, Dmitry Talisman, Enam Hoque, Faiz Surani, Frank Fagan, Galit Sarfaty, Gregory M. Dickinson, Haggai Porat, Jason Hegland, and 21 others. 2023. Legalbench: A collaboratively built benchmark for measuring legal reasoning in large language models. *Preprint*, arXiv:2308.11462.

Shibo Hao, Yi Gu, Haotian Luo, Tianyang Liu, Xiyan Shao, Xinyuan Wang, Shuhua Xie, Haodi Ma, Adithya Samavedhi, Qiyue Gao, Zhen Wang, and Zhiting Hu. 2024. Llm reasoners: New evaluation, library, and analysis of step-by-step reasoning with large language models. *Preprint*, arXiv:2404.05221.

Aaron Hurst, Adam Lerer, Adam P Goucher, Adam Perelman, Aditya Ramesh, Aidan Clark, AJ Ostrow, Akila Welihinda, Alan Hayes, Alec Radford, and et al. 2024. Gpt-4o system card. *Preprint*, arXiv:2410.21276.

Rohan Jha, Bo Wang, Michael Günther, Saba Sturua, Mohammad Kalim Akram, and Han Xiao. 2024. Jina-colbert-v2: A general-purpose multilingual late interaction retriever. *Preprint*, arXiv:2408.16672.

Gregory Kamradt. 2023. Needle in a haystack. github.com.

Omar Khattab and Matei Zaharia. 2020. Colbert: Efficient and effective passage search via contextualized late interaction over bert. In *Proceedings of the 43rd International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '20, page 39–48, New York, NY, USA. Association for Computing Machinery.

Philippe Laban, Alexander R. Fabbri, Caiming Xiong, and Chien-Sheng Wu. 2024. Summary of a haystack: A challenge to long-context llms and rag systems. *Preprint*, arXiv:2407.01370.

Chankyu Lee, Rajarshi Roy, Mengyao Xu, Jonathan Raiman, Mohammad Shoeybi, Bryan Catanzaro, and Wei Ping. 2024a. Nv-embed: Improved techniques for training llms as generalist embedding models. *Preprint*, arXiv:2405.17428.

Jinhyuk Lee, Anthony Chen, Zhuyun Dai, Dheeru Dua, Devendra Singh Sachan, Michael Boratko, Yi Luan, Sébastien M. R. Arnold, Vincent Perot, Siddharth Dalmia, Hexiang Hu, Xudong Lin, Panupong Pasupat, Aida Amini, Jeremy R. Cole, Sebastian Riedel, Iftekhar Naim, Ming-Wei Chang, and Kelvin Guu. 2024b. Can long-context language models subsume retrieval, rag, sql, and more? *Preprint*, arXiv:2406.13121.

Patrick Lewis, Ethan Perez, Aleksandra Piktus, Fabio Petroni, Vladimir Karpukhin, Naman Goyal, Heinrich Küttler, Mike Lewis, Wen tau Yih, Tim Rocktäschel, Sebastian Riedel, and Douwe Kiela. 2021. Retrieval-augmented generation for knowledge-intensive nlp tasks. *Preprint*, arXiv:2005.11401.

LexisNexis. 2023. LexisNexis Launches Lexis+ AI, a Generative AI Solution with Hallucination-Free Linked Legal Citations. lexisnexis.com. [Accessed 13-08-2024].

Zhuowan Li, Cheng Li, Mingyang Zhang, Qiaozhu Mei, and Michael Bendersky. 2024. Retrieval augmented generation or long-context llms? a comprehensive study and hybrid approach. *Preprint*, arXiv:2407.16833.

Varun Magesh, Faiz Surani, Matthew Dahl, Mirac Suzgun, Christopher D. Manning, and Daniel E. Ho. 2024. Hallucination-free? assessing the reliability of leading ai legal research tools. *Preprint*, arXiv:2405.20362.

Hung Phan, Anurag Acharya, Sarthak Chaturvedi, Shivam Sharma, Mike Parker, Dan Nally, Ali Jannesari, Karl Pazdernik, Mahantesh Halappanavar, Sai Munikoti, and Sameera Horawalavithana. 2024. Rag vs. long context: Examining frontier large language models for environmental review document comprehension. *Preprint*, arXiv:2407.07321.

Kunat Pipatanakul, Potsawee Manakul, Natapong Nitarach, Warit Sirichotedumrong, Surapon Nonesung, Teetouch Jaknamon, Parinthapat Pengpun, Pittawat Taveekitworachai, Adisai Na-Thalang, Sittipong Sripaisarnmongkol, Krisanapong Jirayoot, and Kasima Tharnpipitchai. 2024. Typhoon 2: A family of open text and multimodal thai large language models. *Preprint*, arXiv:2412.13702.

Archiki Prasad, Swarnadeep Saha, Xiang Zhou, and Mohit Bansal. 2023. Receval: Evaluating reasoning chains via correctness and informativeness. *Preprint*, arXiv:2304.10703.

Machel Reid, Nikolay Savinov, Denis Teplyashin, Dmitry Lepikhin, Timothy Lillicrap, Jean baptiste Alayrac, Radu Soricut, Angeliki Lazaridou, Orhan Firat, Julian Schrittwieser, and et al. 2024. Gemini 1.5: Unlocking multimodal understanding across millions of tokens of context. *Preprint*, arXiv:2403.05530.

Stephen Robertson and Hugo Zaragoza. 2009. The probabilistic relevance framework: Bm25 and beyond. *Foundations and Trends in Information Retrieval*, 3:333–389.

Sarah Strumberger. 2023. AI-powered legal research: Where legal research meets generative AI. legal.thomsonreuters.com. [Accessed 13-08-2024].

Saba Sturua, Isabelle Mohr, Mohammad Kalim Akram, Michael Günther, Bo Wang, Markus Krimmel, Feng Wang, Georgios Mastrapas, Andreas Koukounas, Andreas Koukounas, Nan Wang, and Han Xiao. 2024. jina-embeddings-v3: Multilingual embeddings with task lora. *Preprint*, arXiv:2409.10173.

Akash Takyar. AI agents for legal: Applications, benefits, implementation and future trends — leewayhertz.com. leewayhertz.com. [Accessed 13-08-2024].

The Kingdom of Thailand. 2022. Section 260 of the Criminal Code B.E. 2565. Office of the Council of State of Thailand. Author's translation.

Kobkrit Viriyayudhakorn. 2024. Thanoy AI Chatbot - genius AI lawyer. iapp.co.th. [Accessed 13-08-2024].

Nirmalie Wiratunga, Ramitha Abeyratne, Lasal Jayawardena, Kyle Martin, Stewart Massie, Ikechukwu Nkisi-Orji, Ruvan Weerasinghe, Anne Liret, and Bruno Fleisch. 2024. Cbr-rag: Case-based reasoning for retrieval augmented generation in llms for legal question answering. *Preprint*, arXiv:2404.04302.

Jingtao Zhan, Jiaxin Mao, Yiqun Liu, Min Zhang, and Shaoping Ma. 2020. Repbert: Contextualized text embeddings for first-stage retrieval. *CoRR*, abs/2006.15498.

A XYZ-ThaiLegal-CCL Dataset Curation

A.1 Curating Training Data

This section outline the data collection process of XYZ-ThaiLegal-CCLdataset. Consider dataset notations from §3.1.1. Questions q_i are generated using Gemini 1.5 Pro (Reid et al., 2024) based on the given section sampled from L. Then, we retrieve relevant candidate sections p_k for each question using BGE-M3 (Chen et al., 2024) resulting in positive documents T_i. The label y was generated using Llama-3-70B (Dubey et al., 2024) (or Claude 3 Sonnet (Anthropic, 2024a) if Llama-3-70B reject the answer). Finally, the generated answer y and positive sections T are further validated by legal experts for assuring data quality. The legal experts either remove irrelevant section, add more relevant sections, or rerank sections in T and adjust y to ensure phrases are all correct. Thus, for our training data, queries q correspond to T_i where $|T_i| \geq 1$ and are considered multi-label. The legislation list for XYZ-ThaiLegal-CCLdataset curation is in Table 9. Figure 1 shows the data collection process for XYZ-ThaiLegal-CCL's training split.

A.2 Curating Test Data

For the test dataset, all queries q_i and generated answer y_i were manually crafted by legal experts given a single section sampled from L. Each manually crafted question was carefully quality-assured by a second legal expert. As a result, the test data are single-labeled ($|T_i| = 1$), whereas the training data are multi-labeled.

A.3 Annotator Profile and Cost

Since we are curating a dataset specifically in the Thai legal domain, it is important to ensure that our annotators have a strong background in Thai legal knowledge. To achieve this, we recruited legal experts through law school professors via their available channels, such as their social networks . We received a total of 97 applications and selected 34 annotators. Their occupations include law students, recent law school graduates, and employees at law firms. Furthermore, all annotators were informed that the data would be used for an open-source research project, and their participation implied consent to this usage.

We compensate annotators per completed task, which includes curating the training set, conducting quality checks, and curating the test set. Tasks are randomly assigned, and we adjust the distribu-

Legislation	Legal Terminology	Training	Test
Organic Act on Counter Corruption, B.E. 2561	organic law	✓	
Civil and Commercial Code	code	✓	✓
Revenue Code	code	✓	✓
Accounting Act, B.E. 2543	act	✓	✓
Accounting Profession Act, B.E. 2547	act	✓	✓
Act on Disciplinary Offenses of Government Officials Performing Duties in Agencies Other than Government Agencies, B.E. 2534	act	✓	
Act on Offences of Officials Working in State Agencies or Organizations, B.E. 2502	act	✓	
Act on Offenses Relating to Registered Partnerships, Limited Partnerships, Companies Limited, Associations and Foundations, B.E. 2499	act	✓	✓
Act on the Establishment of Government Organizations, B.E. 2496	act	✓	
Act on the Management of Shares and Stocks of Ministers, B.E. 2543	act	✓	
Act Repealing the Agricultural Futures Trading Act, B.E. 2542 B.E. 2558	act	✓	
Budget Procedure Act, B.E. 2561	act	✓	
Business Registration Act, B.E. 2499	act	✓	✓
Chamber of Commerce Act, B.E. 2509	act	✓	✓
Derivatives Act, B.E. 2546	act	✓	✓
Energy Conservation Promotion Act, B.E. 2535	act	✓	✓
Energy Industry Act, B.E. 2550	act	✓	✓
Financial Institutions Business Act, B.E. 2551	act	✓	✓
Fiscal Discipline Act, B.E. 2561	act	✓	
Foreign Business Act, B.E. 2542	act	✓	✓
Government Procurement and Supplies Management Act, B.E. 2560	act	✓	
National Economic and Social Development Act, B.E. 2561	act	✓	
Petroleum Income Tax Act, B.E. 2514	act	✓	✓
Provident Fund Act, B.E. 2530	act	✓	✓
Public Limited Companies Act, B.E. 2535	act	✓	✓
Secured Transactions Act, B.E. 2558	act	✓	✓
Securities and Exchange Act, B.E. 2535	act	✓	✓
State Enterprise Capital Act, B.E. 2542	act	✓	
State Enterprise Committee and Personnel Qualifications Standards Act, B.E. 2518	act	✓	
State Enterprise Development and Governance Act, B.E. 2562	act	✓	
State Enterprise Labor Relations Act, B.E. 2543	act	✓	
Trade Association Act, B.E. 2509	act	✓	✓
Trust for Transactions in Capital Market Act, B.E. 2550	act	✓	✓
Emergency Decree on Digital Asset Businesses, B.E. 2561	emergency decree	✓	
Emergency Decree on Special Purpose Juristic Person for Securitization, B.E. 2540	emergency decree	✓	✓

Table 9: ThaiLegal-CCL Legislation (High to Low Legislative Rank, Alphabetical): Training and Test Set Distribution

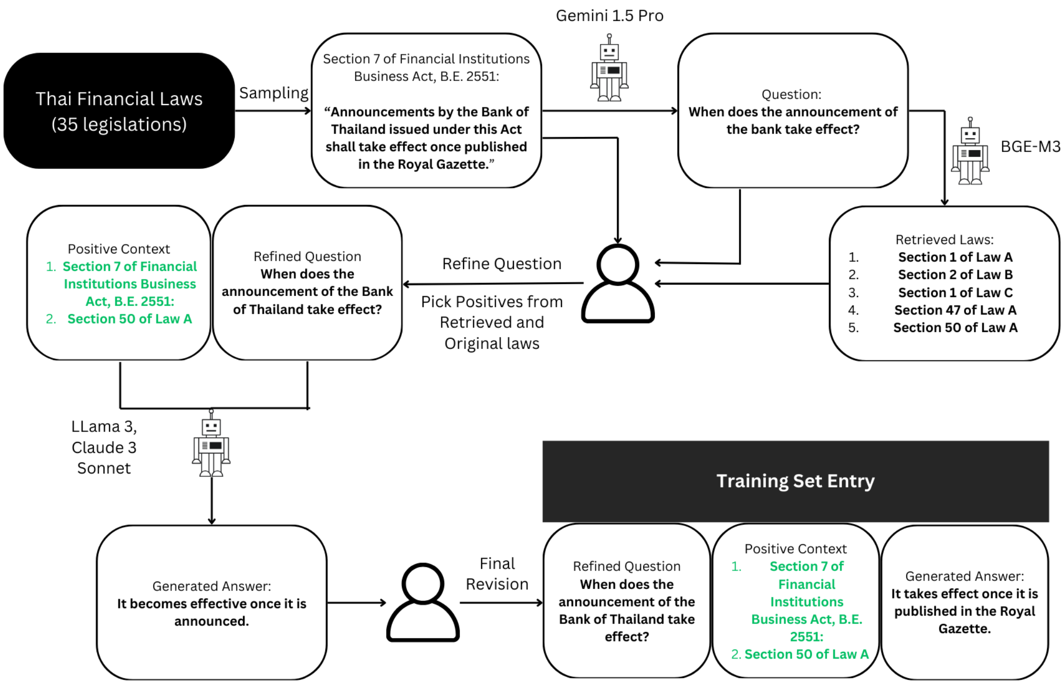


Figure 1: Overall dataset construction pipeline for training set of ThaiLegal-CCL

tion based on each annotator's speed of completion. Payment is determined per task[9], with each task compensated differently based on its difficulty. The tasks are as follows:

1. Rerank retrieved documents for the fine-tuning dataset: 5 THB (approximately $0.15)

[9]To simplify the calculations, we use a fixed conversion rate of 34 Thai baht per $1.

per task.

2. Validate, correct, and reject the generated answers for both training and test data: 10 THB (approximately $0.30) per task.

3. Create a question and answer based on a given law section (for the test set): 30 THB (approximately $0.89) per task.

The total cost spent solely on annotators is approximately 105,120 THB (roughly $3092).

B ThaiLegal-CCL Dataset Curation

ThaiLegal-CCL extends the original XYZ-ThaiLegal-CCL's test set by applying additioanl postprocessing step. Since the annotated contextual information includes the full content of relevant legal sections, we further preprocess the test set by extracting only the names of the referenced legal sections from the annotations and deduplicate entries with the same questions. Figure 2 illustrates the data collection process for ThaiLegal-CCL .

C ThaiLegal-Tax Dataset Curation

To evaluate the generalization capability of the system, we curated an additional dataset derived from publicly available resources in the Thai financial legal domain. Specifically, this dataset was created by scraping tax-related cases from the Revenue Department's official website[10]. These cases represent authentic inquiries or requests (with personally identifiable information removed) submitted to the department. Each case includes the original inquiry or request, the official response, and metadata such as the case ID and submission date. We extracted references to legislative sections mentioned in both the inquiry and the response as case attributes using `gpt-4o-mini-2024-07-18` for any preprocessing steps involving the use of LLM used during constructing ThaiLegal-Tax . The dataset was filtered to retain only cases referencing laws within the 35 Thai financial law codes and to eliminate duplicate references within individual entries. Some cases, however, involve inquiries requesting discretionary decisions from the department-such as extensions for tax deadlines or tax exemptions-rather than informational responses based on statutory interpretation. Since these cases are outside the scope of our work, which focuses on law-based reasoning, they were identified using an LLM and subsequently removed.

Additionally, to align with our evaluation objectives, the department's responses were condensed to essential answers, excluding detailed explanations and rationales. Finally, we restricted the dataset to cases from 2021 onward, reflecting the most recent legislative updates. The resulting ThaiLegal-Tax consists of 50 cases, predominantly related to the Revenue Code, with an average of three referenced legal sections per case. This dataset provides a challenging testbed for evaluating system performance in a specialized domain requiring nuanced legal reasoning and multi-label retrieval.

The complete dataset construction pipeline of ThaiLegal-Tax is outlined in Figure 3.

[10]https://www.rd.go.th

D Judge LLM Performance

Table 10 showed the final agreement score between human-annotated coverage and contradiction score compared to judge LLM-generated ones. LLM-as-a-judge is used for automatic evaluation, with prompts refined to achieve high agreement with human annotations (F1 > 0.8). The LLM-as-a-judge score is generated by `gpt-4o-2024-08-06` (Hurst et al., 2024) model with temperature of 0.3.

Metric	Dataset	Precision	Recall	F1-score	Support
Coverage	ThaiLegal-CCL	.88	.88	.88	200
	ThaiLegal-Tax	.83	.83	.83	150
Contradiction	ThaiLegal-CCL	.98	.97	.98	200
	ThaiLegal-Tax	.92	.91	.91	150

Table 10: Table displaying the weighted average precision, recall, and F1-score between metrics computed by LLM and annotated by human experts

To further analyze this agreement, we present confusion matrices for ThaiLegal-CCL and ThaiLegal-Tax in Tables 11 and 12, respectively. As observed in the confusion matrices, it is rare for the LLM-as-a-judge to misclassify a ground truth score of 0 as 100 or vice versa. Most errors occur in the confusion between 50 and 100, as well as between 0 and 50. We consider this acceptable since the boundaries between these scores can sometimes be subjective. Although the agreement scores did not reach our initial expectations after multiple iterations, we conclude that it remains reliable, achieving at least 80% accuracy for the coverage score and at least 90% accuracy for the contradiction score.

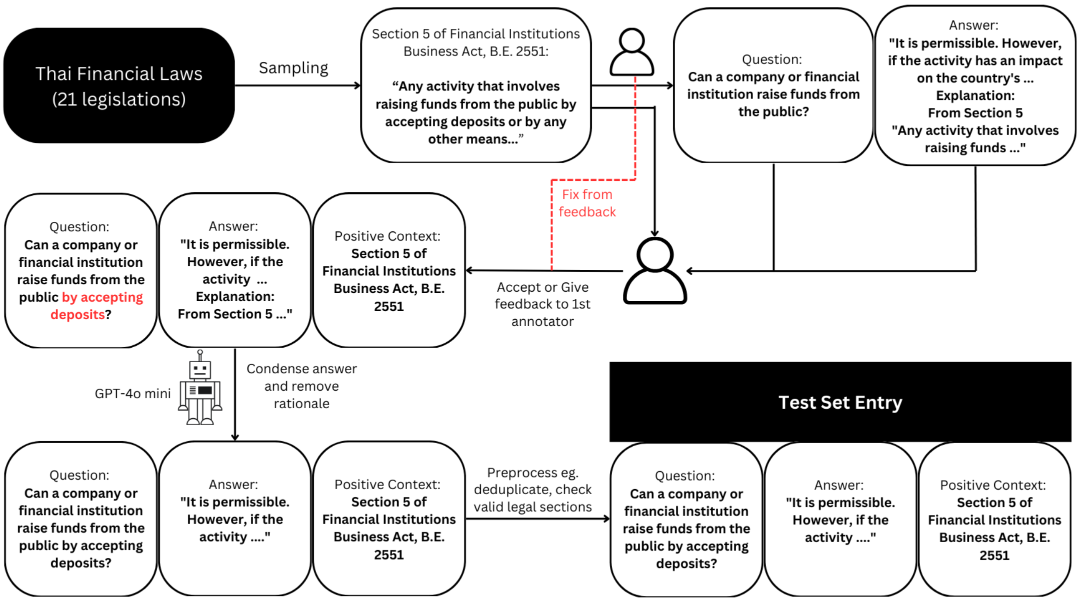


Figure 2: Overall dataset construction pipeline for test set of ThaiLegal-CCL

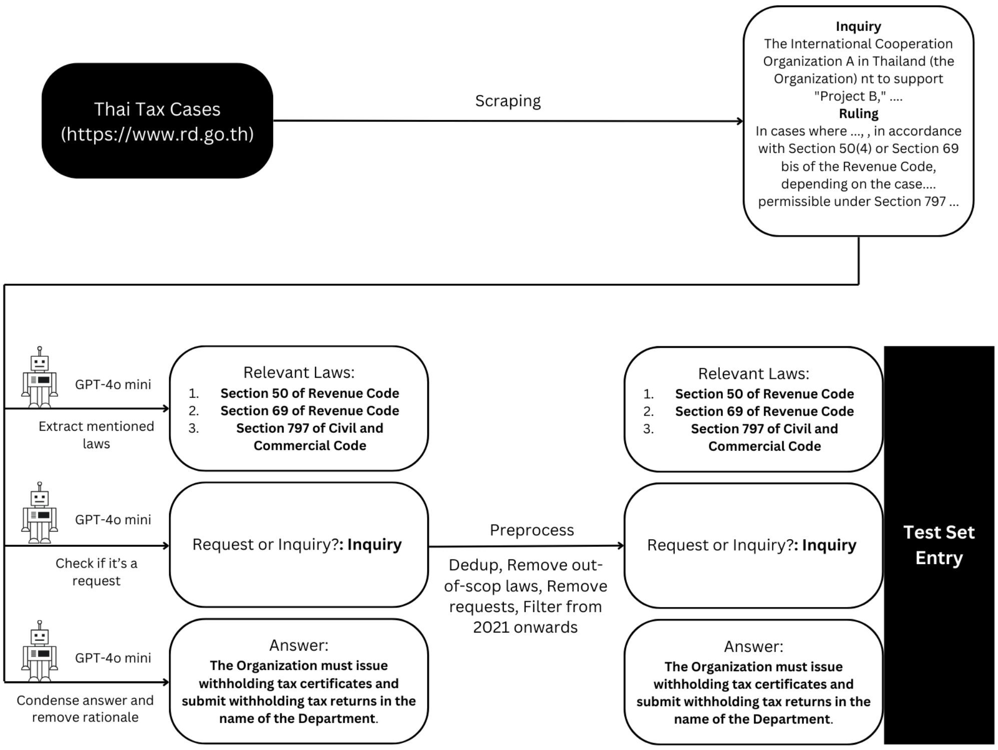


Figure 3: Overall dataset construction pipeline for ThaiLegal-Tax

E Thai Legal System

Thailand's legal system operates within a hierarchical structure (Chuathai, 2023), where lower-level laws must not contradict higher ones. The hierarchy includes the Constitution, Organic Laws, Acts/Codes, Emergency Decrees, Royal Decrees,

	Predicted 0	Predicted 50	Predicted 100
Ground Truth 0	8	2	3
Ground Truth 50	2	29	7
Ground Truth 100	1	9	139

Table 11: Confusion matrix for coverage agreement score on 200 ThaiLegal-CCL samples

	Predicted 0	Predicted 50	Predicted 100
Ground Truth 0	43	5	1
Ground Truth 50	6	35	6
Ground Truth 100	2	5	47

Table 12: Confusion matrix for coverage agreement score on 150 ThaiLegal-Tax samples

Ministerial Regulations, and Local Ordinances. Acts and Codes are primary legislation, with Acts encompassing individual laws and Codes structuring related provisions, such as the Criminal Code.

Acts and Codes are structured hierarchically. The structure proceeds from broad categories to increasingly specific details (Book, Title, Chapter, Division, Section, Subsection, Clause), with **Sections** being the fundamental legal units. This structure is designed for efficient navigation but creates challenges for RAG systems, specifically regarding how to chunk legislative documents while preserving the meaning. Furthermore, Thai legal text often utilizes inter-section references. For instance, understanding Section 260 of the Criminal Code

> "Whoever uses, sells, offers for sale, exchanges, or offers to exchange a ticket arising from the acts described in *section 258* or *section 259* shall be liable to imprisonment not exceeding one year or a fine not exceeding twenty thousand baht, or both." (The Kingdom of Thailand, 2022)

requires the context from section 258 and 259, which are not included in the same text segment. This raises questions about automatic retrieval and augmentation of referenced sections.

F Naive Chunking

We define naive chunking strategy as the best traditional chunking method that minimized "information loss" compared to our proposed hierarchical-aware chunking. Traditional chunking methods such as

- **Character Chunking:** Chunking is based purely on a fixed number of characters.
- **Recursive Chunking:** Chunking using various document structure-related separators.
- **Line Chunking:** Chunking based solely on newline characters.

often split sections naively via naive heuristic, leading to contextual "information loss" in section information. We quantify "information loss" via following metrics:

1. **Sections/Chunk**: Average sections per chunk.
2. **Chunks/Section**: Average chunks covering a section.
3. **Fail Chunk/Section Ratio**: Chunks/sections which are not fully covered.
4. **Uncovered Section Ratio**: Sections which are not covered at all.

Table 13 showed the information loss of different traditional chunking strategy. Notably, we decompose the problem of finding the best naive chunking strategy into two steps. First, we seek to find the best traditional chunking algorithm with the default parameter settings. After that, we further tune the chunking parameters-chunk size and overlap size-that further minimized the information loss. The best setups that will be referred as "naive chunking strategy" is line chunking using chunk size of 553 and overlap size of 50.

G Full Retrieval Model Performance

In addition to BM25 and BGE-M3 variants showed in the main experiment, we also conduct this experiments on various embeddings as well. The results is showed in Table 14. We choose 8 embeddings models for this experiment as follows:

1. BM25 (Robertson and Zaragoza, 2009)
2. JinaAI Colbert V2 (Jha et al., 2024)
3. JinaAI Embeddings V3 (Sturua et al., 2024)
4. NV-Embed V1 (Lee et al., 2024a)
5. BGE-M3 (Chen et al., 2024)
6. Human-Finetuned BGE-M3
7. Auto-Finetuned BGE-M3
8. Cohere Embeddings [11]

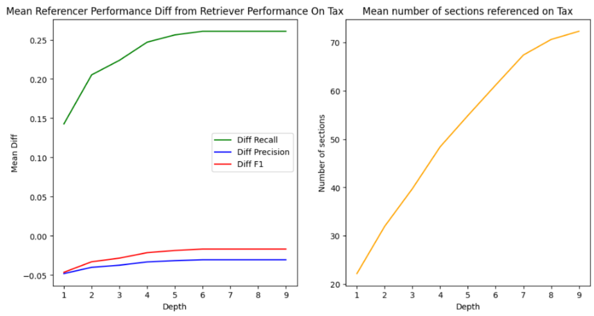
[11] https://cohere.com/blog/introducing-embed-v3

Chunking Strategy	Section/Chunk →1	Chunk/Section →1	Fail Chunk Ratio ↓	Fail Section Ratio ↓	Uncovered Section Ratio ↓
A. Chunking Result by Type of Chunking					
Hierarchy-awared	**1.000**	**1.000**	**0.000**	**0.000**	**0.000**
Character	3.098	1.710	0.819	0.675	0.397
Line	1.689	1.234	0.658	0.417	0.294
Recursive	1.793	1.270	0.741	0.504	0.381
B. Chunking Comparison between Hierarchy-aware and Best Naive Chunking					
Hierarchy-aware chunking	1.000	1.000	0.000	0.000	0.000
Line chunking (553 chunk size and 50 chunk overlap)	1.956	1.180	0.521	0.323	0.156

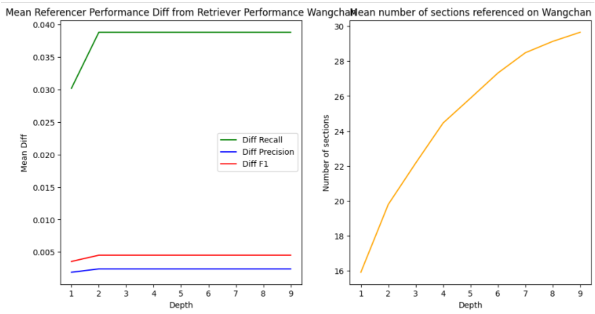
Table 13: **A.** The table showed the comparison of different naive chunking strategies compared ot our proposed hierarchy-awared chunking strategy. **B.** Using the best perform naive chunking strategy (notebly line chunking), we showed the line chunking with best parameter information loss (see §4.1) compared to hierarchy-awared chunking.

H Adding More Reference Depth

Adding more reference depth improves retrieval performance when the question requires extensive legal reasoning. To further investigate the effect of increasing LegalRefdepth towards performance, we examined the relationship between *LegalRef's maximum depth*, *retrieval performance gains* (`Mean Diff` on the y-axis), and *the total number of sections LegalRefresolves* (see Figures 4). For the Tax dataset, retrieval performance improves as reference depth increases, peaking at a depth of 6. However, this comes at the cost of increased context length, reaching approximately 60 sections per query. While the improvement in retrieval performance could be attributed to retrieving more sections—thereby increasing the hit rate—after extensive recursive reference resolution in ThaiLegal-Tax dataset, the results for the ThaiLegal-CCL dataset indicate that this is not always the case. For the ThaiLegal-CCL dataset, retrieval gains remain minimal and plateau after a depth of 2, despite resolving up to 30 sections at a depth of 9. We suspect this is due to the ThaiLegal-CCL dataset requiring only one relevant law per entry, eliminating the need for complex legal reasoning during retrieval.



(a) ThaiLegal-Tax dataset.



(b) ThaiLegal-CCL dataset.

Figure 4: Plots showing the relationship between depth of LegalRef and retrieval performance and number of sections per query on two datasets. (a) ThaiLegal-Tax dataset: `Mean Diff` shows the average retrieval metric difference when increasing section depth compared to retrieval performance without LegalRef. The right plot shows the number of sections cited when resolving more reference depth. (b) ThaiLegal-CCL dataset.

I LCLM Performance Analysis

The effect of the relevant context position in the overall documents on the performance of the system is analyzed on the sampled WCX dataset under the LCLM setting. The resulting performance is binned every 100,000 characters by the maximum depth of the relevant laws that need to be retrieved, and the coverage, contradiction, and E2E F1 of each bin are averaged and plotted in figure 5.

From the resulting plot, there is only a slight decrease in the coverage score and a slightly greater increase in the contradiction score as the depth increases. However, there is a significant drop in the E2E F1 score as the depth increases. Therefore, it can be concluded that **the depth of the relevant laws only mildly affects the coverage and contradiction score while its ability to cite applicable laws clearly has a negative impact.** Furthermore, the gains in performance in LCLM-as-a-retriever when increasing the number of retrieved documents are lower as compared to the gains of conventional

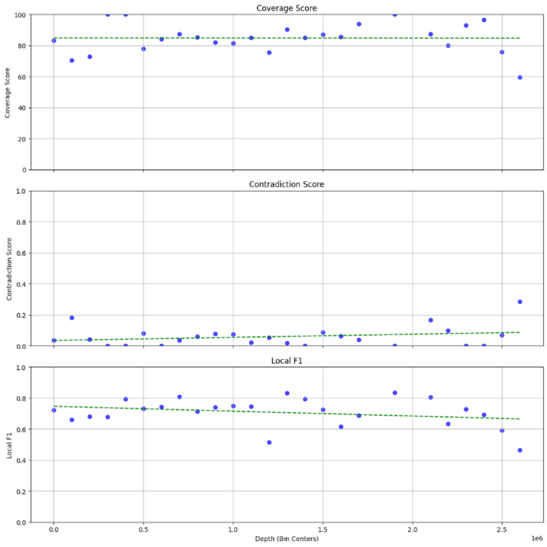


Figure 5: Plot of performance grouped by the maximum depth of relevant context in the long context

retrievers. We suspect that this is due to the next-token nature of LLM which limits its ability to retrieve meaningful sections at the lower ranks which are distant from the context and query.

ThaiLegal-CCL Dataset

Top-K	Model	HR/Recall@k	MRR@k
k=1	BM25	.481	.481
	JINA V2	.681	.681
	JINA V3	.587	.587
	NV-Embed V1	.492	.492
	BGE-M3	.700	.700
	Human-Finetuned BGE-M3	**.735**	**.735**
	Auto-Finetuned BGE-M3	<u>.731</u>	<u>.731</u>
	Cohere	.676	.676
k=5	BM25	.658	.548
	JINA V2	.852	.750
	JINA V3	.821	.681
	NV-Embed V1	.713	.579
	BGE-M3	.880	.773
	Human-Finetuned BGE-M3	**.906**	**.805**
	Auto-Finetuned BGE-M3	<u>.900</u>	<u>.800</u>
	Cohere	.870	.754
k=10	BM25	.715	.556
	JINA V2	.889	.755
	JINA V3	.875	.688
	NV-Embed V1	.776	.587
	BGE-M3	.919	.778
	Human-Finetuned BGE-M3	**.938**	**.809**
	Auto-Finetuned BGE-M3	<u>.934</u>	<u>.804</u>
	Cohere	.912	.760

ThaiLegal-Tax Dataset

Top-K	Model	HR@k	Multi HR@k	Recall@k	MRR@k	Multi MRR@k
k=1	BM25	.220	.080	.118	.220	.118
	JINA V2	.140	.040	.068	.140	.068
	JINA V3	.400	.100	.203	.400	.203
	NV-Embed V1	.100	.020	.035	.100	.035
	BGE-M3	<u>.500</u>	<u>.140</u>	<u>.269</u>	<u>.500</u>	<u>.269</u>
	Human-Finetuned BGE-M3	.480	<u>.140</u>	.255	.480	.255
	Auto-Finetuned BGE-M3	**.520**	**.160**	**.281**	**.520**	**.281**
	Cohere	.340	.100	.179	.340	.179
k=5	BM25	.480	.120	.254	.318	.171
	JINA V2	.200	.080	.114	.165	.085
	JINA V3	<u>.720</u>	**.260**	**.448**	.508	.297
	NV-Embed V1	.200	.020	.081	.126	.050
	BGE-M3	<u>.720</u>	<u>.240</u>	<u>.435</u>	<u>.580</u>	**.337**
	Human-Finetuned BGE-M3	**.740**	.220	.411	.565	.320
	Auto-Finetuned BGE-M3	.700	.200	.382	**.587**	<u>.329</u>
	Cohere	.620	.200	.363	.447	.256
k=10	BM25	.540	.160	.320	.327	.183
	JINA V2	.240	.100	.147	.171	.091
	JINA V3	**.840**	<u>.340</u>	<u>.549</u>	.524	.311
	NV-Embed V1	.220	.040	.097	.128	.052
	BGE-M3	<u>.820</u>	**.360**	**.555**	<u>.593</u>	**.354**
	Human-Finetuned BGE-M3	.800	.280	.499	.574	.333
	Auto-Finetuned BGE-M3	.780	.260	.483	**.600**	<u>.345</u>
	Cohere	.680	.200	.414	.454	.263

Table 14: Retrieval Evaluation Results on ThaiLegal-CCL Dataset and ThaiLegal-Tax Dataset with hierarchy-aware chunking.