A Additional Experiments

A.1 KBL

Here, we provides additional experimental results. Table 6 shows the results of the KBL Legal Knowledge benchmark conducted using the Llama-3.1-8B model. Similar to the result in Table 1, retrieving 5 documents from KoPS consistently yields better performance compared to the case of `kowiki`.

Table 6: Additional evaluation result on Legal Knowledge subtasks from KBL Benchmark (Kim et al., 2024b). Llama-3.1-8B and BM25 retriever were used. The number indicates the average scores over 7 subtasks.

Corpus	Reranker	Top-k=3	Top-k=5
w/o RAG		23.9	
KoPS	-	31.0	33.5
	colbert	31.6	33.0
	cross_encoder	29.4	31.8
	t5	30.0	31.6
kowiki	-	15.1	14.5
	colbert	15.1	14.5
	cross_encoder	14.9	14.9
	t5	14.3	15.7

Table 7 presents the 2025 Korean Bar Exam results with additional models.

Table 7: Evaluation results w/o RAG on 2024 Korean Bar Exam from KBL Benchmark (Kim et al., 2024b) with more various models.

Acc (%, ↑)	civil	public	criminal
Gemma-3-4b-it[a]	37.1	22.5	25.0
Gemma-3-12b-it[a]	28.6	35.0	42.5
EXAONE-3.0-7.8B-Instruct[b,*]	20.0	20.0	22.5
GPT-4o-mini-2024-07-18[*]	31.4	32.5	25.0

[a] Team et al. (2025). [b] Research et al. (2024).
[*] Results cited from Kim et al. (2024b).

A.2 LegalBench

Here we present our initial experiments on LegalBench. Table 8 presents the results of RAG experiments on LegalBench-tiny with Pile-of-Law-mini corpus. LegalBench Tiny is a subset of LegalBench (Guha et al., 2023), constructed by randomly sampling 10 instances per subtask, with sampling stratified to ensure a balanced distribution of correct answers. Pile-of-Law-mini corpus consists of 10% of randomly sampled documents from the original corpus.

Retrieval corpus We first evaluate RAG systems while varying their retrieval corpus: (1) no retrieval (Table 8 1st panel), (2) Wikipedia (2nd panel), and (3) Pile-of-Law-mini (3rd panel). The result highlight the clear importance of using domain specific legal corpus. Interestingly, the accuracy of GPT-4o-mini decreases the most with Pile-ofLaw-mini (3rd panel, 4th row). To investigate this, we altered the input order from instruction + retrieved-documents + examples + questions to retrieved-documents + instruction + examples + questions and observed an increase in accuracy (indicated by diff. prompt, final row of each panel). We suspect this behavior is due to the unique structure of legal documents and GPT-4o-mini's limited adaptation to the legal domain.

LLM backbones Next we demonstrate how the choice of LLM backbones, which generate the final answers, impacts performance (Table 8). The results reveals significant variance between models.

Table 8: Evaluation results of LegalBench-Tiny.

Model	Avg	Interpretation	Issue	Rhetorical	Rule
w/o RAG					
Llama3.1-8B	66.7	68.7	64.6	65.6	67.9
Qwen2.5-7B	67.7	72.5	70.5	62.9	64.7
SaulLM-7B	57.4	60.7	52.9	50.3	60.9
GPT-4o-mini	64.9	65.9	54.6	67.6	71.3
+ diff. prompt	72.7	75.2	73.3	74.2	67.9
BM25 for Wikipedia					
Llama3.1-8B	67.4 (+0.7)	70.4	70.7	60.0	68.4
Qwen2.5-7B	66.3 (-1.4)	72.1	70.1	60.5	62.3
SaulLM-7B	56.3 (-1.1)	62.9	50.0	53.5	58.8
GPT-4o-mini	58.9 (-6.0)	53.0	58.0	63.2	61.4
+ diff. prompt	71.4 (-1.3)	75.3	72.1	66.9	71.1
BM25 for Pile-of-Law-mini					
Llama3.1-8B	68.1 (+1.4)	69.6	71.5	63.5	67.8
Qwen2.5-7B	66.5 (-1.2)	72.7	66.6	60.4	66.1
SaulLM-7B	58.2 (+0.8)	63.3	52.9	55.6	61.1
GPT-4o-mini	55.6 (-9.3)	50.8	55.8	50.1	65.5
+ diff. prompt	73.1 (+0.4)	73.3	69.7	75.2	74.2

Retrieval algorithms Next, we examine the effect of the retrieval algorithm by replacing BM25 baseline with a dense retriever. We use LegalBERT-base (Chalkidis et al., 2020), and LexLM-base (Chalkidis* et al., 2023), as encoder backbone (Table 9). Using original DPR, fine-tuend for general-domain tasks, we observed similar performance to BM25 (1st vs 2nd rows). However, with domain-specialized encoders, there was a significant improvement in accuracy (3rd and 4th rows). When dense retriever finetuned on legal retrieval tasks was used, performance increased further (5th row), consistent with previous findings (Hou et al., 2024).

We also evaluated the effect of introducing ColBERT-based reranker(Khattab and Zaharia, 2020). Interestingly, the ColBERT reranker did not improve performance. This result suggests that using a reranker trained in the general domain can reduce the accuracy of RAG system, aligining with recent findings from (Pipitone and Alami, 2024).

Table 9: Performance table under varying retrieval algorithms (model fixed to GPT-4o-mini), The subset of Pile-of-Law is used as a retrieval pool. LegalBERT-C, LegalBERT-CR, and LegalBERT-CR$_{GPT}$ Stand for "LegalBERT-DPR-CLERC", "LegalBERT-DPR-CLERC + Reranker", "LegalBERT-DPR-CLERC + Reranker (GPT-4o)" respectively.

Retrieval Algorithms	**Avg**	**Interpretation**	**Issue**	**Rhetorical**	**Rule**
BM25	55.6	50.8	55.8	50.1	65.5
DPR	55.1	54.7	53.4	45.6	66.7
LegalBERT	60.4	54.7	55.6	59.0	72.3
LexLM-base	60.3	57.0	55.8	57.9	70.5
LegalBERT-C	63.7	60.0	54.2	71.4	69.0
BM25 + Reranker	55.1	50.4	54.2	51.2	64.5
LegalBERT-CR	63.5	58.7	54.8	70.9	69.7
LegalBERT-CR$_{GPT}$	63.5	58.4	58.2	67.5	69.7
LexLM-base + Reranker	60.4	58.8	53.3	63.9	65.7

Experiments shown in Table 2 LegalBench also include non-knowledge-intensive subtasks where external documents are not required to answer the questions. Additionally, Pile-of-Law comprises a wide range of legal documents, many of which may not be directely relevant. To better evaluate LRAGE in a more controlled setting and to enhance interpretability, we focus on the three knowledge-intensive subtasks from LegalBench and use all corresponding examples. Similarly, instead of random sampling, we construct subsets of Pile-of-Law by categorizing documents based on type.

A.3 LawBench

Table 10 presents additional results on LawBench for models not included in the main text. For InternLM2 (Cai et al., 2024), RAG improves performance in sections 3-3 and 3-4, but leads to lower scores in section 1-2. In contrast, for Qwen2.5-7B (Qwen Team, 2024), RAG improves scores in section 1-2 but results in lower scores in sections 3-3 and 3-4.

A.4 PLAT

Table 11 presents the rubric used in the PLAT experiment. The content was machine-translated from Korean to English.

Table 10: Evaluation of additional LLMs on LawBench (Fei et al., 2024). Three knowledge-intensive subtasks were evaluated here. 1-2: Knowledge Question Answering; 3-3: Charge Prediction; 3-4: Preson Term Prediction w.o. Article. We adopted Chinese Wikipedia (zhwiki) and the CAIL (Xiao et al., 2018) train set for the retrieval corpus.

LawBench	1-2	3-3	3-4
	ACC (%, ↑)	F1 (%, ↑)	-log distance (↑)
internlm2-chat-7b			
w/o RAG	**39.4**	50.0	62.1
CAIL	24.6	**52.0**	**74.5**
zhwiki	25.8	48.0	69.0
Qwen2.5-7B-Instruct-1M			
w/o RAG	29.4	**56.0**	**75.0**
CAIL	**54.5**	48.4	64.7
zhwiki	45.0	43.8	65.7

Table 11: Examples of rubrics used in the taxation dataset (PLAT)

Rubric Type	**Content**
Structural	"Below are 5 evaluation criteria (total 5 points) for the answer on ’The Legitimacy of the Penalty Tax Imposition’ based on the above case and explanation. 1. Structure and length of the writing (1 point): Evaluates whether the writing follows a logical order (introduction-main-conclusion, etc.) and is written concisely without unnecessarily excessive length (verbose description). 2. Formal completeness (1 point): Evaluates whether paragraphs are divided according to the logical flow without unnecessarily verbose expressions. 3. Clarity of introduction and problem statement (1 point): Whether the facts given in the case are concisely summarized and the issue (legitimacy of penalty tax imposition) is clearly presented. 4. Accuracy of citing relevant laws and precedents (1 point): Evaluates whether the laws and precedents necessary for problem-solving such as Value Added Tax Act, Enforcement Decree, Enforcement Rules, Framework Act on National Taxes, precedents, etc. are appropriately cited and properly connected to the necessary parts. 5. Adherence to expression (1 point): Evaluates whether the case overview and the requirements of the problem are faithfully reflected."
Semantic and Structural	"Below are 5 evaluation criteria (total 5 points) for the answer on ’The Legitimacy of the Penalty Tax Imposition’ based on the above case and explanation. Lower points are allocated to items evaluating form, while higher points are allocated to items evaluating content. 1. (Form) Structure and length of the writing (0.5 points) - Whether the writing follows a logical order (introduction-main-conclusion, etc.) - Whether it is written concisely without unnecessarily excessive length (verbose description) 2. (Content) Summary of facts and presentation of main issues (1 point) - Whether the facts appearing in the case are accurately identified and key issues are concisely presented - Whether it clearly emphasizes that the legitimacy of the penalty tax imposition is at issue 3. (Content) Appropriateness of relevant laws and interpretation (1 point) - Whether appropriate citations are made to the Corporate Tax Act (provisions regarding investment trusts being considered domestic corporations), Framework Act on National Taxes, or necessary tax law provisions - Whether it specifically explains how these provisions can/cannot be applied to impose penalty tax 4. (Content) Judgment of legitimate reasons and thoroughness of argumentation (1.5 points) - Whether the plaintiff’s argument (’investment trust is not a taxpayer’, ’excessive refund was inevitable’) and the defendant’s argument (’penalty tax imposition is justified for excessive refund application’) are compared and examined - Whether the existence of ’legitimate reasons’ that could excuse the plaintiff from negligence in the refund procedure is logically analyzed 5. (Content) Validity and clarity of conclusion presentation (1 point) - Whether a clear conclusion is drawn on whether the penalty tax imposition is legitimate or illegitimate - Whether the reasons supporting the conclusion (key issues and results of legal review) are presented concisely and clearly"