KoLegalQA: A Korean Legal QA Dataset for Trustworthy and Explanation-Grounded Legal AI

Yongtae Lee[1], Surin Lee[1], Sumin Kim[1], S M Wahidur Rahman[2], Heung-No Lee[1,2,*]
[1]Department of AI Convergence, Gwangju Institute of Science and Technology, Korea
[2]Department of EECS, Gwangju Institute of Science and Technology, Korea
{lyt98313,leesurin,smkim6927,sm.wahidur}@gm.gist.ac.kr, heungno@gist.ac.kr
* Corresponding author

Abstract

Legal QA systems may benefit from training data that is expert-verified and associated with statutory provisions, as fluent generation alone cannot guarantee legally relevant and citation-supported outputs. However, existing Korean legal datasets provide limited support for legal QA and statute-associated response generation. To address this gap, we introduce KoLegalQA, a large-scale Korean legal question–answer corpus designed for research on legal QA and explanation-oriented legal response generation in real-world consultation scenarios. The dataset comprises 19k consultations collected from government-operated services, with all responses originally authored or verified by licensed legal professionals. Unlike prior resources, KoLegalQA provides explicit statutory references and clause-level summaries, enabling research on citation-associated and explanation-oriented legal response generation. We benchmark six Korean-capable LLMs using both automated evaluation (G-Eval) and human assessment across multiple criteria, including legal correctness, reasoning quality, and citation relevance. Experimental results show that fine-tuning on KoLegalQA generally improves legal reasoning validity and statute-associated response generation across most evaluated models. We present this resource as a practical benchmark dataset for Korean legal NLP research. Dataset splits, preprocessing scripts, and evaluation code will be publicly released to support reproducible research.

1 Introduction

Access to legal information remains a persistent challenge for many individuals in Korea, particularly among socially or legally vulnerable groups (Min and Gil, 2026; Lim et al., 2024). Legal systems rely on complex statutory language and procedural rules, making it difficult for non-experts to interpret their rights and obligations without professional assistance. Although public institutions have introduced various digital services—including online legal consultation platforms and legislative databases—these systems primarily provide static information and limited interactivity (Lim et al., 2024; Ministry of Science and ICT and National Information Society Agency, 2025). As a result, many individuals still struggle to obtain accessible and understandable legal guidance.

Recent advances in large language models (LLMs) have demonstrated strong capabilities in natural language understanding and question answering. These models offer the potential to support interactive legal assistance systems. However, their deployment in legal domains raises concerns regarding factual reliability and unsupported claims. In high-stakes domains such as law, incorrect or misleading responses may lead to serious real-world consequences (Huang et al., 2025). Therefore, developing legal QA systems for such settings may benefit from datasets that include statutory references and explanation-oriented annotations.

Despite growing interest in legal NLP, existing Korean legal datasets remain limited in their support for legal QA and statute-associated response generation. Most available resources focus on tasks such as document classification, statute retrieval, or court judgment analysis. They rarely provide aligned question–answer structures, practitioner-authored explanations, or annotations that reflect the reasoning process behind legal guidance. Compared with English-language legal NLP ecosystems—where standardized datasets and expert-curated benchmarks are more widely available—the Korean legal NLP landscape still faces limitations in data accessibility, availability of expert-written resources, and the large-scale supervised QA datasets derived from real legal consultations (Chalkidis et al., 2022; Guha et al., 2023).

To address these limitations, we introduce **KoLegalQA**, a comprehensive Korean legal ques-

Proceedings of the 6th Workshop on Trustworthy NLP (TrustNLP 2026), pages 240–255

tion–answer dataset designed for research on legal QA and statute-associated response generation. The dataset contains 19,266 real-world consultations collected from government-operated legal services, including the Korean Legal Aid Corporation (KLAC), the Easy to Find Practical Law (EFPL) platform, and web resources from the Ministry of Government Legislation (MOLEG). The responses were originally authored or verified by licensed legal professionals, ensuring both legal reliability and practical relevance. In addition, many instances include statutory references and concise clause-level summaries associated with the relevant legal provisions. These annotations provide supervision signals that associate legal answers with relevant statutory context and explanation-oriented information.

Using KoLegalQA, we conduct supervised fine-tuning experiments on multiple Korean-capable LLMs and evaluate them using both automated and human assessment protocols. Our experiments examine whether fine-tuning on the dataset improves performance on legal QA tasks involving statutory references and explanation-oriented responses. To support reproducibility, the dataset splits, preprocessing scripts, and evaluation code will be publicly released upon publication. Our contributions are summarized as follows:

- **An expert-verified large-scale Korean legal QA dataset derived from public legal consultation services.** We introduce **KoLegalQA**, a Korean legal QA dataset consisting of more than 19k real-world consultation cases collected from government-operated legal services. The dataset provides a large-scale resource for training and evaluating Korean legal QA models.
- **Statute-associated annotations for explanation-oriented legal QA.** Beyond simple QA pairs, many instances include explicit statutory references and clause-level summaries associated with relevant legal provisions. These annotations provide additional supervision signals beyond plain QA pairs. This dataset can also be used to support research on statute-associated and explanation-oriented legal response generation.
- **Benchmark experiments and evaluation settings for Korean legal QA.** We validate model outputs through automated assessment on held-out test data not used during training. To validate machine evaluation, we conduct blinded human assessment across 100 sampled questions. The experiments provide empirical observations on model behavior after fine-tuning with KoLegalQA. Moderate-to-strong alignment between G-Eval and human ratings, together with fair-to-moderate inter-rater reliability, suggests that the evaluation protocol provides reasonably consistent assessments across the evaluated legal QA responses.

2 Related Work

Legal NLP Benchmarks in English. Legal NLP has seen substantial progress in English, supported by a variety of benchmark datasets. **LexGLUE** (Chalkidis et al., 2022) provides a multi-task benchmark covering case outcome prediction, statute classification, and contract understanding. **LegalBench** (Guha et al., 2023) introduces a broad suite of legal reasoning tasks for large language models, including multilingual legal question answering. In addition, **MultiLegalPile** (Niklaus et al., 2024) offers a large multilingual legal corpus for pretraining language models, enabling cross-jurisdictional modeling at scale.

Korean Legal NLP Resources. In contrast, Korean legal NLP remains relatively underdeveloped. General-purpose Korean QA datasets such as **KorQuAD 1.0** (Lim et al., 2019) and domain-specific corpora like **KLAID** [1] provide only partial coverage of legally oriented information needs. These datasets lack expert-authored legal QA pairs, structured legal domain labels, and resources designed for generative legal reasoning.

Recent work by Hwang et al. (Hwang et al., 2022) introduced **LBOX OPEN**, a large-scale Korean legal benchmark containing court judgments, classification tasks, and legal judgment prediction. They also released **LCUBE**, a Korean legal language model pretrained on judicial texts. While LBOX OPEN significantly advances Korean legal NLP by covering diverse judicial decision-making processes, it focuses primarily on court-authored documents and judgment outcomes.

Trustworthy Legal AI and Verifiable Reasoning. Recent research has emphasized the impor-

[1] https://huggingface.co/lawcompany/KLAID

tance of developing reliable language models, particularly in high-stakes domains such as law and medicine (Huang et al., 2025). In legal applications, incorrect or unsupported answers may mislead users about their rights or obligations, making reliability and citation relevance important considerations for legal AI systems. However, large language models are prone to generating hallucinated or unsupported claims, which can lead to severe ethical and professional risks. To mitigate these issues, prior studies have explored approaches for citation-associated generation, citation-aware reasoning, and explanation-oriented responses (Zhou et al., 2025; Es et al., 2024). These efforts highlight the need for datasets that provide not only answers but also statutory references and associated explanations linked to authoritative legal sources.

Expert-Verified Data vs. Crowdsourcing. The reliability of legal NLP systems depends heavily on the quality of the underlying supervision signals. While many NLP benchmarks rely on crowdsourced annotations, such data often lack the nuance and technical accuracy required for complex legal reasoning (Guha et al., 2023). This limitation is particularly critical in the legal domain, where subtle differences in statutory interpretation may lead to substantially different conclusions. TrustNLP initiatives advocate for "human-in-the-loop" systems and expert-curated corpora to improve accountability and reliability. **KoLegalQA** aligns with these principles by utilizing consultations exclusively authored or verified by licensed legal professionals. This expert-driven approach addresses the inherent limitations of general-purpose or crowdsourced datasets, providing a high-quality resource for training and evaluating legal QA models.

Positioning of KoLegalQA. Existing Korean legal resources do not include citizen-facing legal questions, natural-language answers, or explicit mappings to statutory provisions derived from public consultations. **KoLegalQA** addresses this gap by providing high-quality QA pairs collected from government-operated legal counseling services, reflecting authentic legal questions from the general public. Each QA instance is mapped to a bar-exam-aligned category scheme and accompanied by curated statutory explanations. This design supports explanation-oriented legal QA generation and enables more comprehensive evaluation of Korean legal language models in real-world consultation scenarios. Consequently, KoLegalQA provides a practical benchmark for studying legal QA generation with statutory references in Korean.

3 KoLegalQA Dataset

3.1 Data Sources and Collection

KoLegalQA is constructed from publicly accessible legal consultation records provided by three government-operated platforms in South Korea. All sources publish real-world legal inquiries and expert-authored responses on open web pages without access restrictions. The collected data preserve the original question–answer structure of each consultation.

KLAC. The Korean Legal Aid Corporation (KLAC) provides legal consultation services primarily for economically or socially disadvantaged individuals (Korea Legal Aid Corporation). The platform publishes consultation records in which citizens submit legal questions and licensed attorneys provide written responses. After removing HTML artifacts, formatting inconsistencies, and non-substantive boilerplate expressions, 9,618 high-quality QA pairs were retained. Legal references embedded in the responses were preserved to maintain the statutory grounding of each answer.

EFPL. EFPL is a government-operated portal maintained by the Ministry of Government Legislation that provides accessible legal guidance on everyday legal issues (Korea Ministry of Government Legislation, a). Its content includes structured explanatory materials as well as question–answer entries addressing practical legal concerns. For QA entries, preprocessing involved textual extraction from structured web content, normalization of formatting and punctuation, and removal of visually oriented elements (e.g., tables or image references) that are not directly compatible with text-only modeling. Statutory references and legally meaningful markers were preserved. After processing, 2,156 QA pairs were included.

MOLEG. MOLEG publishes official legal interpretations addressing inquiries submitted by citizens, public officials, or legal professionals (Korea Ministry of Government Legislation, b). Each entry follows a question–answer format in which the inquirer presents a factual scenario and identifies the relevant statute requiring interpretation. Responses are drafted by legally qualified personnel within the

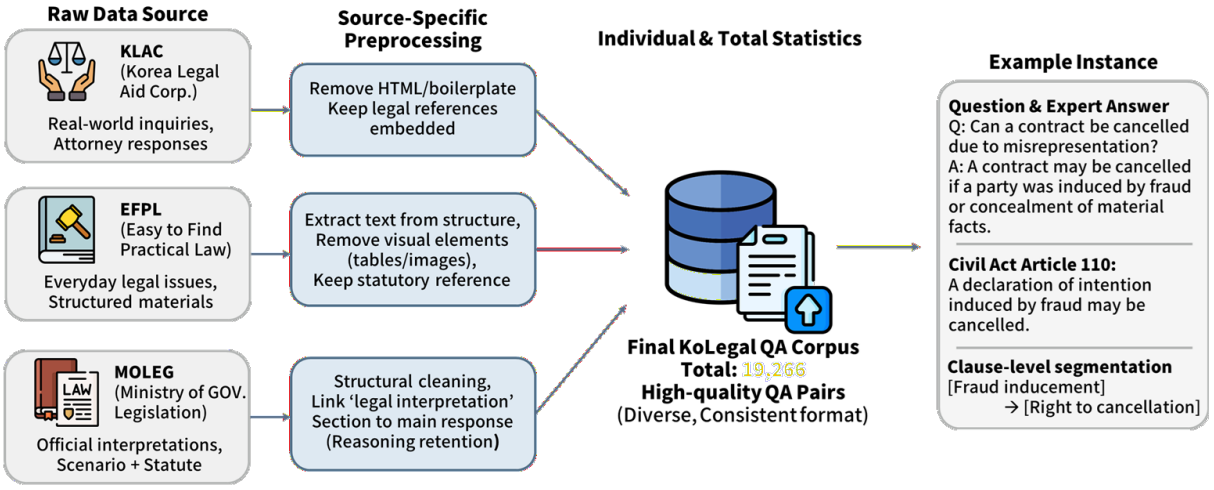

Figure 1: **KoLegalQA dataset construction pipeline and instance illustration.** Multi-source legal resources from KLAC, EFPL, and MOLEG are normalized through shared preprocessing, with MOLEG documents undergoing an additional interpretation-linking stage that preserves reasoning structure. The processed data are aggregated into a structured dataset comprising 19,266 instances. A miniature example on the right illustrates the composition of a single instance, including question–answer pairs, grounded statutory text, and clause-level segmentation, supporting explanation-grounded legal reasoning.

Ministry's legal interpretation division. Preprocessing for MOLEG data followed procedures similar to those applied to other sources, including structural cleaning and normalization. In addition, the section labeled "legal interpretation" was explicitly linked to the main response to ensure that the reasoning and statutory grounds were retained as an integral part of each answer. After processing, 7,492 QA pairs were incorporated into the dataset.

3.2 Clause-Level Summary Annotation

Clause-level summaries were constructed as auxiliary explanations linked to statutory references and consultation answers. They are based on expert-authored legal consultations provided by government-operated legal services, where legal professionals originally produced structured explanations as part of their official advisory process. These explanations were extracted from the archived records during preprocessing and stored together with the corresponding question–answer pairs and related statutory provisions.

The summaries provide fine-grained interpretative signals that complement statutory citations by describing relevant legal clauses in natural language. This enables models to learn how specific statutes are applied in practical consultation scenarios.

All instances containing clause-level summaries were included without additional paraphrasing, preserving the original expert-provided structure as much as possible.

3.3 Dataset Statistics

In total, KoLegalQA contains more than 19k high-quality question–answer pairs collected from three government-operated legal services. The dataset is entirely composed of Korean-language legal consultations and statutory materials. Any English examples presented in this paper and appendix are translated versions provided solely for illustrative and readability purposes.

The dataset covers a wide range of real-world legal issues, including civil disputes, administrative procedures, and consumer protection. Each instance preserves the original consultation structure and is augmented with statutory explanations and clause-level summaries, enabling models to learn explanation-grounded legal reasoning patterns.

To further characterize the distribution of statutory grounding across sources, we analyzed statutory annotation density by measuring both the proportion of answers containing at least one statutory annotation and the number of statutory annotations per answer. MOLEG exhibited substantially higher annotation density (mean: 11.71, median: 10) than EFPL (mean: 0.86, median: 0) and KLAC (mean: 0.48, median: 0), reflecting the statute-centric nature of legal interpretation cases. In contrast, EFPL and KLAC primarily contain consumer-oriented legal guidance and consultation responses, where explicit statutory citations are less frequent.

4 Experiments

4.1 QA Generation Task

To evaluate the effectiveness of KoLegalQA for explanation-grounded legal reasoning, we formulate a supervised question–answer generation task. Given a legal inquiry, the model is required to generate a formal legal response that (i) provides a legally sound conclusion and (ii) articulates the underlying statutory reasoning supporting that conclusion. This task evaluates whether language models can generate legally grounded and verifiable responses rather than unsupported or hallucinated legal claims, which is critical in high-stakes domains such as law.

4.2 Experimental Setup

All experiments were implemented using the Hugging Face Transformers library (`transformers==4.57.3`) and the PEFT/4-bit quantization stack. Models were loaded with automatic device mapping and executed in a CUDA environment when available.

Data Split. The full dataset was randomly divided into a training split (80%) and a held-out evaluation split (20%) using a fixed random seed. Duplicate questions had been removed during earlier preprocessing stages, and the split was performed without explicit stratification by data source. The held-out split was used both for checkpoint selection (based on evaluation loss) and for downstream automated and human evaluation. All reported evaluation results were obtained exclusively on the held-out split, which was not used for gradient updates during training. Human evaluation samples were also drawn only from this held-out portion. Since preprocessing-stage duplicate removal was performed before splitting, the risk of direct train–test leakage through repeated consultation entries was reduced.

Models. We evaluated six publicly available 7B–9B LLM checkpoints with strong Korean capabilities, selected from public Korean LLM leaderboards and widely used Hugging Face models. Four are Llama-family models (Beomi[2], Bllossom[3], Princeton[4], and Seokdong[5]), while two models use non-Llama backbones (Gemma[6] and Monarch[7]). Full Hugging Face identifiers for all models are provided in Table 5, Appendix C.

All models were fine-tuned using parameter-efficient instruction tuning under a unified recipe (4-bit quantization with LoRA adapters), enabling a controlled comparison of KoLegalQA supervision across heterogeneous model backbones.

All models contain between 7B and 9B parameters. Larger models were not considered due to hardware constraints and to examine the effectiveness of KoLegalQA under realistic resource-limited settings. However, we additionally tested `openai/gpt-oss-20b` as an ablation for parameter scaling.

Training Configuration. We adopted QLoRA-style parameter-efficient fine-tuning with 4-bit NF4 quantization (compute dtype: bfloat16) and LoRA adaptation. LoRA adapters were configured with rank r=16, scaling α=32, dropout 0.05, and no bias terms, and were inserted into both attention and MLP projection layers.

The maximum sequence length was set to 512 tokens, with the prompt truncated to at most 256 tokens. For each instance, loss was computed only over answer tokens by masking prompt and padding tokens with `-100`. We enabled gradient checkpointing and disabled KV caching (`use_cache=False`) to reduce memory usage.

Models were trained for 10 epochs with learning rate 2×10^{-5} using paged AdamW 8-bit optimization (`paged_adamw_8bit`). The effective batch size was 16 (per-device batch size 4 with gradient accumulation steps of 4). We evaluated and saved checkpoints at the end of each epoch, and selected the best checkpoint based on the lowest evaluation loss.

Prompting Strategies. Each training instance was formatted using a structured instruction template that explicitly includes the legal domain label:

[2] https://huggingface.co/beomi/Llama-3-Open-Ko-8B

[3] https://huggingface.co/MLP-KTLim/llama-3-Korean-Bllossom-8B

[4] https://huggingface.co/princeton-nlp/Llama-3-8B-ProLong-512k-Base

[5] https://huggingface.co/SEOKDONG/llama3.1_korean_v1.1_sft_by_aidx

[6] https://huggingface.co/recoilme/recoilme-gemma-2-9B-v0.4

[7] https://huggingface.co/mlabonne/NeuralMonarch-7B

```
### Question:
{question} [Domain: {category}]

### Answer:
```

In implementation, we used the same template structure with Korean headers ('질문', '답변') to match the dataset language. The model was trained to autoregressively generate the answer conditioned on the question prompt. The prompt portion was masked during training, so that optimization targets only the generated response. No chain-of-thought prompting or external statute injection was applied during inference, isolating the effect of supervised fine-tuning on KoLegalQA.

Inference Configuration. To ensure fair comparison, all base and fine-tuned models were decoded under the same deterministic generation setup. We used greedy decoding with `do_sample=False` and `num_beams=1`, generating up to 512 new tokens per question. The same prompt template as training was used at inference time, and the generated response was extracted by decoding only the continuation tokens beyond the prompt. For efficiency, inference was performed in batches (batch size 8) with explicit `pad_token_id` and `eos_token_id`.

4.3 Evaluation Metrics

Automatic Evaluation (G-Eval). We adopt the G-Eval framework (Liu et al., 2023) with GPT-4o as an automated evaluator to assess the reliability and reasoning quality of generated legal responses. For each test question, both the base model output and the fine-tuned model output were scored independently.

We evaluate responses along four criteria, consisting of three legal-ability dimensions and one non-legal dimension:

- **Conclusion correctness (CC):** whether the response reaches a legally valid conclusion.
- **Legal reasoning validity (LR):** whether the response provides logically coherent legal justification supporting the conclusion.
- **Citation relevance (CR):** whether cited statutes or precedents are directly relevant and meaningfully support the reasoning.
- **Clarity and tone (CL):** whether the response is clear and maintains an appropriate tone.

Each criterion was rated on a 1–5 scale with explicit anchors (1: incorrect/irrelevant, 3: partially correct, 5: fully correct). To ensure deterministic scoring, the temperature was set to zero and the evaluator was constrained to produce JSON-only outputs.

G-Eval was implemented using a structured evaluation prompt that presents the question and the model-generated answer, followed by the scoring rubric and formatting constraints. The evaluation prompt template is shown in Figure 2.

In implementation, we used the Korean version of the same template to match the dataset language, while preserving identical criteria definitions, score anchors, and the JSON-only constraint to ensure consistency across evaluation settings.

Human Evaluation. To complement automated scoring, we conducted a human evaluation on 100 randomly sampled test questions. The human evaluation set was sampled from the held-out evaluation split described in Section 4.2. The sampled subset approximately preserves the original source distribution of the full dataset: KLAC (52% vs. 49.9%), EFPL (9% vs. 11.2%), and MOLEG (39% vs. 38.9%). This sampling procedure was intended to reduce source-specific sampling bias during human assessment.

From the six models, three representative models were selected based on G-Eval results: (i) the highest-performing fine-tuned model, (ii) the model with the largest improvement after fine-tuning, and (iii) a model showing balanced performance.

For each question, evaluators assessed both base and fine-tuned outputs from the selected models. A total of 12 graduate students in computer science and AI engineering participated as evaluators, following detailed annotation guidelines. Before annotation, evaluators were provided with instructions and example responses for each score level and completed a short calibration round to ensure consistent understanding of the evaluation criteria (Van der Lee et al., 2021; Gehrmann et al., 2023).

Each question–model pair was independently rated by three evaluators in a blinded setting, where both the model identity and the training condition were concealed. Instead of adjudicating disagreements, we report the mean score across raters to reduce individual bias and capture consensus judgments.

```
You are a strict and objective evaluator for Korean legal question - answering.

Read the [Question] and [Answer] below,
then assign a score from 1 to 5 for each criterion.

Scoring rubric:
- 1: strongly disagree
- 3: partially agree
- 5: strongly agree
(2 and 4 indicate intermediate levels between adjacent anchors.)

[Criteria]
1. Conclusion correctness:
   Does the answer provide a legally correct conclusion to the issue raised?
2. Legal reasoning validity:
   Is the legal justification sufficient to support the conclusion?
3. Citation relevance:
   Are the cited statutes or precedents directly relevant to the issue?
4. Clarity and tone:
   Is the response clearly written and easy to understand?

[Question]
{question}

[Answer]
{prediction}

Respond only in the following JSON format.
Do not include any additional text.

{
  "conclusion_correctness": <1-5>,
  "legal_reasoning": <1-5>,
  "citation_relevance": <1-5>,
  "clarity_and_tone": <1-5>
}
```

Figure 2: Prompt template used for G-Eval-based automatic evaluation.

Responses were scored on a 1–5 scale across the same four dimensions used in G-Eval (Liu et al., 2023; Zhong et al., 2022). Final scores were obtained by averaging across raters (Fabbri et al., 2021). Inter-rater reliability was measured using a two-way random-effects intraclass correlation coefficient for absolute agreement (ICC$(2, k)$) (Koo and Li, 2016; Shrout and Fleiss, 1979). To examine the alignment between automated and human evaluation, we computed Spearman rank correlations between G-Eval and human scores (Spearman, 1904; Ruscio, 2008).

4.4 Results

Outputs were rated on four criteria: (a) conclusion correctness, (b) legal reasoning validity, (c) citation relevance, and (d) clarity and tone (1–5 scale; higher is better). We report *Legal-score* as the arithmetic mean of (a–c) and *Overall* as the mean of all four criteria.

G-Eval Scores. Automated evaluation using the G-Eval framework (Liu et al., 2023) shows that fine-tuning on KoLegalQA generally improves legal response quality across models. Table 1 reports Legal-score, clarity, and overall scores for the base and fine-tuned models.

Notably, Gemma exhibited a performance drop after fine-tuning across all reported dimensions, indicating that dataset-specific supervision does not uniformly benefit all model architectures. This suggests that the effectiveness of dataset supervision varies across model architectures, highlighting that a one-size-fits-all approach may be insufficient for developing reliable legal AI.

Across models, fine-tuning substantially improved legal-score in most cases. The largest improvement was observed for Princeton (overall: 1.70 $\rightarrow$ 2.68; +0.98), while Seokdong achieved the best overall score among fine-tuned models (3.21). In contrast, Gemma showed a consistent drop after

Type	Model	Legal-score	Clarity	Overall
Base	Beomi	2.37	2.52	2.41
	Bllossom	1.73	1.83	1.75
	Gemma	**3.48**	**3.61**	**3.51**
	Monarch	2.45	3.00	2.59
	Princeton	1.76	1.55	1.70
	Seokdong	2.91	3.53	3.06
FT	Beomi	3.18	2.36	2.98
	Bllossom	2.77	2.07	2.59
	Gemma	3.29	2.58	3.11
	Monarch	2.97	2.59	2.88
	Princeton	2.86	2.14	2.68
	Seokdong	**3.35**	**2.79**	**3.21**

Table 1: G-Eval scores across models (1–5; higher is better).

Metric	Spearman's ρ	ICC(2,3)
CC	0.53 [.47, .59]	0.44 [.31, .55]
LR	0.58 [.52, .63]	0.47 [.29, .60]
CR	0.56 [.51, .62]	0.43 [.21, .58]
CL	0.64 [.59, .69]	0.48 [.39, .56]
Legal-score	0.60 [.55, .65]	0.47 [.28, .60]
Overall	0.64 [.59, .69]	0.50 [.35, .60]

Table 2: Alignment between G-Eval and human ratings (Spearman's ρ) and inter-rater reliability among three human evaluators (ICC(2,3)). Both are reported with 95% CIs. Metric abbreviations follow Section 4.3. Legal-score averages CC, LR, and CR. Overall averages CC, LR, CR, and CL. Scores are on a 1–5 scale (higher is better). All correlations and ICCs are significant ($p < .001$). 95% CIs for Spearman's ρ were computed via bootstrapping.

fine-tuning ($3.51 \rightarrow 3.11$), suggesting that the effectiveness of dataset supervision may vary across model architectures. This observation highlights that developing trustworthy legal AI may require architecture-aware adaptation strategies rather than uniform fine-tuning pipelines.

Clarity improvements were less consistent than legal-score gains, indicating that optimizing for legally grounded explanations may not always translate into higher perceived readability.

Human Rating Scores. Human evaluation was conducted on a random sample of 100 test questions for three representative models, comparing both base and fine-tuned outputs. Each question–system output was independently rated by three evaluators using the same four criteria as G-Eval, and final human scores were computed by averaging across raters.

To assess the convergent validity of G-Eval scores, we examined the correlation between automated and human ratings across evaluation criteria (Table 2). Spearman's rank correlation coefficients showed moderate-to-strong monotonic relationships across all criteria ($\rho = .53$–$.64$, all $p < .001$). The legal-score and overall quality scores showed particularly strong associations (legal-score: $\rho = .60$, 95% CI [.55, .65]; overall quality: $\rho = .64$, 95% CI [.59, .69]; both $p < .001$), indicating that G-Eval reliably captures relative differences in response quality. Confidence intervals excluded zero for all criteria.

Inter-rater reliability among the three human evaluators was assessed using a two-way random-effects ICC for absolute agreement (ICC$(2, k)$; Table 2). Agreement fell in the fair-to-moderate range across criteria (ICC(2,3) = .43–.48), suggesting that averaging ratings across three evaluators yields reasonably stable human scores, although some variability in absolute scale usage remains. The overall quality score yielded an ICC of .50 (95% CI [.36, .60], $p < .001$), suggesting moderate consistency in absolute rating levels across evaluators. Taken together, these results suggest that G-Eval provides a reasonably reliable proxy for comparative legal QA assessment, and that averaging across three evaluators yields stable absolute scores despite residual variability in individual-criterion scale usage.

As an additional parameter-scaling analysis, we evaluated a larger model, `openai/gpt-oss-20b`. This analysis further examines whether dataset-grounded supervision remains effective as model scale increases. The corresponding G-Eval results are reported in Appendix D.1.

5 Conclusion

This paper presented KoLegalQA, a large-scale Korean legal QA dataset of more than 19k real-world consultations drawn from government-operated platforms. These sources institutionally ensure that all published responses are authored and validated by licensed legal professionals as part of their official mandate. By providing clause-level summaries and statutory annotations, KoLegalQA offers supervision signals that associate legal responses with relevant statutory reasoning and explanation-

oriented information.

To demonstrate its utility, we fine-tuned six publicly available Korean-capable language models under a unified resource-efficient recipe and evaluated outputs across four criteria using both automated and human assessment. Five out of six models showed consistent improvements in legal reasoning validity and citation relevance after fine-tuning, suggesting that expert-verified and statute-associated supervision can improve performance on legal QA tasks. At the same time, the consistent performance drop observed in Gemma after fine-tuning indicates that dataset supervision does not uniformly benefit all architectures, highlighting the need for model-specific adaptation strategies in legal AI development rather than one-size-fits-all fine-tuning pipelines.

Human evaluation conducted on 100 sampled questions across three representative models further corroborated these findings. Ratings from 12 graduate evaluators in a blinded setting showed that fine-tuned models generally outperformed their base counterparts across the evaluated criteria. Inter-rater reliability fell in the fair-to-moderate range, and moderate-to-strong alignment between G-Eval and human scores suggests that automated evaluation can provide a reasonably consistent proxy for comparative legal QA assessment in this setting. Together, these results position KoLegalQA as a practical benchmark resource for Korean legal NLP research. Dataset splits, preprocessing scripts, training configurations, and evaluation code will be publicly released upon publication.

Limitations

While KoLegalQA enables explanation-grounded legal QA fine-tuning and yields consistent gains under automated and human evaluation, several limitations remain.

First, groundedness is evaluated indirectly via citation relevance, without explicit verification of statutory correctness using retrieval or external legal databases. Future work should include statute-linked supervision or retrieval-based validation to better enforce legal grounding.

Second, the evaluation protocol is proxy-based and relies on a combination of G-Eval and non-expert human evaluators. Although moderate-to-strong alignment was observed between automated and human ratings, these evaluations should not be interpreted as equivalent to professional legal assessment.

Third, the dataset does not systematically cover broader legal sources such as judicial opinions or procedural documents. This limits evaluation of document-intensive legal reasoning and broader legal decision-making scenarios.

Fourth, direct comparison with existing legal QA benchmarks remains limited because KoLegalQA is derived from Korean-language government legal consultation services and includes source-specific statutory annotations and explanation-oriented structures that are not directly aligned with most existing English-language legal NLP benchmarks.

Finally, the experimental protocol uses a held-out split for evaluation instead of separate validation and test split. Additionally, human evaluation is limited to 100 sampled questions. Larger-scale evaluation settings and broader human assessment would strengthen the generalizability of the findings.

Acknowledgments

This work was supported by the IITP (Institute of Information & Communications Technology Planning & Evaluation)-ITRC (Information Technology Research Center) grant funded by the Korea government (Ministry of Science and ICT) (IITP-2026-RS-2021-II211835). This work was also supported by the IITP grant funded by the Korea government (MSIT) (No. 2019-0-01842, Artificial Intelligence Graduate School Program (GIST)). This work was supported by the National Research Foundation of Korea (NRF) grant funded by the Korea government (MSIT) (RS-2025-22932973). We also appreciate the high-performance GPU computing support of HPC-AI Open Infrastructure via GIST SCENT.

References

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.

S. Es, J. James, L. E. Anke, and S. Schockaert. 2024. Ragas: Automated evaluation of retrieval augmented generation. In *Proceedings of the 18th Conference of*

the European Chapter of the Association for Computational Linguistics: System Demonstrations, pages 150–158.

Alexander R Fabbri, Wojciech Kryściński, Bryan McCann, Caiming Xiong, Richard Socher, and Dragomir Radev. 2021. Summeval: Re-evaluating summarization evaluation. *Transactions of the Association for Computational Linguistics*, 9:391–409.

Sebastian Gehrmann, Elizabeth Clark, and Thibault Sellam. 2023. Repairing the cracked foundation: A survey of obstacles in evaluation practices for generated text. *J. Artif. Int. Res.*, 77.

Yeong Hyeon Gu, Xianghua Piao, Helin Yin, Dong Jin, Ri Zheng, and Seong Joon Yoo. 2022. Domain-specific language model pre-training for Korean tax law classification. *IEEE Access*, 10:46342–46353.

Neel Guha, Julian Nyarko, Daniel Ho, Christopher Ré, Adam Chilton, Aditya K, Alex Chohlas-Wood, Austin Peters, Brandon Waldon, Daniel Rockmore, Diego Zambrano, Dmitry Talisman, Enam Hoque, Faiz Surani, Frank Fagan, Galit Sarfaty, Gregory Dickinson, Haggai Porat, Jason Hegland, and 21 others. 2023. Legalbench: A collaboratively built benchmark for measuring legal reasoning in large language models. In *Advances in Neural Information Processing Systems*, volume 36, pages 44123–44279. Curran Associates, Inc.

Lei Huang, Weijiang Yu, Weitao Ma, Weihao Zhong, Zhangyin Feng, Haotian Wang, and 1 others. 2025. A survey on hallucination in large language models: Principles, taxonomy, challenges, and open questions. *ACM Transactions on Information Systems*, 43(2):1–55.

Wonseok Hwang, Dongjun Lee, Kyoungyeon Cho, Hanuhl Lee, and Minjoon Seo. 2022. A multi-task benchmark for Korean legal language understanding and judgement prediction. In *Advances in Neural Information Processing Systems*, volume 35, pages 32537–32551. Curran Associates, Inc.

Terry K Koo and Mae Y Li. 2016. A guideline of selecting and reporting intraclass correlation coefficients for reliability research. *Journal of chiropractic medicine*, 15(2):155–163.

Korea Legal Aid Corporation. Legal counseling cases. https://www.klac.or.kr/legalinfo/counsel.do. Accessed: 2026-02-20.

Korea Ministry of Government Legislation. a. Easy-to-find, practical law: Q&a (100 questions, 100 answers). https://easylaw.go.kr/CSP/OnhunqueansLstRetrieve.laf. Accessed: 2026-02-20.

Korea Ministry of Government Legislation. b. Legal interpretation cases. https://www.moleg.go.kr/lawinfo/nwLwAnList.mo?mid=a10106020000. Accessed: 2026-02-20.

Seong-min Lim, Gwan-seon Jeong, Seung-uk Yang, and Seong-hwa Kim. 2024. A study on enhancing access to justice for older adults (고령자의 사법접근권 제고 방안에 관한 연구). Research report, Judicial Policy Research Institute (JPRI), Goyang, South Korea.

Seungyoung Lim, Myungji Kim, and Jooyoul Lee. 2019. KorQuAD1. 0: Korean QA dataset for machine reading comprehension. *arXiv e-prints*, pages arXiv–1909.

Yang Liu, Dan Iter, Yichong Xu, Shuohang Wang, Ruochen Xu, and Chenguang Zhu. 2023. G-eval: NLG evaluation using GPT-4 with better human alignment. In *Proceedings of the 2023 conference on empirical methods in natural language processing*, pages 2511–2522.

Kyung-ah Min and Yu-mi Gil. 2026. A basic study for the statistical measurement of the vulnerable under the legal and institutional framework (법제도상 취약계층의 통계적 측정을 위한 기초 연구). Technical Report Research Report 2025-07, National Data Research Institute, South Korea.

Ministry of Science and ICT and National Information Society Agency. 2025. 2024 report on the digital divide (2024 디지털정보격차 실태조사 보고서). Research report, Ministry of Science and ICT. Survey conducted by Gallup Korea.

Joel Niklaus, Veton Matoshi, Matthias Stürmer, Ilias Chalkidis, and Daniel Ho. 2024. MultiLegalPile: A 689GB multilingual legal corpus. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 15077–15094, Bangkok, Thailand. Association for Computational Linguistics.

John Ruscio. 2008. Constructing confidence intervals for spearman's rank correlation with ordinal data: A simulation study comparing analytic and bootstrap methods. *Journal of Modern Applied Statistical Methods*, 7(2):416–434.

Alice Saebom Kwak, Cheonkam Jeong, Ji Weon Lim, and Byeongcheol Min. 2024. A Korean legal judgment prediction dataset for insurance disputes. *arXiv e-prints*, pages arXiv–2401.

Patrick E Shrout and Joseph L Fleiss. 1979. Intraclass correlations: uses in assessing rater reliability. *Psychological bulletin*, 86(2):420.

Charles Spearman. 1904. The proof and measurement of association between two things. *The American Journal of Psychology*, 15(1):72–101. Original introduction of Spearman's rank correlation.

Chris Van der Lee, Albert Gatt, Emiel Van Miltenburg, and Emiel Krahmer. 2021. Human evaluation of automatically generated text: Current trends and best practice guidelines. *Computer Speech & Language*, 67:101151.

Ming Zhong, Yang Liu, Da Yin, Yuning Mao, Yizhu Jiao, Pengfei Liu, Chenguang Zhu, Heng Ji, and Jiawei Han. 2022. Towards a unified multi-dimensional evaluator for text generation. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 2023–2038, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Xiaoling Zhou, Mingjie Zhang, Zhemg Lee, Wei Ye, and Shikun Zhang. 2025. Hademif: Hallucination detection and mitigation in large language models. In *The Thirteenth International Conference on Learning Representations*.

A Existing Datasets

LBOX OPEN (Hwang et al., 2022) have advanced research in legal document classification and similar case retrieval, respectively. KorQuAD 1.0 (Lim et al., 2019) provides a strong foundation for reading comprehension in Korean. Table 3 shows the overview of known Korean Legal Datasets.

B Dataset Detailed

The dataset under consideration is partitioned into two primary categories, namely KLAC and EFPL, as specified in the ensuing example data. And you can specifically check if the category of EFPL is mapped in the table 4.

B.1 KLAC.

B.1.1 Civil Procedure Law Example

Question: While crossing a crosswalk, I was hit by an uninsured vehicle driven by person A who ignored a stop sign, resulting in an 8-week injury. Person A made a deposit payment as criminal settlement money, and I received a deposit notice but lost it. Can I withdraw the deposit money without the deposit notice?

Answer: In principle, when a deposit beneficiary wishes to request payment of deposited funds, the deposit notice must be attached (Article 33, Paragraph 1 of the Deposit Regulations). However, in cases like yours where the deposit notice has been lost, the Deposit Regulations provide the following provisions to allow you to retrieve the deposited funds.

Specifically, you can retrieve the deposited funds by obtaining either the original deposit certificate or a consent form from the depositor (a document indicating consent to payment request without attaching the deposit notice) (Article 33, Paragraph 1, Item B of the Deposit Regulations). If you cannot obtain such a deposit certificate or consent form from the depositor, you can retrieve the deposited funds by submitting a handwritten guarantee signed by two or more people recognized by the deposit officer, stating that they will jointly compensate for any damages that may arise from payment without the deposit notice, along with their property certificates (such as certified copies of register entries), and copies of their identification cards (Article 41, Paragraph 1 of the Deposit Regulations).

Additionally, if a qualified representative such as an attorney or a judicial scrivener acts as a proxy for the payment request, they can submit a guarantee in the name of the qualified representative (which must be signed and sealed by the qualified representative) instead of the guarantee under Article 41, Paragraph 1 of the Deposit Regulations (Article 41, Paragraph 3 of the Deposit Regulations).

Furthermore, if the amount of the deposit being requested is 50 million won or less, you can retrieve the deposited funds without attaching the deposit notice (Article 33, Paragraph 1, Item A of the Deposit Regulations).

Source: Korea Legal Aid Corporation
Type: Legal Information
Dataset: Legal Consultation Cases
Original Category: Civil Execution
Category: Civil Procedure Law

B.1.2 Labor Law Example

Question: What is the minimum wage system, and how is the minimum wage amount determined?

Answer: The minimum wage system refers to a system where the state establishes a minimum limit for wages and requires employers to comply with it. Therefore, when a minimum wage amount is determined and announced, even if an employer and worker agree to pay less than the minimum wage, such an agreement is automatically invalid, and the employer must pay at least the minimum wage amount announced by the Minister of Employment and Labor (Minimum Wage Act, Article 6).

The minimum wage system applies to all businesses or workplaces that employ workers, including regular workers as well as temporary workers, daily workers, and part-time workers. However, it does not apply to businesses that employ only family members living together and domestic workers. It also does not apply to seafarers covered by the Seafarers Act and shipowners who employ them (Article 3 of the same Act).

Workers who are in a probationary period, within three months from the start of their probation, may be paid 90% of the hourly minimum wage, and workers engaged in surveillance or intermittent work (such as guards, security personnel, private drivers, etc.) who are approved by the Minister of Employment and Labor may be paid 80% of the hourly minimum wage (Article 5, Paragraph 2 of the same Act, Article 3 of the Enforcement Decree of the same Act).

However, Article 6 of the Enforcement Decree of the same Act stipulates that employers may, with

Dataset Name	Main Tasks	Size and Features
LBOX OPEN (Hwang et al., 2022)	Classification, Judgment Prediction, Summarization	147,000 legal cases; includes 6 tasks; **Offers detailed metadata for each case.**
KorQuAD 1.0 (Lim et al., 2019)	Machine Reading Comprehension QA	Over 70,000 QA pairs; based on Wikipedia; **General domain QA, not specific to legal texts.**
KTL-BERT (Gu et al., 2022)	Tax Law Query Classification	327,735 tax law questions; **Specialized in tax law.**
Insurance Dispute LJP(Law Judgment Prediction) (Saebom Kwak et al., 2024)	Judgment Prediction	Small-scale insurance dispute cases; **Includes verdicts and reasoning.**

Table 3: Overview of Korean Legal Datasets

the approval of the Minister of Employment and Labor, exclude from the application of the minimum wage 'persons whose mental or physical disabilities clearly and directly hinder the performance of the work to which they are assigned.'

The Minister of Employment and Labor must request a review of the minimum wage by March 31 each year from the Minimum Wage Commission, which consists of worker representatives, employer representatives, and public interest representatives. The commission reviews the minimum wage proposal considering:

- Workers' cost of living
- Wages of similar workers
- Labor productivity
- Income distribution rates

When the commission submits a minimum wage proposal, the Minister must promptly announce the minimum wage proposal by industry type and the scope of applicable businesses, and must determine the minimum wage by August 5 of each year (Articles 4, 5, 8, 9 of the same Act, Article 7 of the Enforcement Decree of the same Act).

For reference, the minimum wage announced by the Ministry of Employment and Labor for the period from January 1, 2016 to December 31, 2016 is set at 6,030 won per hour.

Source: Korea Legal Aid Corporation
Type: Legal Information
Dataset: Legal Consultation Cases
Original Category: Labor
Category: Labor Law

B.2 EFPL Classification Mapping

B.3 EFPL.

B.3.1 Civil Law Example

Question: We were living together and suddenly we got a notice to break up, can I get alimony if we are common-law married?
Answer: Yes, you can receive alimony. A common-law marriage can be dissolved by agreement between the couple or by unilateral termination by either party. In this case, you can claim alimony from the spouse who unilaterally terminated the common-law marriage without just cause (a cause that falls under Article 840 of the **Civil Code**) or from a third party who caused the breakdown of the common-law marriage (for example, the spouse's parents).If there is no agreement on alimony, you can file an alimony claim with the court to receive alimony. The following are the grounds for divorce (trial grounds for divorce):

- **Civil Code** Article 840
  1. The spouse has committed an unfaithful act
  2. The spouse has abandoned the other party in bad faith
  3. The spouse or his/her immediate dependents have been treated grossly unfairly
  4. When his or her immediate dependent has been treated grossly unfairly by his or her spouse
  5. When the life or death of his or her spouse is not certain for more than three years
  6. When there are other serious reasons that make it difficult to continue the marriage

Source: Legal Affairs
Type: Find Easy Life Law Information

Original EFPL Classification	Legal Field Mapping
Family Law	Civil Law
Traffic/Driving	- Driving accidents, unlicensed/drunk driving, traffic violation penalties: **Criminal Law** - Administrative penalties (points, license revocation) or accident compensation systems, insurance: **Administrative Law** - Used Cars/Rental Cars: **Civil Law** - International Driving Permit: **International Law**
Government/Local Authorities	Administrative Law
National Defense/Veterans	Administrative Law
Labor/Employment	Labor Law
Finance/Financial Affairs	Commercial Law
Trade/Immigration	International Law
Culture/Leisure	Administrative Law
Civil/Criminal Litigation	- Administrative appeals/litigation: **Administrative Law** - Medical law related: **Administrative Law** - Jury trials, indictment-related matters: **Criminal Procedure Law** - Others: **Civil Procedure Law**
Welfare	Administrative Law
Real Estate/Lease	Civil Law
Business	Commercial Law
Public Safety/Crime	Criminal Law
Consumer Affairs	Commercial Law
Children/Youth/Education	- Employment, part-time work: **Labor Law** - Consumer damages, privacy infringement, copyright: **Civil Law** - Fraud, school violence, sexual offenses: **Criminal Law** - Others: **Administrative Law**
Information and Communication Technology	Administrative Law
Entrepreneurship	Commercial Law
Environment/Energy	Environmental Law

Table 4: Mapping of EFPL classification items to legal fields. Some items are subdivided into multiple legal fields based on their content.

Dataset: Family Law White Paper
Original Category: Family Law
Category: Civil Law

B.3.2 Administrative Law Example

Question: I think I left my wallet on the subway when I went to work. It has my credit card and ID card in it, and I'm quite concerned. What should I do?

Answer: If you left something on the subway during your commute, don't panic. First, contact the terminal station of the subway you were on and the station where you got off. You can also inquire at a nearby police station or patrol district, or check if your lost item has been registered on the "National Police Agency's Lost and Found Comprehensive Information Website." Additionally, you can access the websites of subway operating agencies in each region or public transportation lost and found centers to see if they are holding your lost item.

For lost credit cards, call the card company immediately to request a suspension of use. It's also advisable to report the loss of your ID card promptly, as it could be used for identity theft or other crimes.

- If you left something on the subway:
  - First, contact the terminal station of the subway you were on and the station where you got off. If you transferred, it's best to contact the terminal stations of all

sections you traveled through.

    * Since subway systems handle an enormous amount of lost items daily, it's advisable to describe the characteristics of your lost item in as much detail as possible (time of loss, appearance, contents, distinguishing features, etc.) when you contact them.

  – You can also inquire at a nearby police station or patrol district, or access the "National Police Lost and Found Portal" to check if your lost item has been registered or to report it.

  – Additionally, you can check if the Seoul Metro website or public transportation lost and found centers have your lost item.

- Reporting the loss of identification cards and getting replacements:

  – Reporting the loss of and replacing a Resident Registration Card:

    * First, visit your local town/township/district office or access the "Government 24 (www.gov.kr)" website to report the loss of your Resident Registration Card, complete a reissuance application form, and apply for a replacement.

  – Reporting the loss of and replacing a Driver's License:

    * First, visit a nearby driver's license examination office or access the "Korea Road Traffic Authority's Safe Driving Integrated Civil Service (www.safedriving.or.kr)" website to report the loss of your driver's license, complete a reissuance application form, and apply for a replacement.

Source: Ministry of Government Legislation
Type: Find Easy Life Law Information
Dataset: Life Law Q&A
Original Category: Traffic/Driving
Category: Administrative Law

C Model Configuration and Compute Resources

Table 5 lists the six LLM checkpoints used in our experiments, and the compute environment is described below.

Compute Environment: All models were trained and evaluated using NVIDIA A100-SXM4-40GB GPUs (Ampere architecture). A total of 6 GPUs were used in parallel by three researchers.

D Additional Experiments

D.1 Parameter-Scaling Ablation: `openai/gpt-oss-20b`

Experimental Conditions. We conducted a parameter-scaling ablation using `openai/gpt-oss-20b`. Training was performed with data-parallel distributed training across four GPUs using `accelerate` (4 processes, one process per GPU). In contrast to the main 7B–9B runs, which can be executed on a single device with automatic device mapping, the 20B model was loaded on each rank with an explicit single-device mapping to avoid cross-device sharding under DDP. Unless otherwise stated, we kept the overall objective (supervised instruction tuning on KoLegalQA with answer-only loss masking) and the downstream evaluation protocol identical to the main experiments. Inference generation and G-Eval scoring were performed using the same code and deterministic decoding configuration as in the main experiments.

The training pipeline for `gpt-oss-20b` required minimal but model-specific adaptations. First, model loading used the `Mxfp4Config` quantization path (when available in `transformers`) rather than the NF4 quantization stack used for the 7B–9B models; computation used bf16 when supported (otherwise fp16). Second, inputs were formatted using the model's chat template via `tokenizer.apply_chat_template` with a lightweight system message and user message (question and domain), and prompt tokens were masked so that loss is computed only on assistant response tokens. Third, the prompt budget was set to a smaller maximum (`PROMPT_MAX=156`; `MAX_LENGTH=512`) to satisfy memory constraints at 20B scale. Finally, we used a larger global effective batch size under DDP (per-device train batch size 8 with gradient accumulation 16) with paged AdamW 8-bit optimization, gradient check-

Model Alias	Hugging Face ID	Details
Beomi	beomi/Llama-3-Open-Ko-8B	8B parameters LLaMA-3 architecture
Bllossom	MLP-KTLim/llama-3-Korean-Bllossom-8B	8B parameters LLaMA-3 architecture
Gemma	recoilme/recoilme-gemma-2-9B-v0.4	9B parameters Gemma-2 architecture
Monarch	mlabonne/NeuralMonarch-7B	7B parameters Merged model based on Mistral models and others
Princeton	Llama-3-8B-Prolong-512k-Base	8B parameters LLaMA-3 architecture
Seokdong	SEOKDONG/llama3.1_korean_v1.1_sft_by_aidx	8B parameters LLaMA-3.1 architecture

Table 5: Model configurations used in KoLegalQA experiments. We refer to each model by the alias shown here throughout the paper.

Setting	Legal-score	Clarity & tone	Overall
Base	1.88 ± 1.05	1.74 ± 0.97	1.85 ± 0.98
Fine-tuned	2.56 ± 1.29	1.99 ± 1.07	2.42 ± 1.18

Table 6: G-Eval results for the parameter-scaling ablation using openai/gpt-oss-20b. Values are mean $\pm$ standard deviation over the held-out evaluation split (1–5; higher is better). Training was conducted with DDP across four GPUs; inference and G-Eval scoring followed the same deterministic protocol as the main experiments.

pointing enabled, and DDP stabilization options (e.g., ddp_find_unused_parameters=False and max_grad_norm=0.3). Tokenized datasets were cached to disk to reduce repeated preprocessing overhead.

Fine-tuning improves both the legal-score and overall quality for openai/gpt-oss-20b, consistent with the trend observed in the 7B–9B models. The improvement is primarily driven by gains in legal-score, while clarity also increases modestly. Compared to the 7B–9B results, we do not observe an obvious reduction in the magnitude of fine-tuning gains at this larger scale; the improvement is broadly comparable to that of several 7B–9B baselines. Nonetheless, since this observation is based on a single 20B ablation with minor implementation differences, we refrain from making stronger scaling claims. Overall, these results suggest that KoLegalQA supervision remains beneficial even at larger parameter scales under the same deterministic evaluation protocol.

E Use of AI Assistants

We used ChatGPT-4 during the research process to support code debugging, LaTeX formatting, and early-stage drafting. All outputs were manually reviewed, edited, and integrated by the authors. No content was directly copied without modification.