LEGALMIDM: USE-CASE-DRIVEN LEGAL DOMAIN SPECIALIZATION FOR KOREAN LARGE LANGUAGE MODEL

Youngjoon Jang[*], **Chanhee Park**[*], **Hyeonseok Moon, Heuiseok Lim**[†]
Department of Computer Science and Engineering, Korea University
{dew1701, pch7678, glee889, limhseok}@korea.ac.kr

Young-kyoung Ham, Jiwon Moon, Jinhyeon Kim, JuKyung Jung
KT (Korea Telecom)
{youngkyoung.ham, jiwon.moon, jin_hyeon.kim, jukyung.jung}@kt.com

ABSTRACT

In recent years, the rapid proliferation of open-source large language models (LLMs) has spurred efforts to turn general-purpose models into domain specialists. However, many domain-specialized LLMs are developed using datasets and training protocols that are not aligned with the nuanced requirements of real-world applications. In the legal domain, where precision and reliability are essential, this lack of consideration limits practical utility. In this study, we propose a systematic training framework grounded in the practical needs of the legal domain, with a focus on Korean law. We introduce LEGALMIDM, a Korean legal-domain LLM, and present a methodology for constructing high-quality, use-case-driven legal datasets and optimized training pipelines. Our approach emphasizes collaboration with legal professionals and rigorous data curation to ensure relevance and factual accuracy, and demonstrates effectiveness in key legal tasks.

1 INTRODUCTION

In recent years, state-of-the-art large language models (LLMs) such as OpenAI's ChatGPT OpenAI et al. (2024) and Meta's Llama Grattafiori et al. (2024) have demonstrated remarkable performance across a wide range of tasks, including mathematics, programming, and logical reasoning Lightman et al. (2024); Gu et al. (2024). These models are typically trained for general-purpose use, aiming for broad applicability across diverse scenarios Das et al. (2024). However, off-the-shelf LLMs often require additional adaptation to meet the specific needs of specialized fields such as legal Jeong (2024); Jung & Jung (2025), finance Li et al. (2023); Lee et al. (2025), and medical Liu et al. (2024); Singhal et al. (2022), leading to increased industry interest in domain-specific LLMs.

Although recent research highlights the potential of domain-specialized LLMs Bousetouane (2025), we find that current domain specialization strategies often lack rigorous consideration of real-world use-case requirements Mumcuoğlu et al. (2021); Greco & Tagarelli (2024). Especially in the legal domain, there is a high demand for AI-assisted workflows, yet current LLM training frameworks often fail to fully address these needs. Instead of taking actual requirements and demands of the legal field into account, models are trained based on naive assumptions of potential utility within this domain Padiu et al. (2024); Yao et al. (2024b); Valerio et al. (2024); Baack et al. (2025). To fill this gap, we present a use-case-driven framework for developing LLMs in the legal domain, with a focus on Korean law, where there is strong demand for specialized AI tools but limited exploration of practical adaptation strategies. With domain experts' guidance, we construct a high-quality training dataset grounded in real cases. Building on this dataset, we design a practical domain-adaptation pipeline that can be applied to open LLMs.

[*]Equal contribution.
[†]Corresponding author.

Concretely, this paper (i) describes an end-to-end recipe for transforming open LLMs into domain experts, with a focus on data curation and training pipelines, and (ii) conducts ablation studies that quantify how key design choices (data composition, synthetic data formatting, and prompt-based instruction-tuning) affect both legal and general domain performance. All experiments are conducted using Mi:dm[1], a Korean-specialized LLM, to develop LEGALMIDM, a vertical LLM for the Korean legal domain. Through this research, we hope to catalyze the development of more robust and practical vertical LLM construction frameworks.

2 RELATED WORK

With technological advancements, a wide variety of general domain LLMs have become publicly available, including Llama Grattafiori et al. (2024), Mistral Jiang et al. (2023), and Gemma Team et al. (2024). Several attempts have been made to create domain-specific LLMs based on these foundational models, particularly in fields such as finance and law Shen et al. (2024); Langa et al. (2024); Satterfield et al. (2024); Zhu et al. (2024). These efforts involve post-training on domain-specific corpora Jeong (2024); Xie et al. (2024).

However, these attempts often rely on collecting domain-related texts or operate under simplistic assumptions about domain requirements Yao et al. (2024b); Huang et al. (2023); Zhang & Yang (2023). In the legal domain, high-demand tasks often involve significant ambiguity Macey-Dare (2023); Siino et al. (2025). Recent studies have shown the great potential of LLMs' application in the legal domain Shu et al. (2024); Yao et al. (2024a); Zhou et al. (2024), yet, we observe that few of these existing studies justify their data compositions or optimal formats of the inputs and outputs. In response, we propose a pipeline for constructing vertical LLMs that takes into account actual domain needs and specific LLM use cases.

3 REQUIREMENT INVESTIGATION

In this section, we analyze the key considerations for building a legal domain-specific LLM for service applications. We design guidelines for constructing training data tailored to legal domain models by examining use cases of LLMs in the legal domain and general LLM services.

3.1 LEGAL TECH INDUSTRIES

First, we analyze the currently operational legal tech services to identify high-demand tasks within the legal domain. We examine a total of 20 legal tech companies in the United States and South Korea, focusing on the types of services they provide. Table 1 lists the AI-based legal tech services available in the United States and South Korea as of June 1, 2025, along with their technological features. These services were identified using the Google Search API with "legal AI" as the search query, selecting the top 10 search results for companies in each country. We adjusted the search region setting for each country during our investigation.

Our analysis reveals that most legal tech services aim to address tasks related to Question Answering (QA), Document-based QA (DQA), Document Summarization (DS), and Document Drafting (DD). Based on these findings, we aim to set tasks with high practical demand in the legal domain. We consult two practicing attorneys to assess the suitability of AI for each task and finalize our task selection. Through this, we select two high-stakes tasks in document drafting: complaints and petitions. In addition, we include the multiple choice question task, often considered a conventional tool even in the legal domain Siino et al. (2025); Guha et al. (2023), to facilitate the training and assessment of fundamental legal understanding. Consequently, we finalize six tasks: Summary, Document-based QA, Open QA, Complaint, Petition, and Multiple Choice QA, for subsequent learning and evaluation.

3.2 HUMAN-CURATED DATA COMPOSITION

To address these needs, we construct both training and evaluation datasets for the legal domain. Legal experts actively participate in the data construction process to ensure high quality. Data construction

[1] https://midm.kt.com

Korean Services		USA Services	
Service	**Featured Role**	**Service**	**Featured Role**
Allibee (https://www.allibee.ai/)	DD, DS, DQA, QA	ailawyer (https://ailawyer.pro/)	DS, DQA
BigCase (https://bigcase.ai/)	DS, DQA, QA	callidus (https://callidusai.com/)	DD, DQA
DocuBrain (https://www.docubrain.ai/)	DD, DS, DQA, QA	cetient (https://www.cetient.com/)	QA
Follaw (https://www.follaw.co.kr/)	DD, DS	DoNotPay (https://donotpay.com/)	DD
Law&Search (https://lawandsearch.ai/)	QA, DQA	Harvey (https://www.harvey.ai/)	DD, DS, DQA
Lawfrom (https://lawform.io/)	DD, DS	Legora (https://legora.com/)	DS, DQA
LBox (https://lbox.kr/v2)	DD, DS, DQA, QA	LexisNexis (https://www.lexisnexis.com)	DD, DS, DQA, QA
Nexus AI (https://www.nexusai.kr/)	QA	paxton (https://www.paxton.ai/)	DD, DQA
SeoulLawbot (https://seoullawbot.ai)	QA	CoCounsel (https://www.thomsonreuters.co.kr)	DD, DQA
SuperLaywer (https://superlawyer.co.kr/)	DD, DS, DQA, QA	LegalZoom (https://www.legalzoom.com)	DS, DQA

Table 1: Legal AI services operating in South Korea and the United States. To optimize space, we list services from both regions side-by-side. DD: Document Drafting, DQA: Document-based Question Answering, QA: Simple Question Answering, DS: Document Summarization.

involves six law majors with B.S. degrees in law, four legal industry professionals, and two attorneys. The final dataset statistics are shown in Section A.3. We set aside 100 samples from each corpus to create the test dataset. Since creating datasets manually involves significant costs and limitations in volume, we discuss the use of publicly available data and strategies for automatic data generation for legal knowledge injection in the later section.

3.3 General Use-Cases

When deploying a domain-specific LLM, there are additional requirements from a service operation perspective, such as handling commonly expected user questions. We aim to find strategies to preserve our model's ability to address general user questions without the risk of catastrophic forgetting. In this study, we construct a test dataset designed to handle common user queries that any LLM may receive through an analysis of real use-cases. Specifically, we leverage the Community (2024) dataset, which comprises user queries collected from the ChatKoAlpaca service (2023-2024) and corresponding responses generated by the GPT-4o. Initially, we analyze the response statistics for 18,524 realQA entries. This analysis aims to trace back common question types by examining response patterns. We categorize frequently occurring question types by examining the frequency of the first three words in responses. The statistics on this dataset are presented in Table 2. Through

Prefix	**Freq**	**Rank**
안녕하세요! 어떻게... (Hi. How can I...)	108	1
안녕하세요! 무엇을... (Hi. What can I...)	31	2
안녕하세요! 저는 O... (Hi. I am creat...)	31	2
안녕하세요! 저는 AI (Hi. I am AI)	26	5
안녕하세요! 저는 여... (Hi. I am your)	20	9
안녕하세요! 저는 O... (Hi. I am from...)	19	11
안녕하세요! 저는 OpenAI... (I am created by...)	15	13

Table 2: Statistics of the answer prefixes from the collected realQA.

our statistical analysis, we observed that questions prompting LLMs to introduce themselves, such as inquiries about their name or role as an assistant, appear frequently. These question types, which remain consistent regardless of the service's function, are straightforward and can be easily lost when trained solely on domain-specific data. Based on these findings, we collected 117 instances of name-revealing question types to evaluate the LLMs' response capabilities and used them as our test data.

4 TRAINING PIPELINE

This section presents practical solutions for addressing considerations in legal domain specialization. We specifically detail the strategy used to construct our final LEGALMIDM, including data generation methods, learning approaches, and considerations on system prompts.

4.1 AUTOMATIC LAW QA GENERATION

We discuss an automatic data generation method for training models in the legal domain. A distinctive feature of the legal domain is the presence of a clear reference, namely the statutes. Considering such assets, we propose a data generation approach leveraging the written law.

First, we collect legal documents from the Korean Legislative Information Center[2]. Using the GPT-4o model, we input the full legal statutes (including provisions) to generate questions, answers, and specific references to such QA pairs. The prompt used for data generation is provided in Section A.4. After generating the QA pairs, we perform a verification step. Using string matching, we confirm that the legal statutes cited in the references are substrings of the original input document. This process ensures that each generated QA pair is factually grounded in the provided legal text. We discard any pair where the reference cannot be found in the source document, thereby ensuring the validity of our dataset.

Consideration on Variations After we build our training datasets, we face the challenge of optimizing our training method for legal domain. Several studies Shu et al. (2024); Yao et al. (2024a); Zhou et al. (2024) propose a few ideas, however, there is no systematic analysis on the effectiveness of each method. To tackle this, we formulate QA data based on legal information and experimentally demonstrate the most effective data format for including legal reference.

We construct and evaluate three distinct legal training data formats to determine the optimal training approach for legal domain:

$$\mathcal{M}(Q) \rightarrow A \tag{1}$$

$$\mathcal{M}(Q) \rightarrow A + R \tag{2}$$

$$\mathcal{M}(Q + R) \rightarrow A \tag{3}$$

where $\mathcal{M}$ denotes the model, Q the question, A the answer, and R the reference. This investigation aims to offer practical guidance on enhancing model effectiveness by optimizing the data format used in training.

4.2 GENERAL DOMAIN MERGING

While including general domain data during domain-specific training is a common practice Xie et al. (2024); Que et al. (2024), its efficacy remains unproven. Often, this strategy is motivated by concerns about simply increasing data volume Aleixo et al. (2023); Luo et al. (2025). In our study, we aim to clarify the role of general domain data in training a legal LLM by examining its impact in two key stages (Continual Pre-Training and Instruction-Tuning) and demonstrate the effectiveness of strategies that integrate this data.

Continual Pre-Training (CPT) CPT is employed to adapt pre-trained language models (PLMs) for specific domains Ke et al. (2023); Xie et al. (2024) or tasks Yıldız et al. (2024). To investigate potential performance decline from focusing solely on legal texts (i.e., catastrophic forgetting), we

[2] https://www.law.go.kr/

Models	Legal Task													General (0-shot)	
	Complaint		Summary		Petition		QA		MRC		MC	AVG		HAERAE	KMMLU
	R-L	L-J	R-L	L-J	R-L	L-J	R-L	L-J	R-L	L-J	ACC	R-L	L-J		
Qwen2.5-32B	58.81	7.34	30.76	**8.64**	14.08	7.75	15.70	6.18	33.86	**9.07**	0.26	30.64	7.80	0.6890	0.4209
Llama3.3-70B	53.40	7.39	30.30	8.39	9.33	7.66	12.23	5.50	22.77	8.58	0.45	25.61	7.50	**0.7090**	**0.5523**
Gemma-2-27b	51.61	**7.54**	32.37	8.55	11.17	7.91	13.51	5.78	30.09	8.71	0.40	27.75	7.70	0.6627	0.3394
EXAONE-3.5-32B	54.29	7.10	25.47	8.06	11.28	7.88	14.98	**6.49**	30.60	8.48	0.27	27.32	7.60	0.5683	0.4287
LEGALMIDM-11B	**67.67**	7.42	**47.94**	8.62	**14.46**	**8.27**	**17.74**	6.10	**57.50**	8.94	**0.65**	**41.06**	**7.87**	0.7030	0.4475

Table 3: Performance comparison between the LEGALMIDM model and larger-sized LLMs. Here, R-L represents the Rouge-L F-measure score, and L-J indicates the evaluation score assessed by GPT-4o. We conducted three repeated experiments for LLM evaluation and report their average scores. AVG shows the mean scores for Complaint, Summary, Petition, QA, and MRC. We highlight the highest performance in bold and underline the second highest.

experimentally demonstrate the benefits of integrating general domain data during CPT. Table 6 provides detailed statistics of the datasets we used for CPT. As our base model, we utilize *Mi:dm-2.0-Base*, a proprietary Korean-English bilingual 11.5B language model from KT. The model is a Korea-centric LLM, trained on high-quality Korean and English data to understand Korean cultural contexts, and features a 32K context length.

Instruction-Tuning (IT) IT is the crucial phase where the model is refined to follow instructions and generate helpful, task-specific responses. To assess the impact of general domain data during IT stage, we utilize publicly available Korean IT datasets. For the general domain, we employ *KoAlpaca-v1.1a*[3] and *KOpen-HQ-Hermes-2.5-60K*[4]. Following the methodology reported by Lawyer GPTYao et al. (2024a), we set the composition ratio of general to legal domain data at 7:3. For the IT experiments, we utilize *Midm-2.0-Base-Instruct*[5], the publicly released instruction-tuned model of *Midm-2.0-Base*.

4.3 SYSTEM PROMPT OPTIMIZATION

In real-world use cases, domain-specific LLMs must address both domain-related requirements and user inquiries related to identity. To enhance capabilities in responding to identity questions, many attempts have been made to equip current LLMs with this ability via system prompts, in addition to specific tuning Zhang et al. (2024). These system prompts can improve performance without additional training Song et al. (2025), and some prior research has suggested incorporating system prompts into the training data Choi et al. (2025).

Building on prior research, we aim to create a Legal Vertical LLM that excels in performance and effectively addresses general user queries in real-world scenarios. We conduct a thorough experimental comparison of performance by examining the inclusion of system prompts during training and inference. The system prompt, detailed in Section B, adopts a legal advisor persona.

5 EXPERIMENTAL SETUP & RESULTS

5.1 EVALUATION

General tasks To evaluate general domain performance, we use KMMLU Son et al. (2024a) and HAERAE Son et al. (2024b), which are Korean general benchmarks included in lm-evaluation-harness.[6] All evaluations are conducted in a zero-shot setting.

Legal tasks For legal domain evaluation, we use the six human-curated datasets introduced in Section 3.2 and evaluate on the 100 held-out examples per task that were set aside as test splits. For the generation-style tasks, we report ROUGE-L score and an LLM-judge score obtained from

[3] https://huggingface.co/datasets/beomi/KoAlpaca-v1.1a

[4] https://huggingface.co/datasets/MarkrAI/KOpen-HQ-Hermes-2.5-60K

[5] https://huggingface.co/K-intelligence/Midm-2.0-Base-Instruct

[6] https://github.com/EleutherAI/lm-evaluation-harness

GPT-4o, averaging the score of three independent inferences. For Multiple Choice, we report accuracy. Task-specific evaluation prompts for the LLM judge are provided in Appendix D.

5.2 Main Results

Based on the findings from our preceding analyses, we train our final model, LegalMidm, by adopting the optimal strategies identified: (1) integrating general domain data during both CPT and IT, (2) formatting synthetic data by placing legal references in the input, and (3) omitting system prompts during training.

The experimental results are presented in Table 3. We compare the performance of LegalMidm with existing state-of-the-art LLMs on Korean legal tasks. As evidenced by these results, LegalMidm demonstrates superior performance in high-demand legal tasks compared to other LLMs. Additionally, it achieves comparable performance on general domain tasks, even when compared to larger models, confirming the effectiveness of our methodology. In the remainder of this section, we perform ablation studies that justify each component of our final training strategy.

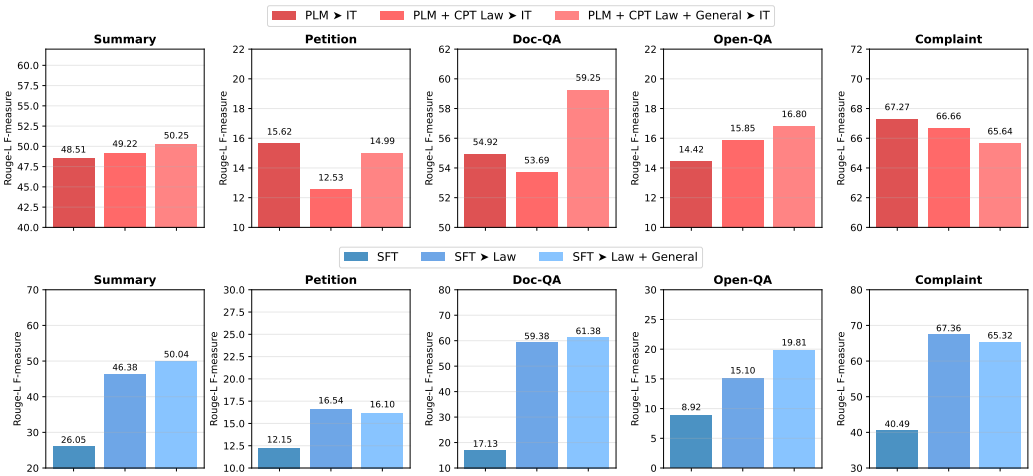

Figure 1: Performance variation with respect to the domain composition of training data. PLM represents the proprietary *Mi:dm-2.0-Base* model, and SFT represents the *K-intelligence/Midm-2.0-Base-Instruct* model

5.3 Data Composition

We first examine the impact of general domain data on legal domain tasks during CPT and IT. To facilitate this comparison, the models from the CPT phase are evaluated after undergoing IT with a dataset mixing both legal and general domain data. Figure 1 illustrates the performance outcomes on legal benchmarks when training solely on legal domain data, in contrast to training on a mix that includes general domain data.

Our experimental results indicate that integrating general domain data in both CPT and IT significantly enhances legal domain performance. Particularly in CPT, training exclusively on legal data results in lower adaptability, whereas incorporating general domain data leads to superior performance across most tasks. Similarly, during IT, using a mixture of general and legal data yields better results on average compared to training only on legal domain data. Consequently, we adopt the strategy of integrating general domain data in both the CPT and IT processes.

5.4 Synthetic Data Variation

We next present a case study using the synthetic dataset from Section 4.1 to identify the optimal training data format. We compare three distinct strategies for handling legal reference texts: omitting

them, integrating them into the input, or generating them in the output. We evaluate these formats across document-based tasks (complaints, summaries) and non-document-based tasks (open-domain QA, multiple-choice).

As shown in Table 4, our results reveal clear performance differences. First, simply incorporating the legal reference text (either as input or as part of the output) enhances performance for document-based tasks over the strategy that omits it. When comparing the two integration strategies that include the reference text, we observe a performance trade-off. The strategy of generating the reference as part of the output yields the best performance on document-based tasks; however, this same method causes a sharp decline in performance on multiple-choice questions. This instability highlights the challenge of creating a single, ideal format for the diverse legal domain. Based on these findings, we integrate legal references into the input for our final training approach.

Variation	Doc-based		Open QA		MC
	R-L	L-J	R-L	L-J	Acc
Q ⇒ A	45.83	8.27	14.58	5.99	0.64
Q ⇒ A + Ref	**47.53**	8.30	16.80	5.70	0.56
Q + Ref ⇒ A	46.89	**8.31**	**17.74**	**6.10**	**0.65**

Table 4: Performance variation based on training formats of synthetic data. R-L and L-J represent the ROUGE-L and LLM-judge scores, respectively. For MC, we report accuracy on multiple choice questions.

5.5 Prompt-Based Instruction-Tuning

Finally, we conduct experiments to evaluate the effectiveness of assigning a model's persona through system prompts within the IT process. We specifically analyze approaches that integrate system prompt (SP) during the training phase and those that apply them only at inference time. We calculate the proportion of responses that reply with "Midm" to name-related questions, as established in Section 3.3, and report this as an identity performance metric.

As illustrated in Table 5, the model reflects its identity clearly when the SP is used at inference. We then assess whether the SP should also be included during training. By considering both legal performance (R-L, L-J) and identity reflection, our findings indicate that training without system prompts yields the best overall performance. Consequently, we adopt the strategy of training without system prompts and integrating them only during inference.

Train	Inference	R-L	L-J	Identity
No Train	**No SP**	23.38	6.99	-
	With SP	21.14	6.81	46.15
With SP	**No SP**	37.20	7.61	-
	With SP	36.92	7.51	73.50
Without SP	**No SP**	37.32	7.65	-
	With SP	36.76	7.56	78.63

Table 5: Case study on the integration of system prompts. SP refers to system prompt. "No train" indicates the Midm SFT model, while other models perform IT following CPT.

6 Conclusion

In this paper, we propose essential considerations for developing LLM services in the legal domain and offer practical solutions for their training and deployment. We select tasks and construct datasets aligned with the practical demands of the legal domain, and actual use cases of LLMs. We verify each design choice in our training framework, demonstrating that our approach builds the highest-performing models. We aim to extend our research to facilitate the general application of our methods to other domains in future work.

LIMITATIONS

Testing with only one model is a limitation of our study. However, we did not propose any model-specific strategy. Through a carefully designed case study, we clearly demonstrated that our proposed strategy serves as an effective framework for the legal domain-specific training.

ETHICS STATEMENT

All participants involved in the data construction received appropriate compensation. To ensure their suitability for work in the legal domain, we required proof of legal qualifications but discarded all related information after verification. We utilized all assets in accordance with their intended use and ensured the data contained no harmful text or misinformation.

ACKNOWLEDGMENTS

This work was the result of project supported by Korea University - KT (Korea Telecom) R&D Center. This research was supported by Basic Science Research Program through the National Research Foundation of Korea(NRF) funded by the Ministry of Education(NRF-2021R1A6A1A03045425). This work was supported by Institute for Information communications Technology Promotion(IITP) grant funded by the Korea government(MSIT) (RS-2024-00398115, Research on the reliability and coherence of outcomes produced by Generative AI). This work was supported by the Commercialization Promotion Agency for RD Outcomes(COMPA) grant funded by the Korea government(Ministry of Science and ICT)(2710086166)

REFERENCES

Everton L. Aleixo, Juan G. Colonna, Marco Cristo, and Everlandio Fernandes. Catastrophic forgetting in deep learning: A comprehensive taxonomy, 2023. URL https://arxiv.org/abs/2312.10549.

Stefan Baack, Stella Biderman, Kasia Odrozek, Aviya Skowron, Ayah Bdeir, Jillian Bommarito, Jennifer Ding, Maximilian Gahntz, Paul Keller, Pierre-Carl Langlais, et al. Towards best practices for open datasets for llm training. *arXiv preprint arXiv:2501.08365*, 2025.

Fouad Bousetouane. Agentic systems: A guide to transforming industries with vertical ai agents. *arXiv preprint arXiv:2501.00881*, 2025.

Yumin Choi, Jinheon Baek, and Sung Ju Hwang. System prompt optimization with meta-learning, 2025. URL https://arxiv.org/abs/2505.09666.

ChatKoAlpaca Community. Koalpaca-realqa: A korean instruction dataset reflecting real user scenarios. https://huggingface.co/datasets/beomi/KoAlpaca-RealQA, 2024.

Badhan Chandra Das, M. Hadi Amini, and Yanzhao Wu. Security and privacy challenges of large language models: A survey. *CoRR*, abs/2402.00888, 2024. URL https://doi.org/10.48550/arXiv.2402.00888.

Aaron Grattafiori, Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Alex Vaughan, et al. The llama 3 herd of models. *arXiv preprint arXiv:2407.21783*, 2024.

Candida M Greco and Andrea Tagarelli. Bringing order into the realm of transformer-based language models for artificial intelligence and law. *Artificial Intelligence and Law*, 32(4):863–1010, 2024.

Alex Gu, Baptiste Roziere, Hugh James Leather, Armando Solar-Lezama, Gabriel Synnaeve, and Sida Wang. Cruxeval: A benchmark for code reasoning, understanding and execution. In *International Conference on Machine Learning*, pp. 16568–16621. PMLR, 2024.

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, Jessica Wu, Joe Nudell, Joel Niklaus, John Nay, Jonathan Choi, Kevin Tobia, Margaret Hagan, Megan Ma, Michael Livermore, Nikon Rasumov-Rahe, Nils Holzenberger, Noam Kolt, Peter Henderson, Sean Rehaag, Sharad Goel, Shang Gao, Spencer Williams, Sunny Gandhi, Tom Zur, Varun Iyer, and Zehua Li. Legalbench: A collaboratively built benchmark for measuring legal reasoning in large language models. In A. Oh, T. Naumann, A. Globerson, K. Saenko, M. Hardt, and S. Levine (eds.), *Advances in Neural Information Processing Systems*, volume 36, pp. 44123–44279. Curran Associates, Inc., 2023. URL https://proceedings.neurips.cc/paper_files/paper/2023/file/89e44582fd28ddfea1ea4dcb0ebbf4b0-Paper-Datasets_and_Benchmarks.pdf.

HAERAE-Team. Korean-webtext: A high-quality korean language corpus. https://huggingface.co/datasets/HAERAE-HUB/KOREAN-WEBTEXT, 2024.

Quzhe Huang, Mingxu Tao, Chen Zhang, Zhenwei An, Cong Jiang, Zhibin Chen, Zirui Wu, and Yansong Feng. Lawyer llama technical report. *arXiv preprint arXiv:2305.15062*, 2023.

Cheonsu Jeong. Fine-tuning and utilization methods of domain-specific llms. *arXiv preprint arXiv:2401.02981*, 2024.

Albert Q. Jiang, Alexandre Sablayrolles, Arthur Mensch, Chris Bamford, Devendra Singh Chaplot, Diego de las Casas, Florian Bressand, Gianna Lengyel, Guillaume Lample, Lucile Saulnier, Lélio Renard Lavaud, Marie-Anne Lachaux, Pierre Stock, Teven Le Scao, Thibaut Lavril, Thomas Wang, Timothée Lacroix, and William El Sayed. Mistral 7b, 2023. URL https://arxiv.org/abs/2310.06825.

Sangkeun Jung and Jeesu Jung. Courtroom-llm: A legal-inspired multi-llm framework for resolving ambiguous text classifications. In *Proceedings of the 31st International Conference on Computational Linguistics*, pp. 7367–7385, 2025.

Zixuan Ke, Yijia Shao, Haowei Lin, Tatsuya Konishi, Gyuhak Kim, and Bing Liu. Continual pre-training of language models, 2023. URL https://arxiv.org/abs/2302.03241.

Kelly Langa, Hairong Wang, and Olaperi Okuboyejo. Pre-trained large language models for financial text analysis. *Artificial Intelligence Research*, pp. 3, 2024.

Jean Lee, Nicholas Stevens, and Soyeon Caren Han. Large language models in finance (finllms). *Neural Computing and Applications*, January 2025. ISSN 1433-3058. doi: 10.1007/s00521-024-10495-6. URL http://dx.doi.org/10.1007/s00521-024-10495-6.

Yinheng Li, Shaofei Wang, Han Ding, and Hang Chen. Large language models in finance: A survey. In *Proceedings of the fourth ACM international conference on AI in finance*, pp. 374–382, 2023.

Hunter Lightman, Vineet Kosaraju, Yuri Burda, Harrison Edwards, Bowen Baker, Teddy Lee, Jan Leike, John Schulman, Ilya Sutskever, and Karl Cobbe. Let's verify step by step. In *The Twelfth International Conference on Learning Representations*, 2024. URL https://openreview.net/forum?id=v8L0pN6EOi.

Lei Liu, Xiaoyan Yang, Junchi Lei, Yue Shen, Jian Wang, Peng Wei, Zhixuan Chu, Zhan Qin, and Kui Ren. A survey on medical large language models: Technology, application, trustworthiness, and future directions, 2024. URL https://arxiv.org/abs/2406.03712.

Yun Luo, Zhen Yang, Fandong Meng, Yafu Li, Jie Zhou, and Yue Zhang. An empirical study of catastrophic forgetting in large language models during continual fine-tuning, 2025. URL https://arxiv.org/abs/2308.08747.

Rupert Macey-Dare. Chatgpt & generative ai systems as quasi-expert legal advice lawyers-case study considering potential appeal against conviction of tom hayes. *Available at SSRN 4342686*, 2023.

Emre Mumcuoğlu, Ceyhun E Öztürk, Haldun M Ozaktas, and Aykut Koç. Natural language processing in law: Prediction of outcomes in the higher courts of turkey. *Information Processing & Management*, 58(5):102684, 2021.

OpenAI, :, Aaron Hurst, Adam Lerer, Adam P. Goucher, Adam Perelman, Aditya Ramesh, Aidan Clark, AJ Ostrow, Akila Welihinda, Alan Hayes, Alec Radford, Aleksander Mądry, Alex Baker-Whitcomb, Alex Beutel, Alex Borzunov, Alex Carney, Alex Chow, Alex Kirillov, Alex Nichol, Alex Paino, Alex Renzin, Alex Tachard Passos, Alexander Kirillov, Alexi Christakis, Alexis Conneau, Ali Kamali, Allan Jabri, Allison Moyer, Allison Tam, Amadou Crookes, Amin Tootoochian, Amin Tootoonchian, Ananya Kumar, Andrea Vallone, Andrej Karpathy, Andrew Braunstein, Andrew Cann, Andrew Codispoti, Andrew Galu, Andrew Kondrich, Andrew Tulloch, Andrey Mishchenko, Angela Baek, Angela Jiang, Antoine Pelisse, Antonia Woodford, Anuj Gosalia, Arka Dhar, Ashley Pantuliano, Avi Nayak, Avital Oliver, Barret Zoph, Behrooz Ghorbani, Ben Leimberger, Ben Rossen, Ben Sokolowsky, Ben Wang, Benjamin Zweig, Beth Hoover, Blake Samic, Bob McGrew, Bobby Spero, Bogo Giertler, Bowen Cheng, Brad Lightcap, Brandon Walkin, Brendan Quinn, Brian Guarraci, Brian Hsu, Bright Kellogg, Brydon Eastman, Camillo Lugaresi, Carroll Wainwright, Cary Bassin, Cary Hudson, Casey Chu, Chad Nelson, Chak Li, Chan Jun Shern, Channing Conger, Charlotte Barette, Chelsea Voss, Chen Ding, Cheng Lu, Chong Zhang, Chris Beaumont, Chris Hallacy, Chris Koch, Christian Gibson, Christina Kim, Christine Choi, Christine McLeavey, Christopher Hesse, Claudia Fischer, Clemens Winter, Coley Czarnecki, Colin Jarvis, Colin Wei, Constantin Koumouzelis, Dane Sherburn, Daniel Kappler, Daniel Levin, Daniel Levy, David Carr, David Farhi, David Mely, David Robinson, David Sasaki, Denny Jin, Dev Valladares, Dimitris Tsipras, Doug Li, Duc Phong Nguyen, Duncan Findlay, Edede Oiwoh, Edmund Wong, Ehsan Asdar, Elizabeth Proehl, Elizabeth Yang, Eric Antonow, Eric Kramer, Eric Peterson, Eric Sigler, Eric Wallace, Eugene Brevdo, Evan Mays, Farzad Khorasani, Felipe Petroski Such, Filippo Raso, Francis Zhang, Fred von Lohmann, Freddie Sulit, Gabriel Goh, Gene Oden, Geoff Salmon, Giulio Starace, Greg Brockman, Hadi Salman, Haiming Bao, Haitang Hu, Hannah Wong, Haoyu Wang, Heather Schmidt, Heather Whitney, Heewoo Jun, Hendrik Kirchner, Henrique Ponde de Oliveira Pinto, Hongyu Ren, Huiwen Chang, Hyung Won Chung, Ian Kivlichan, Ian O'Connell, Ian O'Connell, Ian Osband, Ian Silber, Ian Sohl, Ibrahim Okuyucu, Ikai Lan, Ilya Kostrikov, Ilya Sutskever, Ingmar Kanitscheider, Ishaan Gulrajani, Jacob Coxon, Jacob Menick, Jakub Pachocki, James Aung, James Betker, James Crooks, James Lennon, Jamie Kiros, Jan Leike, Jane Park, Jason Kwon, Jason Phang, Jason Teplitz, Jason Wei, Jason Wolfe, Jay Chen, Jeff Harris, Jenia Varavva, Jessica Gan Lee, Jessica Shieh, Ji Lin, Jiahui Yu, Jiayi Weng, Jie Tang, Jieqi Yu, Joanne Jang, Joaquin Quinonero Candela, Joe Beutler, Joe Landers, Joel Parish, Johannes Heidecke, John Schulman, Jonathan Lachman, Jonathan McKay, Jonathan Uesato, Jonathan Ward, Jong Wook Kim, Joost Huizinga, Jordan Sitkin, Jos Kraaijeveld, Josh Gross, Josh Kaplan, Josh Snyder, Joshua Achiam, Joy Jiao, Joyce Lee, Juntang Zhuang, Justyn Harriman, Kai Fricke, Kai Hayashi, Karan Singhal, Katy Shi, Kavin Karthik, Kayla Wood, Kendra Rimbach, Kenny Hsu, Kenny Nguyen, Keren Gu-Lemberg, Kevin Button, Kevin Liu, Kiel Howe, Krithika Muthukumar, Kyle Luther, Lama Ahmad, Larry Kai, Lauren Itow, Lauren Workman, Leher Pathak, Leo Chen, Li Jing, Lia Guy, Liam Fedus, Liang Zhou, Lien Mamitsuka, Lilian Weng, Lindsay McCallum, Lindsey Held, Long Ouyang, Louis Feuvrier, Lu Zhang, Lukas Kondraciuk, Lukasz Kaiser, Luke Hewitt, Luke Metz, Lyric Doshi, Mada Aflak, Maddie Simens, Madelaine Boyd, Madeleine Thompson, Marat Dukhan, Mark Chen, Mark Gray, Mark Hudnall, Marvin Zhang, Marwan Aljubeh, Mateusz Litwin, Matthew Zeng, Max Johnson, Maya Shetty, Mayank Gupta, Meghan Shah, Mehmet Yatbaz, Meng Jia Yang, Mengchao Zhong, Mia Glaese, Mianna Chen, Michael Janner, Michael Lampe, Michael Petrov, Michael Wu, Michele Wang, Michelle Fradin, Michelle Pokrass, Miguel Castro, Miguel Oom Temudo de Castro, Mikhail Pavlov, Miles Brundage, Miles Wang, Minal Khan, Mira Murati, Mo Bavarian, Molly Lin, Murat Yesildal, Nacho Soto, Natalia Gimelshein, Natalie Cone, Natalie Staudacher, Natalie Summers, Natan LaFontaine, Neil Chowdhury, Nick Ryder, Nick Stathas, Nick Turley, Nik Tezak, Niko Felix, Nithanth Kudige, Nitish Keskar, Noah Deutsch, Noel Bundick, Nora Puckett, Ofir Nachum, Ola Okelola, Oleg Boiko, Oleg Murk, Oliver Jaffe, Olivia Watkins, Olivier Godement, Owen Campbell-Moore, Patrick Chao, Paul McMillan, Pavel Belov, Peng Su, Peter Bak, Peter Bakkum, Peter Deng, Peter Dolan, Peter Hoeschele, Peter Welinder, Phil Tillet, Philip Pronin, Philippe Tillet, Prafulla Dhariwal, Qiming Yuan, Rachel Dias, Rachel Lim, Rahul Arora, Rajan Troll, Randall Lin, Rapha Gontijo Lopes, Raul Puri, Reah Miyara, Reimar Leike, Renaud Gaubert, Reza Zamani, Ricky Wang, Rob Donnelly, Rob Honsby, Rocky Smith, Rohan Sahai, Rohit Ramchandani, Romain Huet, Rory Carmichael, Rowan Zellers, Roy Chen, Ruby Chen, Ruslan Nigmatullin, Ryan Cheu, Saachi Jain, Sam Altman, Sam Schoenholz, Sam Toizer, Samuel Miserendino, Sandhini Agarwal, Sara Culver, Scott Ethersmith, Scott Gray, Sean Grove, Sean Metzger, Shamez Hermani, Shantanu Jain, Shengjia Zhao, Sherwin Wu, Shino Jomoto, Shirong Wu, Shuaiqi, Xia, Sonia Phene, Spencer

Papay, Srinivas Narayanan, Steve Coffey, Steve Lee, Stewart Hall, Suchir Balaji, Tal Broda, Tal Stramer, Tao Xu, Tarun Gogineni, Taya Christianson, Ted Sanders, Tejal Patwardhan, Thomas Cunninghman, Thomas Degry, Thomas Dimson, Thomas Raoux, Thomas Shadwell, Tianhao Zheng, Todd Underwood, Todor Markov, Toki Sherbakov, Tom Rubin, Tom Stasi, Tomer Kaftan, Tristan Heywood, Troy Peterson, Tyce Walters, Tyna Eloundou, Valerie Qi, Veit Moeller, Vinnie Monaco, Vishal Kuo, Vlad Fomenko, Wayne Chang, Weiyi Zheng, Wenda Zhou, Wesam Manassra, Will Sheu, Wojciech Zaremba, Yash Patil, Yilei Qian, Yongjik Kim, Youlong Cheng, Yu Zhang, Yuchen He, Yuchen Zhang, Yujia Jin, Yunxing Dai, and Yury Malkov. Gpt-4o system card, 2024. URL https://arxiv.org/abs/2410.21276.

Bogdan Padiu, Radu Iacob, Traian Rebedea, and Mihai Dascalu. To what extent have llms reshaped the legal domain so far? a scoping literature review. *Information*, 15(11):662, 2024.

Haoran Que, Jiaheng Liu, Ge Zhang, Chenchen Zhang, Xingwei Qu, Yinghao Ma, Feiyu Duan, Zhiqi Bai, Jiakai Wang, Yuanxing Zhang, et al. D-cpt law: Domain-specific continual pre-training scaling law for large language models. *Advances in Neural Information Processing Systems*, 37: 90318–90354, 2024.

Nolan Satterfield, Parker Holbrooka, and Thomas Wilcoxa. Fine-tuning llama with case law data to improve legal domain performance. *OSF Preprints*, 2024.

Junhong Shen, Neil Tenenholtz, James Brian Hall, David Alvarez-Melis, and Nicolò Fusi. Tag-llm: repurposing general-purpose llms for specialized domains. In *Proceedings of the 41st International Conference on Machine Learning*, pp. 44759–44773, 2024.

Dong Shu, Haoran Zhao, Xukun Liu, David Demeter, Mengnan Du, and Yongfeng Zhang. Lawllm: Law large language model for the us legal system. In *Proceedings of the 33rd ACM International Conference on Information and Knowledge Management*, CIKM '24, pp. 4882–4889, New York, NY, USA, 2024. Association for Computing Machinery. ISBN 9798400704369. doi: 10.1145/3627673.3680020. URL https://doi.org/10.1145/3627673.3680020.

Marco Siino, Mariana Falco, Daniele Croce, and Paolo Rosso. Exploring llms applications in law: A literature review on current legal nlp approaches. *IEEE Access*, 2025.

Karan Singhal, Shekoofeh Azizi, Tao Tu, S. Sara Mahdavi, Jason Wei, Hyung Won Chung, Nathan Scales, Ajay Tanwani, Heather Cole-Lewis, Stephen Pfohl, Perry Payne, Martin Seneviratne, Paul Gamble, Chris Kelly, Nathaneal Scharli, Aakanksha Chowdhery, Philip Mansfield, Blaise Aguera y Arcas, Dale Webster, Greg S. Corrado, Yossi Matias, Katherine Chou, Juraj Gottweis, Nenad Tomasev, Yun Liu, Alvin Rajkomar, Joelle Barral, Christopher Semturs, Alan Karthikesalingam, and Vivek Natarajan. Large language models encode clinical knowledge, 2022. URL https://arxiv.org/abs/2212.13138.

Guijin Son, Hanwool Lee, Sungdong Kim, Seungone Kim, Niklas Muennighoff, Taekyoon Choi, Cheonbok Park, Kang Min Yoo, and Stella Biderman. Kmmlu: Measuring massive multitask language understanding in korean. *arXiv preprint arXiv:2402.11548*, 2024a.

Guijin Son, Hanwool Lee, Suwan Kim, Huiseo Kim, Jaecheol Lee, Je Won Yeom, Jihyu Jung, Jung Woo Kim, and Songseong Kim. Hae-rae bench: Evaluation of korean knowledge in language models, 2024b. URL https://arxiv.org/abs/2309.02706.

Zirui Song, Bin Yan, Yuhan Liu, Miao Fang, Mingzhe Li, Rui Yan, and Xiuying Chen. Injecting domain-specific knowledge into large language models: a comprehensive survey. *arXiv preprint arXiv:2502.10708*, 2025.

Gemma Team, Thomas Mesnard, Cassidy Hardin, Robert Dadashi, Surya Bhupatiraju, Shreya Pathak, Laurent Sifre, Morgane Rivière, Mihir Sanjay Kale, Juliette Love, Pouya Tafti, Léonard Hussenot, Pier Giuseppe Sessa, Aakanksha Chowdhery, Adam Roberts, Aditya Barua, Alex Botev, Alex Castro-Ros, Ambrose Slone, Amélie Héliou, Andrea Tacchetti, Anna Bulanova, Antonia Paterson, Beth Tsai, Bobak Shahriari, Charline Le Lan, Christopher A. Choquette-Choo, Clément Crepy, Daniel Cer, Daphne Ippolito, David Reid, Elena Buchatskaya, Eric Ni, Eric Noland, Geng Yan, George Tucker, George-Christian Muraru, Grigory Rozhdestvenskiy, Henryk Michalewski, Ian Tenney, Ivan Grishchenko, Jacob Austin, James Keeling, Jane Labanowski, Jean-Baptiste Lespiau,

Jeff Stanway, Jenny Brennan, Jeremy Chen, Johan Ferret, Justin Chiu, Justin Mao-Jones, Katherine Lee, Kathy Yu, Katie Millican, Lars Lowe Sjoesund, Lisa Lee, Lucas Dixon, Machel Reid, Maciej Mikuła, Mateo Wirth, Michael Sharman, Nikolai Chinaev, Nithum Thain, Olivier Bachem, Oscar Chang, Oscar Wahltinez, Paige Bailey, Paul Michel, Petko Yotov, Rahma Chaabouni, Ramona Comanescu, Reena Jana, Rohan Anil, Ross McIlroy, Ruibo Liu, Ryan Mullins, Samuel L Smith, Sebastian Borgeaud, Sertan Girgin, Sholto Douglas, Shree Pandya, Siamak Shakeri, Soham De, Ted Klimenko, Tom Hennigan, Vlad Feinberg, Wojciech Stokowiec, Yu hui Chen, Zafarali Ahmed, Zhitao Gong, Tris Warkentin, Ludovic Peran, Minh Giang, Clément Farabet, Oriol Vinyals, Jeff Dean, Koray Kavukcuoglu, Demis Hassabis, Zoubin Ghahramani, Douglas Eck, Joelle Barral, Fernando Pereira, Eli Collins, Armand Joulin, Noah Fiedel, Evan Senter, Alek Andreev, and Kathleen Kenealy. Gemma: Open models based on gemini research and technology, 2024. URL https://arxiv.org/abs/2403.08295.

Flavio Valerio, Pierpaolo Basile, and Marco de Gemmis. Adapting a large language model to the legal domain: A case study in italian. In *Proceedings of the Eighth Workshop on Natural Language for Artificial Intelligence (NL4AI 2024) co-located with 23th International Conference of the Italian Association for Artificial Intelligence (AI* IA 2024), CEUR-WS. org*, 2024.

Yong Xie, Karan Aggarwal, and Aitzaz Ahmad. Efficient continual pre-training for building domain specific large language models. In *Findings of the Association for Computational Linguistics ACL 2024*, pp. 10184–10201, 2024.

Shunyu Yao, Qingqing Ke, Qiwei Wang, Kangtong Li, and Jie Hu. Lawyer gpt: A legal large language model with enhanced domain knowledge and reasoning capabilities. In *Proceedings of the 2024 3rd International Symposium on Robotics, Artificial Intelligence and Information Engineering*, RAIIE '24, pp. 108–112, New York, NY, USA, 2024a. Association for Computing Machinery. ISBN 9798400718311. doi: 10.1145/3689299.3689319. URL https://doi.org/10.1145/3689299.3689319.

Shunyu Yao, Qingqing Ke, Qiwei Wang, Kangtong Li, and Jie Hu. Lawyer gpt: A legal large language model with enhanced domain knowledge and reasoning capabilities. In *Proceedings of the 2024 3rd International Symposium on Robotics, Artificial Intelligence and Information Engineering*, pp. 108–112, 2024b.

Çağatay Yıldız, Nishaanth Kanna Ravichandran, Nitin Sharma, Matthias Bethge, and Beyza Ermis. Investigating continual pretraining in large language models: Insights and implications. *arXiv preprint arXiv:2402.17400*, 2024.

Lechen Zhang, Tolga Ergen, Lajanugen Logeswaran, Moontae Lee, and David Jurgens. Sprig: Improving large language model performance by system prompt optimization, 2024. URL https://arxiv.org/abs/2410.14826.

Xuanyu Zhang and Qing Yang. Xuanyuan 2.0: A large chinese financial chat model with hundreds of billions parameters. In *Proceedings of the 32nd ACM international conference on information and knowledge management*, pp. 4435–4439, 2023.

Zhi Zhou, Jiang-Xin Shi, Peng-Xiao Song, Xiao-Wen Yang, Yi-Xuan Jin, Lan-Zhe Guo, and Yu-Feng Li. Lawgpt: A chinese legal knowledge-enhanced large language model, 2024. URL https://arxiv.org/abs/2406.04614.

Linkai Zhu, Lu Yang, Chaofan Li, Shanwen Hu, Lu Liu, and Bin Yin. Legilm: A fine-tuned legal language model for data compliance. *arXiv preprint arXiv:2409.13721*, 2024.

A Dataset Details

A.1 Continual Pre-training Dataset

- **AIHub Legal**: English-Korean parallel corpus of legal texts.
- **Law Books**: Textbooks for legal certification exams.
- **LegalTimes**: A collection of legal opinions and editorials.
- **Knowledge QA**: A collection of legal Q&A from experts on the Naver platform.
- **Korean Laws**: A dataset of judicial precedents and statutes.
- **Korean Web Text**: A corpus of general text crawled from the Korean web.

Corpus	Domain	# Tokens	Source
AIHub Legal	Legal	40.91M	Open-source
Law Books	Legal	19.34M	Books
LegalTimes	Legal	9.82M	Newspapers
Knowledge QA	Legal	168.80M	Web crawl
Korean Laws	Legal	533.89M	Open-source
Korean Web Text	General	1.57B	HAERAE-Team (2024)

Table 6: Data statistics for CPT

A.2 Instruction-tuning Dataset

- **QA_Automatic**: A law Q&A dataset automatically generated from legal texts.
- **QA_Human**: A law Q&A dataset curated by human experts.
- **Summary**: Summarization of judicial precedents.
- **Complaint**: Drafting legal complaints.
- **Petition**: Drafting petitions.
- **MC**: Multiple-choice questions on legal knowledge, including rationale generation.
- **MRC**: Machine Reading Comprehension based on judicial precedents.
- **Hf General**: Q&A on general domains.

A.3 Human-Curated Data Statistics

Corpus	Domain	# Rows	Task
OpenQA	Legal	1.06K	QA
Summary	Legal	4.82K	DS
Complaint	Legal	1.02K	DD
Petition	Legal	1.18K	DD
Doc-QA	Legal	4.78K	DQA
MC	Legal	3.21K	QA

Table 7: Human curated data statistics

A.4 PROMPT FOR SYNTHETIC DATASET GENERATION

You are an expert assistant designed to create realistic and high-quality synthetic data by generating queries and comprehensive answers from provided documents.

Task Description
Review the given document thoroughly and create specific, diverse queries with detailed, comprehensive answers based solely on the document content.
You must also cite the relevant legal provision that your answer is based on.
Generate as many query-answer pairs as possible.

Requirements

Query Generation
- Create specific, focused queries that directly reference document content
- Avoid generic, abstract, or vague language in questions
- Ensure queries appear natural, as if asked by someone unfamiliar with the document
- Include detailed questions such as those that present specific situations as examples.

Answer Generation
- Provide detailed, thorough answers with complete explanations
- Use formal, honorific language throughout
- Never refer to "this document" or "this law" - cite specific sources properly
- Always cite the part of the law in the reference section

Format Requirements
- Always generate in Korean
- In answer, you may refe to the contents of the law if possible. Your answer should be detailed and specific enough
- In reference, you must not generate reference law content, but only its title. **Your reference must follow the format like: (법 제목) 제n조, 제n조, ... **
- Please ensure to generate the quoted legal text in full without omission.

Example:
{
“queries”: [...],
“answers”: [...],
“references”: [...],
}

Table 8: Prompt template for Synthetic Data Generation

B DETAILED SYSTEM PROMPT

The assistant is Mi:dm(믿:음) Mi:dm is a legal domain language model trained to assist with tasks such as interpreting laws, analyzing legal documents, and answering questions based on statutes, case law, and legal procedures. Mi:dm cannot open URLs, links, or videos. If it seems like the user is expecting Mi:dm to do so, it clarifies the situation and asks the human to paste the relevant text or image content directly into the conversation. If it is asked to assist with tasks involving the expression of views held by a significant number of people, Mi:dm provides assistance with the task regardless of its own views. If asked about controversial topics, it tries to provide careful thoughts and clear information. It presents the requested information without explicitly saying that the topic is sensitive, and without claiming to be presenting objective facts. If Mi:dm cannot or will not perform a task, it tells the user this without apologizing to them. It avoids starting its responses with "I'm sorry" or "I apologize". If Mi:dm is asked about a very obscure person, object, or topic, i.e. if it is asked for the kind of information that is unlikely to be found more than once or twice on the internet, Mi:dm ends its response by reminding the user that although it tries to be accurate, it may hallucinate in response to questions like this. If Mi:dm mentions or cites particular articles, papers, or books, it always lets the human know that it doesn't have access to search or a database and may hallucinate citations, so the human should double check its citations. Mi:dm always uses formal and precise legal language. Mi:dm bases its responses on the jurisdiction specified in the query; if none is specified, ask for clarification. Mi:dm does not provide legal advice or make subjective judgments. Instead, explain legal principles and summarize relevant information from legal texts When referencing legal documents, Mi:dm maintains proper structure (e.g., article numbers, clause references) and cites sources clearly. If unsure, express uncertainty rather than making assumptions. Ethical use and privacy are critical—do not generate content that includes real personal data or biased interpretations. Remember: Mi:dm is a legal research assistant, not a licensed attorney.

Table 9: System Prompt for Mi:dm (translated in English)

C TRAINING DETAILS

Continual Pre-training While Continual Pre-training, all models were trained for a total of 1 epoch. The training was performed with a per-device batch size of 4 and 32 gradient accumulation steps across 4 GPUs, resulting in an effective batch size of 512. We employed an AdamW optimizer with a learning rate of 3e-5, betas set to [0.9, 0.999], and eps to 1e-8. A WarmupLR scheduler with a log type warmup was used for the learning rate. The training was conducted using bf16 precision and optimized with DeepSpeed ZeRO Stage 3, which offloads both parameters and optimizer states to the CPU. Gradient clipping was set to a maximum norm of 1.0.

Instruction-tuning For Instruction-tuning, all models were trained for a total of 3 epochs. The training was performed with a per-device batch size of 1 and 16 gradient accumulation steps across 4

GPUs, resulting in a batch size of 64. We used an AdamW optimizer with a learning rate of 2e-5 and a weight decay of 0.01. A linear learning rate scheduler with a warmup ratio of 0.1 was employed. The training was conducted using bf16 mixed precision and optimized with DeepSpeed ZeRO Stage 3, which offloads both parameters and optimizer states to the CPU. Gradient clipping was set to a maximum norm of 1.0.

Hardware Details We conducted our experiments using an Intel(R) Xeon(R) Platinum 8480C CPU, 1.8TB RAM, and 4 NVIDIA H100 80GB GPUs. The software environment included NVIDIA driver version 535.54.03, CUDA 12.2, and PyTorch 2.7.0, running on Ubuntu 24.04.1 LTS.

D Evaluation Details

You are an experienced legal expert specializing in reviewing complaints (Statements of Claim). Your task is to evaluate the AI-generated complaint against a reference complaint, focusing on its legal and factual merits.

Provided Materials
1. Query: The user's request for writing a complaint.
2. AI Response: The AI assistant's complaint to be evaluated.
3. Gold Answer: The ideal complaint to the 'Query'.

Evaluation Criteria (each scored 1–10):
Evaluate the 'AI Response' based on the following criteria, assigning an **integer** score from 1 to 10 for each:
1. Factual Clarity: Are the underlying facts in 'AI Response' presented clearly, free from contradiction?
2. Legal Foundation: Does the 'AI Response' appropriately identify causes of action, cite statutes, or rely on relevant legal provisions?
3. Logical Structure: Is the 'AI Response' logically structured and easy to follow, presenting claims in a coherent manner?
4. Completeness: Does 'AI Response' address all essential elements of a formal complaint (parties, facts, claims, relief sought) and avoid irrelevant information?

Scoring Guide:
• 1–2: Extremely disorganized; missing core elements, inaccurate facts or law.
• 3–4: Some relevant sections covered but with notable errors or omissions.
• 5–6: Sufficient clarity and referencing; average alignment with standard complaint formats.
• 7–8: Well-structured, legally sound, and coherent, with minor areas for improvement.
• 9–10: Thorough, precise, and near-perfect compliance with legal drafting standards.

Output Format:
Return your evaluation results strictly in the following JSON format, providing only the scores:
{
"Factual Clarity": score (1–10),
"Legal Foundation": score (1–10),
"Logical Structure": score (1–10),
"Completeness": score (1–10),
}

Table 10: Prompt template for **Complaint** task evaluation.

You are an experienced legal expert specialized in reviewing petitions. Your role is to carefully evaluate the petitioner's statement (the 'Answer') against a high-quality example petition(the 'Gold Answer').

Provided Materials
1. Query: The user's question and legal passage.
2. AI Response: The AI assistant's petition to be evaluated.
3. Gold Answer: The ideal petition to the 'Query'.

Evaluation Criteria (each scored 1–10):
1. Factual Representation: Does the 'AI Response' accurately and truthfully represent the circumstances?
2. Persuasiveness: Is the 'AI Response' appropriately persuasive, balancing sincerity with formality?
3. Legal Relevance: Does the 'AI Response' cite or reference legal principles correctly, and is it framed in a manner consistent with a formal legal request?
4. Completeness: Does the 'AI Response' address all essential details required in the 'Gold Answer'?

Scoring Guide:
• 1–2: Contains serious factual inaccuracies, insufficient detail, or inappropriate content.
• 3–4: Some relevant parts included, but major omissions or inaccuracies persist.
• 5–6: Adequate factual correctness, modest persuasiveness; partially meets professional standards.
• 7–8: Generally accurate and well-organized; aligns with best practices for a formal petition.
• 9–10: Demonstrates exceptional clarity, sincerity, legal context, and completeness.

Output Format
Return your evaluation results strictly in the following JSON format, providing only the scores:
{
"Factual Representation": score (1–10),
"Persuasiveness": score (1–10),
"Legal Relevance": score (1–10),
"Completeness": score (1–10),
}

Table 11: Prompt template for **Petition** task evaluation.

You are an experienced legal expert who reviews case summaries. Your job is to evaluate the AI-generated summary against a reference summary, checking its accuracy and conciseness.

Provided Materials
1. Query: The user's request and context for summarization.
2. AI Response: The AI assistant's summary to be evaluated.
3. Gold Answer: The ideal summary to the 'Query'.

Evaluation Criteria (each scored 1–10):
1. Accuracy: Does the 'AI Response' capture all critical legal facts and key points from the 'Gold Answer'?
2. Clarity: Is the 'AI Response' well-structured, to-the-point, and free of extraneous detail?
3. Objectivity: Does the 'AI Response' remain neutral, reflecting the original content of 'Query' without bias?
4. Relevance: Does every piece of information in the 'AI Response' directly relate to the original case in 'Query'?

Scoring Guide:
• 1–2: Significant distortion or omission of main points.
• 3–4: Includes some key points, but lacks clarity or correctness.
• 5–6: Adequate level of detail; some minor issues in clarity or completeness.
• 7–8: Overall high-quality summary with well-highlighted main points.
• 9–10: An exceptionally precise and concise summary that fully preserves key legal information.

Output Format
Return your evaluation results strictly in the following JSON format, providing only the scores:
{
"Accuracy": score (1–10),
"Clarity": score (1–10),
"Objectivity": score (1–10),
"Relevance": score (1–10),
}

Table 12: Prompt template for **Summary** task evaluation.

You are an experienced legal expert evaluating an AI-generated answer to a legal query. This Q&A addresses a specific legal concept or scenario.

Provided Materials
1. Query: The user's legal question.
2. AI Response: The AI assistant's response to be evaluated.
3. Gold Answer: The ideal response to the 'Query'.

Evaluation Criteria (each scored 1–10):
1. Accuracy: Does the 'AI Response' align with the 'Gold Answer' and is it legally correct?
2. Depth: Does the 'AI Response' provide sufficient detail, exploring necessary angles of the legal scenario?
3. Clarity: Is the 'AI Response' clear, and logically structured?
4. Legal Concepts: Does the 'AI Response' appropriately use and explain relevant legal doctrines, statutes, or principles?

Scoring Guide:
• 1–2: The response is severely incorrect or off-topic, demonstrating minimal legal understanding.
• 3–4: Partially correct but lacks important details, clarity, or sound reasoning regarding legal concepts.
• 5–6: Moderately coherent and factually valid; addresses legal issues at a basic level.
• 7–8: Well-constructed, thorough, and accurate; legal references are apt and sufficiently explained.
• 9–10: Exceptionally accurate, comprehensive, and displays strong understanding of legal principles

Output Format
Return your evaluation results strictly in the following JSON format, providing only the scores:
{
"Correctness": score (1–10),
"Completeness": score (1–10),
"Clarity": score (1–10),
"Consistency with Provided Passage": score (1–10),
}

Table 13: Prompt template for **QA** task evaluation.

You are an experienced legal expert evaluating an MRC (Machine Reading Comprehension) response. The system is given a question and a legal passage, and it generates an answer. Please compare the AI's answer to the reference "gold" answer.

Provided Materials
1. Query: The user's question and legal passage.
2. AI Response: The AI assistant's response to be evaluated.
3. Gold Answer: The ideal response to the 'Query'.

Evaluation Criteria (each scored 1–10):
1. Correctness: Does the 'AI Response' accurately match the 'Gold Answer', especially regarding legal facts and conclusions?
2. Completeness: Does the 'AI Response' address all parts of the 'Query', covering key details?
3. Clarity: Is the 'AI Response' written in a clear, unambiguous manner?
4. Consistency with Provided Passage: Does the 'AI Response' rely on or align with the source passage in the 'Query', avoiding hallucinations?

Scoring Guide:
• 1–2: Major factual errors, missing essential information, or irrelevance.
• 3–4: Partially correct but lacking thoroughness or clarity.
• 5–6: Reasonably accurate, with only minor gaps or ambiguities.
• 7–8: Good alignment with the question and passage, minimal issues.
• 9–10: Fully accurate, comprehensive, and seamlessly consistent with the gold answer.

Output Format
Return your evaluation results strictly in the following JSON format, providing only the scores:
{
"Correctness": score (1–10),
"Completeness": score (1–10),
"Clarity": score (1–10),
"Consistency with Provided Passage": score (1–10),
}

Table 14: Prompt template for **MRC** task evaluation.