LRAGE: Legal Retrieval Augmented Generation Evaluation Tool

Minhu Park[1,*] **Hongseok Oh**[1,*] **Eunkyung Choi**[1] **Wonseok Hwang**[1,2,†]

[1] University of Seoul [2] LBOX

{alsgn2003, cxv0519, rmarud202, wonseok.hwang}@uos.ac.kr

Abstract

Recently, building retrieval-augmented generation (RAG) systems to enhance the capability of large language models (LLMs) has become a common practice. Especially in the legal domain, previous judicial decisions play a significant role under the doctrine of stare decisis which emphasizes the importance of making decisions based on (retrieved) prior documents. However, the overall performance of RAG system depends on many components: (1) retrieval corpora, (2) retrieval algorithms, (3) rerankers, (4) LLM backbones, and (5) evaluation metrics. Here we propose LRAGE, an open-source tool for holistic evaluation of RAG systems focusing on the legal domain. LRAGE provides GUI and CLI interfaces to facilitate seamless experiments and investigate how changes in the aforementioned five components affect the overall accuracy. We validated LRAGE using multilingual legal benches including Korean (KBL), English (LegalBench), and Chinese (LawBench) by demonstrating how the overall accuracy changes when varying the five components mentioned above. The source code is available at https://github.com/hoorangyee/LRAGE.

1 Introduction

Recently large language models (LLMs) have demonstrated remarkable performance across a wide range of tasks. However, in expert domains–where average users struggle to assess the accuracy of an LLMs' responses–their performance remains limited due to a tendency to hallucinate (Dahl et al., 2024; Magesh et al., 2024).

To address this limitation, it has become standard practice to employ Retrieval-Augmented Generation (RAG), which integrates LLMs with information retrieval techniques. Although RAG systems have proven effective, they still exhibit hallucinations (Magesh et al., 2024; Niu et al., 2024). This underscores the need for rigorous evaluation before introducing such systems to users, especially within expert domains.

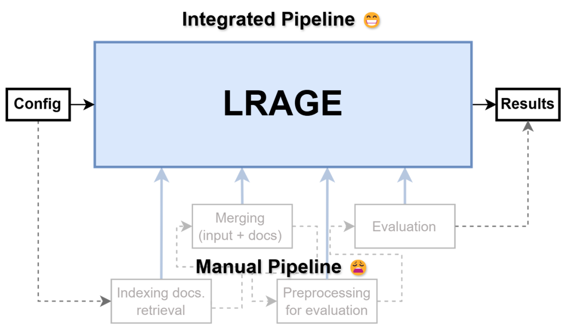

Figure 1: Comparison of conventional RAG evaluation pipeline (bottom) and the LRAGE framework (up) where each process is seamlessly integrated.

Evaluating LLMs on domain-specific benchmark datasets is thus essential for both research and industrial applications. To support this, several evaluation frameworks–such as Language Model Evaluation Harness (Gao et al., 2024b), Holistic Evaluation of Language Models (Liang et al., 2023)–have been developed and widely adopted by the research community.

Despite these developments, there remains a significant gap in the availability of comprehensive evaluation tools tailored for RAG pipelines, where multiple components influence overall accuracy: (1) retrieval corpus, (2) retrieval algorithms, (3) rerankers, (4) LLM backbones, and (5) evaluation metrics. For instance, Magesh et al. (2024) shows 40-50% of hallucinations can originate from the failure in document retrieval steps.

While there are existing tools (Rau et al., 2024; Zhang et al., 2024), their utility is often limited to general benchmark datasets and corpora, such as MMLU (Hendrycks et al., 2021) and Wikipedia while extending to other domain is not straightfor-

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

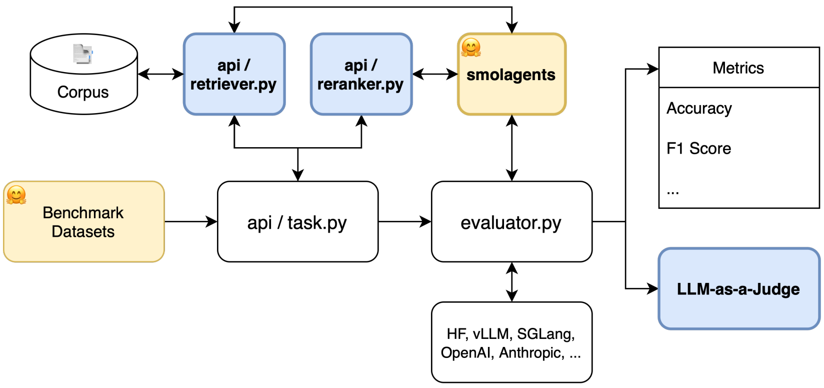

Figure 2: System diagram

ward. Additionally, domain experts may struggle to adapt these tools to their specific goals, as many do not provide a graphical interface (GUI).

Here we propose LRAGE[1], a holistic evaluation tool explicitly designed for assessing RAG systems in the legal domain. LRAGE extends Language Model Evaluation Harness (Gao et al., 2024b) by integrating it with `pyserini` (Lin et al., 2021) for information retrieval, while allowing easy control over individual components. Furthermore, LRAGE supports the use of legal-specific corpora, such as Pile-of-Law (Henderson et al., 2022), and benchmarks like LegalBench (Guha et al., 2023), in an off-the-shelf manner. By providing a user-friendly GUI, LRAGE not only streamlines the evaluation process for legal researchers working on RAG but also enables legal AI practitioners to efficiently assess their models using domain-specific data. In summary, our contributions are as follows.

- We propose LRAGE, an open-source evaluation tool for RAG systems that allows seamless integration of new corpus, tasks, and retrieval components.
- LRAGE features a user-friendly GUI, making it accessibility to domain experts.
- LRAGE provides pre-configured legal datasets for conducting RAG experiments with ease.

2 Related work

2.1 Legal Case Retrieval

Finding relevant previous cases is critical for legal decision-making (Feng et al., 2024). Accordingly, various studies have proposed models and datasets to address legal retrieval tasks (Goebel et al., 2023; Santosh et al., 2024; Hou et al., 2024; Li et al., 2023a,b; Gao et al., 2024a; Zheng et al., 2025). However, no comprehensive evaluation tools have been developed to specialized in how retrieval performance in legal RAG systems is influenced by the choice of (1) retrieval corpus, (2) retreiver, (3) backbone LLMs, (4) reranker, and (5) rubric.

2.2 RAG in legal domain

Magesh et al. (2024) analyzed commercial RAG systems in the U.S. legal domain using 202 examples, revealing that even the most competent system exhibited 17% hallucination rate. Niu et al. (2024) introduced RAGTruth benchmark, built using a subset of LegalBench (Guha et al., 2023). Their evaluation is limited to the retrieval tasks.

Zheng et al. (2025) developes two retrieval and RAG legal benchmarks: Bar Exam QA, and Housing Statute QA based on U.S. precedents and the statutory housing law. They show BM25 and current dense retrievers exhibit limited performance in recognizing gold passages in the legal domain. Notably, they built large scale ground truth passages labeled by law students and legal experts.

2.3 Legal Benchmarks for LLMs

This section briefly reviews legal benchmarks designed for evaluating LLMs. Guha et al. (2023) proposed LegalBench, a benchmark comprising 162 legal language understanding tasks. These tasks are organized according to six types of legal reasoning based on the IRAC framework. LegalBench focuses exclusively on English legal language understanding. Kim et al. (2024b) developed KBL, a benchmark dedicated to Korean legal language understanding. In addition to examples, they also

[1] Legal Retrieval Augmented Generation Evaluation tool

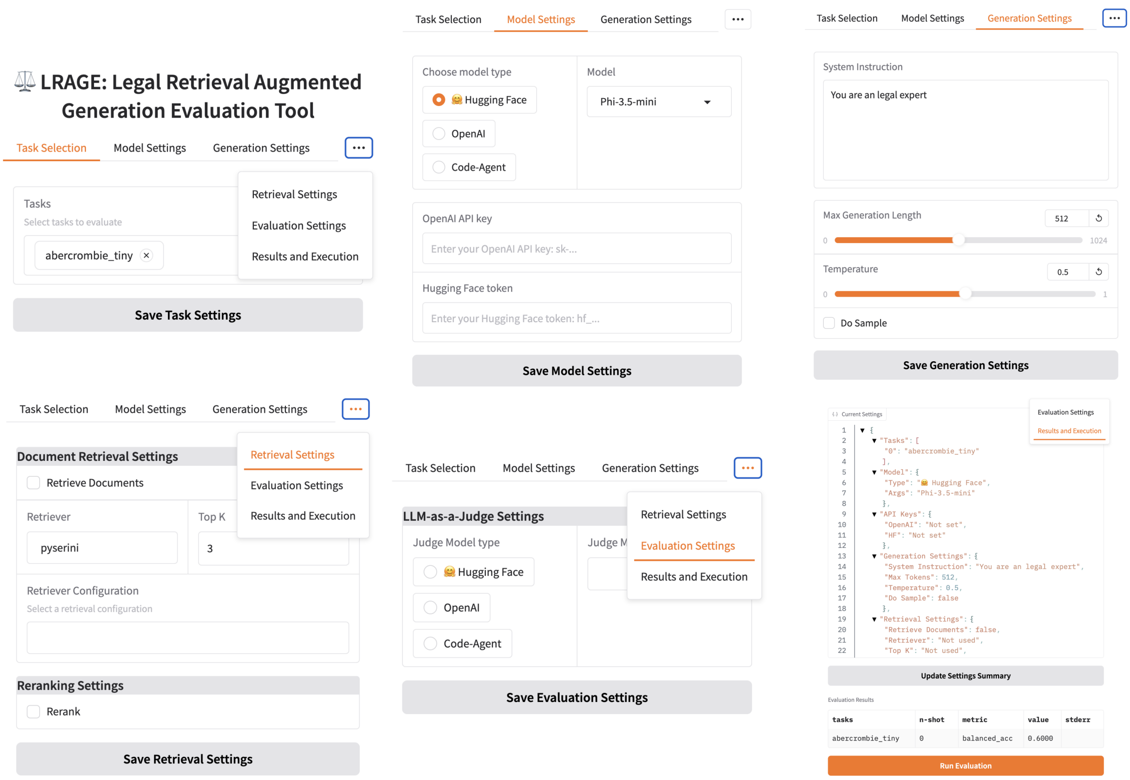

Figure 3: GUI of LRAGE. It consists of six tabs: Task (top-left), Model (top-center), Generation Parameters (top-right), Retriever (bottom-left), LLM-as-a-Judge (bottom-center), and a result tab(bottom-right). Each configuration tab allows users to define settings, which are then used in the final tab to perform experiments and immediately view the results.

provide resources for RAG experiments, including a corpus of Korean statutes and precedents Hwang et al. (2022). Fei et al. (2024) developed LawBench comprising 20 Chinese legal tasks categorized into three levels–Memorization, Understanding, and Applying–based on Bloom's taxonomy.

Except for KBL, these studies evaluated LLMs without incorporating RAG. Also, the KBL benchmark utilized only a basic RAG setup, employing a BM25 retriever without a reranker.

2.4 RAG Evaluation Tools

Rau et al. (2024) developed BERGEN, a tool designed for the systematic evaluation of RAG systems in question-answering (QA) tasks. BERGEN enables users to analyze the impact of individual system components, offering comprehensive support for various retrievers, rerankers, and language models. It employs an abstract class architecture with YAML configurations, allowing users to extend and customize components according to their requirements. Additionally, BERGEN supports multilingual capability by providing Wikipedia indices in 12 languages and offering multilingual versions of benchmark datasets.

Zhang et al. (2024) introduced RAGLAB, a RAG evaluation tool focused on the comparative analysis of different RAG algorithms rather than the individual pipeline components such as retrievers, rerankers, and LLMs. The framework provides six major RAG algorithms and incorporates ten QA benchmarks, using Wikipedia as the retrieval corpus. RAGLAB also supports advanced evaluation metrics, such as ALCE (Gao et al., 2023) and FactScore (Min et al., 2023), for assessing generative tasks. Moreover, the frameworks allows researchers to easily integrate new RAG algorithms.

Despite their strengths, these RAG evaluation frameworks have some notable limitations. First, these frameworks primarily rely on Wikipedia as the sole retrieval source. In domains where specialized knowledge and domain-specific documentation are critical (e.g., legal or medical fields), this reliance is a significant limitation, as current frameworks fail to adequately address specialized retrieval scenarios. Second, while BERGEN provides extensibility across various pipeline components, its flexibility comes at the cost of cumbersome setup process, often requiring complex code implementations and configuration files. This lack of a no-code evaluation environment limits its accessibility and ease of use, particularly for domain experts. In contrast, LRAGE is designed to address these gaps. It offers seamless integration with other retrieval corpora and tasks, a user-friendly GUI for easy accessibility by domain experts, and off-the-shelf legal-specific datasets, as detailed in the next section.

3 System

LRAGE is built on top of the open-source Language Model Evaluation Tool, `lm-evaluation-harness` (Gao et al., 2024b), by incorporating Retriever and Reranker modules for RAG. This allows us to inherit advantages such as extensibility to various task and models and flexible system instruction prompt tuning.

To support various retriever and reranker frameworks and models, LRAGE employs a modular architecture for these components (Fig. 2). The system follows SOLID design principles, particularly leveraging dependency injection to ensure loosely coupling between components, thereby enabling high modularity and extensibility. This modular design facilitates straightforward integration of new models and datasets without modifying the core architecture. It defines abstract classes that specify the essential operations for Retriever and Reranker in constructing the RAG pipeline, which can be implemented at either the framework or model level.

3.1 Retreiver and Reranker Modules

The Retriever module is currently implemented using `pyserini` (Lin et al., 2021), while the Reranker module utilizes `rerankers` (Clavié, 2024). Both frameworks are highly flexible and support various models. By modularizing these components at the framework level, LRAGE significantly reduces the implementation overhead typically required to support multiple models.

3.2 Metric Module

To support the evaluation of generative tasks, we extend the metric module of `lm-evaluation-harness` by integrating a custom LLM-as-a-Judge functionality. This enables flexible, rubric-based evaluation of legal benchmarks. By allowing rubrics to be defined at the instance level (Min et al., 2023; Kim et al., 2024a), LRAGE facilitates more detailed and precise evaluations. The users also can easily access the evaluation results through aggregated final scores while retaining the ability to review detailed instance-level rubric-based assessments via stored sample logs. The module leverages the existing LM class architecture of `lm-evaluation-harness`, ensuring compatibility with various frameworks and models that can serve as judges.

3.3 Legal domain specialization

LRAGE offers pre-configured settings for evaluating RAG systems in the legal domain. It currently supports various legal benchmarks such as KBL (Kim et al., 2024b), LegalBench (Guha et al., 2023), and LawBench (Fei et al., 2024). Beyond benchmark support, LRAGE provides preprocessed resources for legal copora such as Pile-of-Law, including a chunked version, a pre-compiled BM25 index, and a pre-compiled FAISS index (Douze et al., 2024), facilitating immediate use of legal datasets in RAG experiments. The demo video for our system is available at `https://github.com/hoorangyee/LRAGE`.

4 Experiments

We used Llama-3.1-8B (Meta, 2024), GPT-4 (OpenAI, 2023) and various other LLMs during our evaluations. The Pile-of-Law (Henderson et al., 2022) corpus was chunked in a similar manner to that described in (Hou et al., 2024). We converted the CAIL (Xiao et al., 2018) training set into a retrieval corpus by concatenating the fact section with metadata. For both CAIL and Korean precedents and statutes corpora (Hwang et al., 2022; Kim et al., 2024b), we treated each individual judgment as a single document for indexing, following the setup in previous work (Kim et al., 2024b). Unless otherwise specified, we used Llama-3.1-8B, BM25, and the top 3 retrieved documents as the default setting for RAG experiments.

5 Results

To demonstrate LRAGE, we measured the overall performance of RAG systems on legal Benchmarks while varying the following components: retrieval corpus, retrieval algorithm, LLM backbones, and reranker.

Retrieval corpus We first evaluate RAG performance on the multiple-choice questions from Ko-

Table 1: Evaluation result on 2024 Korean Bar Exam subtasks from KBL Benchmark (Kim et al., 2024b). Korean precedents and statues (Hwang et al., 2022; Kim et al., 2024b) (KoPS) or Korean wikipedia (kowiki) were used as the retrieval corpus. The values in parentheses indicate the difference (in percentage points) compared to the score obtained without RAG.

Acc (%, ↑)	civil	public	criminal
Llama-3.1-8B-chat			
w/o RAG	27.1	17.5	**27.5**
RAG w/ KoPS			
BM25	28.6 (+1.5)	35.0 (+17.5)	12.5 (-15.0)
+ ColBERT reranker	27.1 (+0.0)	30.0 (+12.5)	17.5 (-10.0)
+ T5 reranker	31.4 (+4.3)	35.0 (+17.5)	15.0 (-12.5)
+ Cross-Encoder reranker	**31.4 (+4.3)**	**40.0 (+22.5)**	17.5 (-10.0)
mE5-L Dense Retriever[a]	21.4 (-5.7)	27.5 (+10.0)	15.0 (-12.5)
bge-m3 Dense Retriever[b]	15.7 (-11.4)	27.5 (+10.0)	15.0 (-12.5)
RAG w/ kowiki			
BM25	27.1 (+0.0)	27.5 (+10.0)	15.0 (-12.5)
+ ColBERT reranker	27.1 (+0.0)	27.5 (+10.0)	17.5 (-10.0)
+ T5 reranker	31.4 (+4.3)	25.0 (+7.5)	15.0 (-12.5)
+ Cross-Encoder reranker	27.1 (+0.0)	17.5 (+0.0)	17.5 (-10.0)
GPT-4o			
w/o RAG	44.3	**57.5**	32.5
BM25 w/ KoPS	**57.1 (+12.8)**	55.0 (-2.5)	**50.0 (+17.5)**

a: Wang et al. (2024) *b*: Chen et al. (2024)

rean Bar Exam using Llama-3.1-8B and GPT-4o[2] with LRAGE. With Korean precedents and statues (KoPS) as the retrieval corpus, Llama shows improved performance compared to no RAG setting in `civil` and `public` subtasks (Table 1, 3rd vs. 5th–8th rows). In contrast, using kowiki corpus yields little to no improvement, or in some cases results in lower performance (3rd vs. 12th–15th rows). We also conducted RAG evaluation for LegalBench (Guha et al., 2023) (English), and LawBench (Fei et al., 2024) (Chinese). Table 2 demonstrates that, on three selected knowledge-intensive subtasks from LegalBench, RAG generally improves accuracy (see diagonal entries in 3rd–5th rows), although the effectiveness still depends on the choice of retrieval corpus. Similarly, evaluation on three subtasks from LawBench reveals corpus-dependent performance (Table 3). These results align with the intuition that selecting a task-specific corpus is crucial for effective RAG.

Retrieval algorithms In the Korean Bar Exam experiments, the dense retriever underperforms compared to BM25 (Table 1, 9th and 10th rows). This suggests that domain adaptation of dense retreivers is critical in the legal domain consistent with findings from recent studies Zheng et al. (2025); Hou et al. (2024).

[2] gpt-4o-2024-11-20

Table 2: LegalBench (Guha et al., 2023) evaluation result. We adopted wiki and the subsets of Pile of Law (PoL) (Henderson et al., 2022) for the retrieval corpus. PoL-cases includes "courtlistener_opinions", "tax_rulings", "canadian_decisions", and "echr". PoL-study incudes "cc_casebooks". The values in parentheses indicate the difference (in percentage points) compared to w/o RAG.

Acc (%, ↑)	international citizenship questions	nys judicial ethics	personal jurisdiction
w/o RAG	51.3	69.4	54.7
wiki	**59.4 (+8.1)**	68.7 (-0.7)	59.9 (+5.2)
PoL-cases	55.5 (+4.2)	**70.9 (+1.5)**	56.2 (-1.5)
PoL-study-materials	51.3 (+0.0)	69.9 (+0.5)	**66.1 (+11.4)**

Table 3: LawBench (Fei et al., 2024) evaluation result. Three knowledge-intensive subtasks were evaluated here. 1-2: Knowledge Question Answering; 3-3: Charge Prediction; 3-4: Preson Term Prediction w.o. Article. We adopted Chinese Wikipedia (zhwiki) and the CAIL (Xiao et al., 2018) train set for the retrieval corpus. The values in parentheses indicate the difference (in percentage points) compared to the score obtained without RAG.

LawBench	1-2 ACC (%, ↑)	3-3 F1 (%, ↑)	3-4 -log distance (↑)
w/o RAG	34.4	27.2	0.59
CAIL	**34.8 (+0.4)**	**41.2 (+14.0)**	**0.72 (+0.13)**
zhwiki	31.8 (-2.6)	22.8 (-4.4)	0.54 (-0.05)

Reranker Next, we examine how performance varies with the choice of rerankers. Evaluation on the *civil* and *public* subtasks shows that the cross-encoder reranker achieves the best performance in both cases (Table 1, 8th row). However, this trend does not hold for the *kowiki* corpus, where the T5-based reranker performs better on the *civil* subtask. This demonstrates that reranker effectiveness varies not only by task but also by corpus, highlighting the importance of the evaluation tools like LRAGE which allows seamless exploration of different RAG components combinations.

LLM backbones Interestingly, for `criminal` category, both KoPS and kowiki corpora show a significant drop in accuracy *criminal* task (Table 1, 3rd vs 5th–8th and 12th–15th columns). This contrasts with the improvements observed in stronger API model (final two rows), suggesting that the base capability of the model has a substantial impact on RAG performance, again emphasizing the importance of evaluation tools like LRAGE.

Further experiments with other models and sub-

Table 4: LLM-as-a-judge score on PLAT with different rubric settings. $S_{\text{sem,struc}}$ and S_{struc} stand for "semantic and structural rubrics" and "structural rubrics". The average scores from three independent experiments are shown. GPT-4o was used as the judge model.

Model	$S_{\text{sem,struc}}$	S_{struc}
GPT-4o-mini	4.16(±0.02)	4.41(±0.07)
GPT-4o	4.26(±0.06)	4.47(±0.07)

tasks are presented in Appendix, demonstrating that the performance of a RAG system in the legal domain depends on multiple interacting components.

Rubrics We demonstrate the LLM-as-a-judge functionality using PLAT (Choi et al., 2025), a Korean taxation benchmark.

We first convert 50 yes/no questions about the legitimacy of additional tax penalties from PLAT into descriptive questions. To prepare the rubrics, we use the reasoning section of Korean precedents and automatically convert them into rubrics using GPT-o1. After manually revising 10 examples in collaborating with a tax expert, we label the remaining 40 examples using GPT-o1 with a few-shot learning approach. More details will be provided in the paper currently in preparation.

The resulting PLAT rubrics consist of two types of items: (1) semantic and (2) structural. Semantic items evaluate the correctness of specific legal reasoning (e.g. "Is `[question-specific-article]` appropriately cited?"), while structural items focus on general aspects of writing (e.g. "Is the answer written concisely without unnecessary repetition?"). Each question includes four or five items, with a total possible score of 5 points. See Appendix for more examples.

To investigate how the scores depend on the choice of rubrics, we prepare a new set consisting only of structural items. Since structural understanding may not require deep legal knowledge or reasoning skills, less capable LLMs may achieve similar scores with more competent LLMs. The results show that on the original rubrics, GPT-4o-mini achieves -0.1 score compared to GPT-4o (Table 4, 1st column), whereas the gap narrows to -0.06 on the new structural rubrics (final column).

We conduct additional experiments using the BigLaw Bench core samples, an English legal task dataset.[3] The rubrics comprise sixty-four 1-point items and five 2-point items. We evaluate answers generated by GPT-4o-mini or GPT-4o using GPT-4o as a judge, which yields a mean score 5.63 ± 0.45 (from three independent experiments). When the point values are swapped–1-point items changed to 2-points and vice versa–the mean score adjusts to 5.80 ± 0.33. Although this indicate a rise in the average scores, it is difficult to draw a definitive conclusions due to the limited number of examples (five). Nevertheless, combined with the earlier experiment using PLAT, the result highlights the importance of supporting rubric modifications when evaluating free-form text.

[3] We used examples from `https://github.com/harveyai/biglaw-bench/blob/main/blb-core/core-samples.csv`.

Table 5: Evaluation results of Bar Exam QA (Zheng et al., 2025) with agentic RAG (Roucher et al., 2025).

Model	w/o RAG	w/ BM25	w/ Agentic RAG
Llama-3.1-8B-Instruct	47.0	41.0 (-6.0)	47.9 (+0.9)
Llama-3.1-70B-Instruct	76.1	68.4 (-7.7)	61.5 (-14.6)
GPT-4o-mini	48.7	51.3 (+2.6)	51.3 (+2.6)

Agentic RAG To support agentic RAG, LRAGE integrates `smolagents` (Roucher et al., 2025). For the demonstration, we use recent legal RAG benchmark from Zheng et al. (2025). The results show that the off-the-shelf application of agentic RAG does not necessarily improve performance (Table 5, 2nd vs 3rd columns), although stronger model shows relatively more competent results (1st vs 2nd rows). This suggests that, similar to retrievers, domain adaptation of agent components–such as prompts, tools, and reasoning frameworks (Kang et al., 2023; Anthropic, 2025)–may be necessary.

6 Conclusion

We propose LRAGE, a holistic evaluation tool for RAG systems specifically tailored for applications in the legal domain. Building on the widely adapted open-source LLM evaluation tools `lm-evaluation-harness`, LRAGE integrates two core functionalities–Retriever, Reranker–along with additional features for evaluating generative tasks using instance-level custom rubrics. Experiments on legal domain benchmakrs demonstrate how the overall performance of RAG systems depends on individual components, highlighting the effectiveness of LRAGE, which enables analysis with just a few lines of script or GUI. The inclusion of a user-friendly GUI and pre-processed legal corpora for retrieval facilitates seamless adaptation by legal domain experts, making LRAGE highly accessible and practical for specialized use cases.

References

Anthropic. 2025. The "think" tool: Enabling claude to stop and think in complex tool use situations. *Anthropic Engineering Blog*. Accessed: 2025-03-29.

Zheng Cai, Maosong Cao, Haojiong Chen, Kai Chen, Keyu Chen, Xin Chen, Xun Chen, Zehui Chen, Zhi Chen, Pei Chu, Xiaoyi Dong, Haodong Duan, Qi Fan, Zhaoye Fei, Yang Gao, Jiaye Ge, Chenya Gu, Yuzhe Gu, Tao Gui, Aijia Guo, Qipeng Guo, Conghui He, Yingfan Hu, Ting Huang, Tao Jiang, Penglong Jiao, Zhenjiang Jin, Zhikai Lei, Jiaxing Li, Jingwen Li, Linyang Li, Shuaibin Li, Wei Li, Yining Li, Hongwei Liu, Jiangning Liu, Jiawei Hong, Kaiwen Liu, Kuikun Liu, Xiaoran Liu, Chengqi Lv, Haijun Lv, Kai Lv, Li Ma, Runyuan Ma, Zerun Ma, Wenchang Ning, Linke Ouyang, Jiantao Qiu, Yuan Qu, Fukai Shang, Yunfan Shao, Demin Song, Zifan Song, Zhihao Sui, Peng Sun, Yu Sun, Huanze Tang, Bin Wang, Guoteng Wang, Jiaqi Wang, Jiayu Wang, Rui Wang, Yudong Wang, Ziyi Wang, Xingjian Wei, Qizhen Weng, Fan Wu, Yingtong Xiong, Chao Xu, Ruiliang Xu, Hang Yan, Yirong Yan, Xiaogui Yang, Haochen Ye, Huaiyuan Ying, Jia Yu, Jing Yu, Yuhang Zang, Chuyu Zhang, Li Zhang, Pan Zhang, Peng Zhang, Ruijie Zhang, Shuo Zhang, Songyang Zhang, Wenjian Zhang, Wenwei Zhang, Xingcheng Zhang, Xinyue Zhang, Hui Zhao, Qian Zhao, Xiaomeng Zhao, Fengzhe Zhou, Zaida Zhou, Jingming Zhuo, Yicheng Zou, Xipeng Qiu, Yu Qiao, and Dahua Lin. 2024. Internlm2 technical report. *Preprint*, arXiv:2403.17297.

Ilias Chalkidis, Manos Fergadiotis, Prodromos Malakasiotis, Nikolaos Aletras, and Ion Androutsopoulos. 2020. LEGAL-BERT: The muppets straight out of law school. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 2898–2904, Online. Association for Computational Linguistics.

Ilias Chalkidis*, Nicolas Garneau*, Catalina Goanta, Daniel Martin Katz, and Anders Søgaard. 2023. LeXFiles and LegalLAMA: Facilitating English Multinational Legal Language Model Development. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics*, Toronto, Canada. Association for Computational Linguistics.

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

Eunkyung Choi, Young Jin Suh, Hun Park, and Wonseok Hwang. 2025. Taxation perspectives from large language models: A case study on additional tax penalties. *Preprint*, arXiv:2503.03444.

Benjamin Clavié. 2024. rerankers: A lightweight python library to unify ranking methods. *Preprint*, arXiv:2408.17344.

Matthew Dahl, Varun Magesh, Mirac Suzgun, and Daniel E. Ho. 2024. Large legal fictions: Profiling legal hallucinations in large language models. *Preprint*, arXiv:2401.01301.

Matthijs Douze, Alexandr Guzhva, Chengqi Deng, Jeff Johnson, Gergely Szilvasy, Pierre-Emmanuel Mazaré, Maria Lomeli, Lucas Hosseini, and Hervé Jégou. 2024. The faiss library.

Zhiwei Fei, Xiaoyu Shen, Dawei Zhu, Fengzhe Zhou, Zhuo Han, Alan Huang, Songyang Zhang, Kai Chen, Zhixin Yin, Zongwen Shen, Jidong Ge, and Vincent Ng. 2024. LawBench: Benchmarking legal knowledge of large language models. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 7933–7962, Miami, Florida, USA. Association for Computational Linguistics.

Yi Feng, Chuanyi Li, and Vincent Ng. 2024. Legal case retrieval: A survey of the state of the art. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 6472–6485, Bangkok, Thailand. Association for Computational Linguistics.

Cheng Gao, Chaojun Xiao, Zhenghao Liu, Huimin Chen, Zhiyuan Liu, and Maosong Sun. 2024a. Enhancing legal case retrieval via scaling high-quality synthetic query-candidate pairs. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 7086–7100, Miami, Florida, USA. Association for Computational Linguistics.

Leo Gao, Jonathan Tow, Baber Abbasi, Stella Biderman, Sid Black, Anthony DiPofi, Charles Foster, Laurence Golding, Jeffrey Hsu, Alain Le Noac'h, Haonan Li, Kyle McDonell, Niklas Muennighoff, Chris Ociepa, Jason Phang, Laria Reynolds, Hailey Schoelkopf, Aviya Skowron, Lintang Sutawika, Eric Tang, Anish Thite, Ben Wang, Kevin Wang, and Andy Zou. 2024b. A framework for few-shot language model evaluation.

Tianyu Gao, Howard Yen, Jiatong Yu, and Danqi Chen. 2023. Enabling large language models to generate text with citations. In *Empirical Methods in Natural Language Processing (EMNLP)*.

Randy Goebel, Yoshinobu Kano, Mi-Young Kim, Juliano Rabelo, Ken Satoh, and Masaharu Yoshioka. 2023. Summary of the competition on legal information, extraction/entailment (coliee) 2023. In *Proceedings of the Nineteenth International Conference on Artificial Intelligence and Law*, ICAIL '23, page 472–480, New York, NY, USA. Association for Computing Machinery.

Neel Guha, Julian Nyarko, Daniel E. Ho, Christopher Ré, Adam Chilton, Aditya Narayana, Alex Chohlas-Wood, Austin Peters, Brandon Waldon, Daniel N. Rockmore, Diego Zambrano, Dmitry Talisman, Enam Hoque, Faiz Surani, Frank Fagan, Galit

Sarfaty, Gregory M. Dickinson, Haggai Porat, Jason Hegland, Jessica Wu, Joe Nudell, Joel Niklaus, John Nay, Jonathan H. 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. 2023. Legalbench: A collaboratively built benchmark for measuring legal reasoning in large language models. *Preprint*, arXiv:2308.11462.

Peter Henderson, Mark Simon Krass, Lucia Zheng, Neel Guha, Christopher D Manning, Dan Jurafsky, and Daniel E. Ho. 2022. Pile of law: Learning responsible data filtering from the law and a 256GB open-source legal dataset. In *Thirty-sixth Conference on Neural Information Processing Systems Datasets and Benchmarks Track*.

Dan Hendrycks, Collin Burns, Steven Basart, Andy Zou, Mantas Mazeika, Dawn Song, and Jacob Steinhardt. 2021. Measuring massive multitask language understanding. *Proceedings of the International Conference on Learning Representations (ICLR)*.

Abe Bohan Hou, Orion Weller, Guanghui Qin, Eugene Yang, Dawn Lawrie, Nils Holzenberger, Andrew Blair-Stanek, and Benjamin Van Durme. 2024. Clerc: A dataset for legal case retrieval and retrieval-augmented analysis generation. *Preprint*, arXiv:2406.17186.

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 *Thirty-sixth Conference on Neural Information Processing Systems Datasets and Benchmarks Track*.

Xiaoxi Kang, Lizhen Qu, Lay-Ki Soon, Adnan Trakic, Terry Zhuo, Patrick Emerton, and Genevieve Grant. 2023. Can ChatGPT perform reasoning using the IRAC method in analyzing legal scenarios like a lawyer? In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 13900–13923, Singapore. Association for Computational Linguistics.

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

Seungone Kim, Juyoung Suk, Ji Yong Cho, Shayne Longpre, Chaeeun Kim, Dongkeun Yoon, Guijin Son, Yejin Cho, Sheikh Shafayat, Jinheon Baek, Sue Hyun Park, Hyeonbin Hwang, Jinkyung Jo, Hyowon Cho, Haebin Shin, Seongyun Lee, Hanseok Oh, Noah Lee, Namgyu Ho, Se June Joo, Miyoung Ko, Yoonjoo Lee, Hyungjoo Chae, Jamin Shin, Joel Jang, Seonghyeon Ye, Bill Yuchen Lin, Sean Welleck, Graham Neubig, Moontae Lee, Kyungjae Lee, and Minjoon Seo. 2024a. The biggen bench: A principled benchmark for fine-grained evaluation of language models with language models. *Preprint*, arXiv:2406.05761.

Yeeun Kim, Youngrok Choi, Eunkyung Choi, JinHwan Choi, Hai Jin Park, and Wonseok Hwang. 2024b. Developing a pragmatic benchmark for assessing Korean legal language understanding in large language models. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 5573–5595, Miami, Florida, USA. Association for Computational Linguistics.

Haitao Li, Qingyao Ai, Jia Chen, Qian Dong, Yueyue Wu, Yiqun Liu, Chong Chen, and Qi Tian. 2023a. Sailer: Structure-aware pre-trained language model for legal case retrieval. In *Proceedings of the 46th International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '23, page 1035–1044, New York, NY, USA. Association for Computing Machinery.

Haitao Li, Yunqiu Shao, Yueyue Wu, Qingyao Ai, Yixiao Ma, and Yiqun Liu. 2023b. Lecardv2: A large-scale chinese legal case retrieval dataset. *Preprint*, arXiv:2310.17609.

Percy Liang, Rishi Bommasani, Tony Lee, Dimitris Tsipras, Dilara Soylu, Michihiro Yasunaga, Yian Zhang, Deepak Narayanan, Yuhuai Wu, Ananya Kumar, Benjamin Newman, Binhang Yuan, Bobby Yan, Ce Zhang, Christian Alexander Cosgrove, Christopher D Manning, Christopher Re, Diana Acosta-Navas, Drew Arad Hudson, Eric Zelikman, Esin Durmus, Faisal Ladhak, Frieda Rong, Hongyu Ren, Huaxiu Yao, Jue WANG, Keshav Santhanam, Laurel Orr, Lucia Zheng, Mert Yuksekgonul, Mirac Suzgun, Nathan Kim, Neel Guha, Niladri S. Chatterji, Omar Khattab, Peter Henderson, Qian Huang, Ryan Andrew Chi, Sang Michael Xie, Shibani Santurkar, Surya Ganguli, Tatsunori Hashimoto, Thomas Icard, Tianyi Zhang, Vishrav Chaudhary, William Wang, Xuechen Li, Yifan Mai, Yuhui Zhang, and Yuta Koreeda. 2023. Holistic evaluation of language models. *Transactions on Machine Learning Research*. Featured Certification, Expert Certification.

Jimmy Lin, Xueguang Ma, Sheng-Chieh Lin, Jheng-Hong Yang, Ronak Pradeep, and Rodrigo Nogueira. 2021. Pyserini: A Python toolkit for reproducible information retrieval research with sparse and dense representations. In *Proceedings of the 44th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR 2021)*, pages 2356–2362.

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

Meta. 2024. The llama 3 herd of models. *Preprint*, arXiv:2407.21783.

Sewon Min, Kalpesh Krishna, Xinxi Lyu, Mike Lewis, Wen-tau Yih, Pang Koh, Mohit Iyyer, Luke Zettlemoyer, and Hannaneh Hajishirzi. 2023. FActScore: Fine-grained atomic evaluation of factual precision in long form text generation. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 12076–12100, Singapore. Association for Computational Linguistics.

Cheng Niu, Yuanhao Wu, Juno Zhu, Siliang Xu, KaShun Shum, Randy Zhong, Juntong Song, and Tong Zhang. 2024. RAGTruth: A hallucination corpus for developing trustworthy retrieval-augmented language models. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 10862–10878, Bangkok, Thailand. Association for Computational Linguistics.

OpenAI. 2023. Gpt-4 technical report. *Preprint*, arXiv:2303.08774.

Nicholas Pipitone and Ghita Houir Alami. 2024. Legalbench-rag: A benchmark for retrieval-augmented generation in the legal domain. *Preprint*, arXiv:2408.10343.

Qwen Team. 2024. Qwen2.5: A party of foundation models.

David Rau, Hervé Déjean, Nadezhda Chirkova, Thibault Formal, Shuai Wang, Stéphane Clinchant, and Vassilina Nikoulina. 2024. BERGEN: A benchmarking library for retrieval-augmented generation. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 7640–7663, Miami, Florida, USA. Association for Computational Linguistics.

LG AI Research, :, Soyoung An, Kyunghoon Bae, Eunbi Choi, Stanley Jungkyu Choi, Yemuk Choi, Seokhee Hong, Yeonjung Hong, Junwon Hwang, Hyojin Jeon, Gerrard Jeongwon Jo, Hyunjik Jo, Jiyeon Jung, Yountae Jung, Euisoon Kim, Hyosang Kim, Joonkee Kim, Seonghwan Kim, Soyeon Kim, Sunkyoung Kim, Yireun Kim, Youchul Kim, Edward Hwayoung Lee, Haeju Lee, Honglak Lee, Jinsik Lee, Kyungmin Lee, Moontae Lee, Seungjun Lee, Woohyung Lim, Sangha Park, Sooyoun Park, Yongmin Park, Boseong Seo, Sihoon Yang, Heuiyeen Yeen, Kyungjae Yoo, and Hyeongu Yun. 2024. Exaone 3.0 7.8b instruction tuned language model. *Preprint*, arXiv:2408.03541.

Aymeric Roucher, Albert Villanova del Moral, Thomas Wolf, Leandro von Werra, and Erik Kaunismäki. 2025. ‘smolagents‘: a smol library to build great agentic systems. `https://github.com/huggingface/smolagents`.

T. Y. S. S Santosh, Rashid Gustav Haddad, and Matthias Grabmair. 2024. Ecthr-pcr: A dataset for precedent understanding and prior case retrieval in the european court of human rights. *Preprint*, arXiv:2404.00596.

Gemma Team, Aishwarya Kamath, Johan Ferret, Shreya Pathak, Nino Vieillard, Ramona Merhej, Sarah Perrin, Tatiana Matejovicova, Alexandre Ramé, Morgane Rivière, Louis Rouillard, Thomas Mesnard, Geoffrey Cideron, Jean bastien Grill, Sabela Ramos, Edouard Yvinec, Michelle Casbon, Etienne Pot, Ivo Penchev, Gaël Liu, Francesco Visin, Kathleen Kenealy, Lucas Beyer, Xiaohai Zhai, Anton Tsitsulin, Robert Busa-Fekete, Alex Feng, Noveen Sachdeva, Benjamin Coleman, Yi Gao, Basil Mustafa, Iain Barr, Emilio Parisotto, David Tian, Matan Eyal, Colin Cherry, Jan-Thorsten Peter, Danila Sinopalnikov, Surya Bhupatiraju, Rishabh Agarwal, Mehran Kazemi, Dan Malkin, Ravin Kumar, David Vilar, Idan Brusilovsky, Jiaming Luo, Andreas Steiner, Abe Friesen, Abhanshu Sharma, Abheesht Sharma, Adi Mayrav Gilady, Adrian Goedeckemeyer, Alaa Saade, Alex Feng, Alexander Kolesnikov, Alexei Bendebury, Alvin Abdagic, Amit Vadi, András György, André Susano Pinto, Anil Das, Ankur Bapna, Antoine Miech, Antoine Yang, Antonia Paterson, Ashish Shenoy, Ayan Chakrabarti, Bilal Piot, Bo Wu, Bobak Shahriari, Bryce Petrini, Charlie Chen, Charline Le Lan, Christopher A. Choquette-Choo, CJ Carey, Cormac Brick, Daniel Deutsch, Danielle Eisenbud, Dee Cattle, Derek Cheng, Dimitris Paparas, Divyashree Shivakumar Sreepathihalli, Doug Reid, Dustin Tran, Dustin Zelle, Eric Noland, Erwin Huizenga, Eugene Kharitonov, Frederick Liu, Gagik Amirkhanyan, Glenn Cameron, Hadi Hashemi, Hanna Klimczak-Plucińska, Harman Singh, Harsh Mehta, Harshal Tushar Lehri, Hussein Hazimeh, Ian Ballantyne, Idan Szpektor, Ivan Nardini, Jean Pouget-Abadie, Jetha Chan, Joe Stanton, John Wieting, Jonathan Lai, Jordi Orbay, Joseph Fernandez, Josh Newlan, Ju yeong Ji, Jyotinder Singh, Kat Black, Kathy Yu, Kevin Hui, Kiran Vodrahalli, Klaus Greff, Linhai Qiu, Marcella Valentine, Marina Coelho, Marvin Ritter, Matt Hoffman, Matthew Watson, Mayank Chaturvedi, Michael Moynihan, Min Ma, Nabila Babar, Natasha Noy, Nathan Byrd, Nick Roy, Nikola Momchev, Nilay Chauhan, Noveen Sachdeva, Oskar Bunyan, Pankil Botarda, Paul Caron, Paul Kishan Rubenstein, Phil Culliton, Philipp Schmid, Pier Giuseppe Sessa, Pingmei Xu, Piotr Stanczyk, Pouya Tafti, Rakesh Shivanna, Renjie Wu, Renke Pan, Reza Rokni, Rob Willoughby, Rohith Vallu, Ryan Mullins, Sammy Jerome, Sara Smoot, Sertan Girgin, Shariq Iqbal, Shashir Reddy, Shruti Sheth, Siim Põder, Sijal Bhatnagar, Sindhu Raghuram Panyam, Sivan Eiger, Susan Zhang, Tianqi Liu, Trevor Yacovone, Tyler Liechty, Uday Kalra, Utku Evci, Vedant Misra, Vincent Roseberry, Vlad Feinberg, Vlad Kolesnikov, Woohyun Han, Woosuk Kwon, Xi Chen, Yinlam Chow, Yuvein Zhu, Zichuan Wei, Zoltan Egyed, Victor Cotruta, Minh Giang, Phoebe Kirk, Anand Rao, Kat Black, Nabila Babar, Jessica Lo, Erica Moreira, Luiz Gustavo Martins, Omar Sanseviero, Lucas Gonzalez, Zach Gleicher, Tris Warkentin, Vahab Mirrokni, Evan Senter, Eli Collins, Joelle Barral, Zoubin Ghahramani, Raia Hadsell, Yossi Matias, D. Sculley, Slav Petrov, Noah Fiedel, Noam Shazeer, Oriol Vinyals, Jeff Dean, Demis Hassabis, Koray

Kavukcuoglu, Clement Farabet, Elena Buchatskaya, Jean-Baptiste Alayrac, Rohan Anil, Dmitry, Lepikhin, Sebastian Borgeaud, Olivier Bachem, Armand Joulin, Alek Andreev, Cassidy Hardin, Robert Dadashi, and Léonard Hussenot. 2025. Gemma 3 technical report. *Preprint*, arXiv:2503.19786.

Liang Wang, Nan Yang, Xiaolong Huang, Linjun Yang, Rangan Majumder, and Furu Wei. 2024. Multilingual e5 text embeddings: A technical report. *Preprint*, arXiv:2402.05672.

Chaojun Xiao, Haoxi Zhong, Zhipeng Guo, Cunchao Tu, Zhiyuan Liu, Maosong Sun, Yansong Feng, Xianpei Han, Zhen Hu, Heng Wang, et al. 2018. Cail2018: A large-scale legal dataset for judgment prediction. *arXiv preprint arXiv:1807.02478*.

Xuanwang Zhang, Yun-Ze Song, Yidong Wang, Shuyun Tang, Xinfeng Li, Zhengran Zeng, Zhen Wu, Wei Ye, Wenyuan Xu, Yue Zhang, Xinyu Dai, Shikun Zhang, and Qingsong Wen. 2024. RAGLAB: A modular and research-oriented unified framework for retrieval-augmented generation. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 408–418, Miami, Florida, USA. Association for Computational Linguistics.

Lucia Zheng, Neel Guha, Javokhir Arifov, Sarah Zhang, Michal Skreta, Christopher D. Manning, Peter Henderson, and Daniel E. Ho. 2025. A reasoning-focused legal retrieval benchmark. In *Proceedings of the 2025 Symposium on Computer Science and Law*, CSLAW '25, page 169–193, New York, NY, USA. Association for Computing Machinery.