

Retrieval Augmented Reasoning for Question Answering in Legal Domain

Luka Rozgic, Belmont High School, Belmont, MA

Abstract

The dynamic nature of legal systems presents a fundamental challenge for artificial intelligence applications in law: legislation evolves continuously, creating a vast ever-changing corpus of legal knowledge. While large flagship language models have the capacity to absorb comprehensive legal knowledge, their development typically occurs outside the purview of legal experts, resulting in improvements that are substantial yet unpredictable in nature and alignment with legal reasoning requirements.

In this project we address the critical need for robust, maintainable approaches to legal question answering by systematically evaluating retrieval-augmented generation (RAG) methods for complex legal question answering tasks that require reasoning and synthesis of information from multiple legal documents. We (1) Perform assessment of various retrieve-and-answer-question solutions that do not require model retraining on three recent multi-hop reasoning legal datasets: BarExamQA and HousingStatuteQA [ZGA⁺25], and the English version of KOBLEX dataset [LKH⁺25]. (2) Examine applicability of search inference-time scaling methods that perform multi-step retrieval and reasoning (including Monte Carlo Tree Search RAG [HZZC25]) to legal domain and assess if the same scaling strategies are applicable across tasks and datasets.

1 Introduction

AI solutions ^{1 2 3} based on large language models (LLMs) are increasingly used by legal professionals to automate tasks that are typically time-consuming and complex for humans. These tasks span: individual document analysis (e.g., contract review and issue spotting), document summarization, retrieval of relevant legal materials (e.g., case law and statutes pertinent to specific queries), legal drafting, legal research, and predictive analysis (e.g., forecasting potential legal outcomes or judgments).

The application of LLMs to real-world legal tasks faces two main challenges. First, legal tasks require strict adherence to rules and facts, while LLMs are prone to hallucinations. Second, the body of legal knowledge evolves continuously, and when knowledge is embedded in LLM parameters, models require costly maintenance involving unlearning and relearning, unlike domains such as mathematics and coding where unlearning is less critical.

Retrieval-Augmented Generation (RAG) offers an appealing solution to these challenges by enabling systems to retrieve relevant segments from up-to-date legal databases. However, the retrieval task itself is challenging due to limited lexical similarity between queries and legal documents (where queries may range from simple questions to document segments). Beyond retrieval, effective legal reasoning requires processing of long retrieved documents, identifying relevant rules and facts, reasoning over them, formulating and testing hypotheses, and, when necessary, performing sequential retrieval and reasoning.

We hypothesize that an LLM familiar with legal terminology and language, excelling in general reasoning over large input contexts, and paired with a performant retrieval module, can achieve high performance on real-world legal question-answering tasks. To test this hypothesis, we: (1) Select legal benchmarks that require retrieval of multiple pieces of information and reasoning over retrieved content; (2) Analyze benchmark data for potential biases that may impact the relevance of our findings; (3) Perform comprehensive testing of retrieve-and-answer methods on selected benchmarks and analyze the relationship between retrieval and QA performance; and (4) Compare the QA performance of

¹Thomson Reuters CoCounsel

²LexisNexis

³Spellbook

retrieve-and-answer methods with output-focused search methods that perform multiple retrieve and reason steps.

1.1 Related Work

LLM Hallucination Identification and Minimization Hallucination—the generation of fluent and syntactically correct content that is factually inaccurate or unsupported by external evidence—remains a critical barrier to the reliable deployment of Large Language Models (LLMs), particularly in domains requiring factual accuracy [SMG⁺24]. Mitigation techniques span multiple approaches, including prompt optimization [RZG⁺25], retrieval-augmented generation and the use of Knowledge Graphs [NWZ⁺24], and self-refinement [NLS⁺24]. In parallel, new mathematic approaches to detecting and measuring through uncertainty quantification strategies [QYK⁺24] [SML⁺25] and more principled decoding [SHL⁺24]. In legal domain prominent work [DMSH24] introduces a novel benchmark with typology of legal hallucinations and evaluates flagship LLMs on it.

Legal Retrieval and Reasoning Datasets The research on specialized legal retrieval and reasoning systems has been hindered by the lack of realistic benchmarks that capture the complexity of both legal retrieval and downstream question-answering tasks. Recent efforts have introduced novel legal RAG datasets designed to evaluate systems’ ability to retrieve relevant legal documents and answer complex legal questions requiring reasoning, including CLERC [HWQ⁺24] encompassing retrieval and summarization of retrieved documents, Bar Exam QA and Housing Statute QA [ZGA⁺25] encompassing retrieval and multiple-choice question answering, and KOBLEX [LKH⁺25] encompassing retrieval and open-ended question answering. These benchmarks correspond to real-world legal research tasks and were produced through annotation processes that resemble actual legal research and come with corresponding statute databases.

A number of legal datasets focuses on retrieval only. Massive Legal Embedding Benchmark (MLEB) is an aggregate of ten expert-annotated datasets spanning multiple jurisdictions (the US, UK, EU, Australia, Ireland, and Singapore) and document types (cases, legislation, regulatory guidance, contracts, and literature) [BBM25]. LegalBench-RAG [PA24] evaluates legal information retrieval performance and consists of four pre-existing evaluation sets focusing on US contracts.

Finally, a class of legal datasets focuses on reasoning. LegalBench [Guh24] consists of 162 tasks covering six different types of legal reasoning is used to evaluate multiple flagship LLMs. CitaLaw [ZYDX25] addresses a critical gap in legal AI by evaluating large language models’ ability to generate legally sound responses with appropriate verifiable citations. LEXam [FNM⁺25] is derived from 340 law exams spanning 116 law school courses across a range of subjects and degree levels including long-form, open-ended questions and multiple-choice questions [oth26]. While the reasoning benchmarks do not come with legal document database and in a typical evaluation setting do not use RAG they could be used in retrieve-and-reason setting.

Inference-Time Scaling Inference-time scaling is a promising approach that trades extra compute at inference time to improve performance without modifying model parameters. In essence, inference-time scaling is a problem of dynamic compute allocation and step selection (where steps can include additional thinking, task decomposition and planning, as well as tool use) where the system decides which strategy to apply and how much compute to allocate on a per-query basis. Large flagship reasoning-focused LLMs, when prompted, can natively perform inference time scaling as they are trained to do; however, smaller general purpose models require external methods to manage step selection. The scaling methods can be categorized to output-focused and input-focused ones [WWN25].

Output-focused methods include multi-step reasoning strategies Chain-of-Thought [WWS⁺23b], Tree-of-Thought [YYZ⁺23], and ReAct [YZY⁺23] each explicitly modeling reasoning process. Search output-based methods include (a) Best-of-N, (b) Depth/breadth first search that define exploration budget and identify the best solution within budget, and (b) Monte Carlo Tree Search (MCTS) that balances exploration and exploitation, through selection, expansion, simulation, and backpropagation [HGM⁺23, SM22]. Search based methods require assessment of different reasoning trajectories and employ voting, verification models [CKB⁺21], confidence [KZS25] and self-consistency [WWS⁺23a] measures to achieve this.

Input-focused methods include query expansion and transformation [JZQ⁺23] and RAG (retrieval, re-ranking, iterative retrieval). Iterative retrieval methods, Self-RAG [AWW⁺23], Auto-RAG [YZF24] leverage iterative introspection to refine intermediate outputs and are in a way input-focused equivalents of ReAct. RAG-Star [JCL⁺24] uses a fixed search tree for retrieval control, but does not use

dynamic pruning, a shortcoming addressed in MCTS-RAG [HZZC25] that supports parallel expansion of diverse reasoning strategies and incorporates pruning to maximize efficiency.

Inference-time scaling is underexplored on complex legal tasks. Notable exceptions include hard-coded multi-step processes including query expansion, retrieval and reasoning [ZGA+25] [LKH+25] and a recent comparative study [HYG+25] that evaluates 12 LLMs, including both reasoning-focused and general-purpose ones, across Chinese and English legal tasks spanning statutory and case law. The latter work does not implement test time scaling algorithms but exploits built-in thinking capabilities of used LLMs.

2 Datasets

3 Experiments and Results

3.1 Housing Statutes QA

For retrieval task we compare two best open-source encoding models from [ZGA+25], e5-base-v2 and e5-large-v2, with a newer generation encoding model Qwen 3 0.6B Embedding [ZLL+25]. For each retrieval model we evaluate recall@10 performances in five settings (a) original query only, (b,c) legal expert reformulation of the original query using Claude 3 Sonnet and Qwen 2.5 7B models, (d,e) aggregation of the original query and (b,c) expansions. Prompt used for reformulation (expansion) is based on structured legalreasoning prompt [?].

Table 1: Recall@10 Performance by Model and Query Type for the Housing Statutes dataset. Recall@10 value is an upper recall bound - recall is successful if top-10 retrieved statutes contain at least one ground truth statute for the corresponding question.

Model	Query Type	Recall@10 [95% CI]
E5_base_v2	Query	0.457 [0.445, 0.470]
	Claude 3 Sonnet	0.503 [0.491, 0.515]
	Query + Claude 3 Sonnet Expansion	0.526 [0.516, 0.538]
	Qwen 2.5 7B Instruct Expansion	0.369 [0.358, 0.380]
	Query + Qwen 2.5 7B Instruct Expansion	0.423 [0.410, 0.434]
E5_large_v2	Query	0.505 [0.495, 0.517]
	Claude 3 Sonnet Expansion	0.543 [0.531, 0.555]
	Query + Claude 3 Sonnet Expansion	0.551 [0.539, 0.563]
	Qwen 2.5 7B Instruct Expansion	0.432 [0.421, 0.444]
	Query + Qwen 2.5 7B Instruct Expansion	0.443 [0.432, 0.455]
Qwen 3 0.6B	Query	0.400 [0.389, 0.411]
	Claude 3 Sonnet Expansion	0.520 [0.508, 0.532]
	Query + Claude 3 Sonnet Expansion	0.528 [0.516, 0.538]
	Qwen 2.5 7B Instruct Expansion	0.435 [0.423, 0.447]
	Query + Qwen 2.5 7B Instruct Expansion	0.479 [0.467, 0.491]

Query expansion with strong Claude 3 Sonnet model improves retrieval over original query based retrieval for all embedding models (+2-12%) and combination of the original query with the expansion brings small additional (+1-2%) improvement. Retrieval based on query expansion with a smaller Qwen 2.5 7B model degrades performance for e5-base-v2 and e5-large-v2 embedding models, but brings improvement for Qwen 3 0.6B embedding model (+3-8%). Inspection of expansions indicate that Claude 3 Sonnet expansions are more legally proficient and specific than Qwen 2.5 7B Instruct expansions. Additionally, we hypothesize that Qwen 3 0.6B model is trained on text generated by Qwen models, so addition of Qwen based expansions help Qwen encoder while degrade performance of other encoders.

As Housing Statute dataset has questions and statutes from all US states we examined per-state retrieval performances (see 1 and 2)

Further we evaluated QA performance in (1) zero-shot setting (question only, no retrieval) establishing lower bound, (2) answering question with golden passages in the context (Oracle assuming ideal

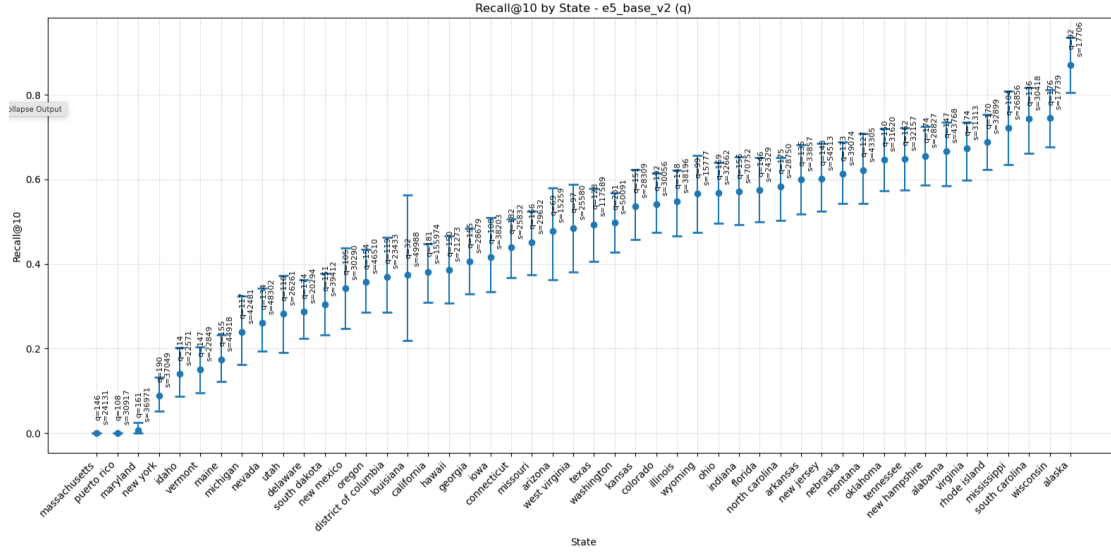


Figure 1: Per-state recall@10 performance with confidence intervals. The best embedding model e5-large-v2 is used on the original queries only.

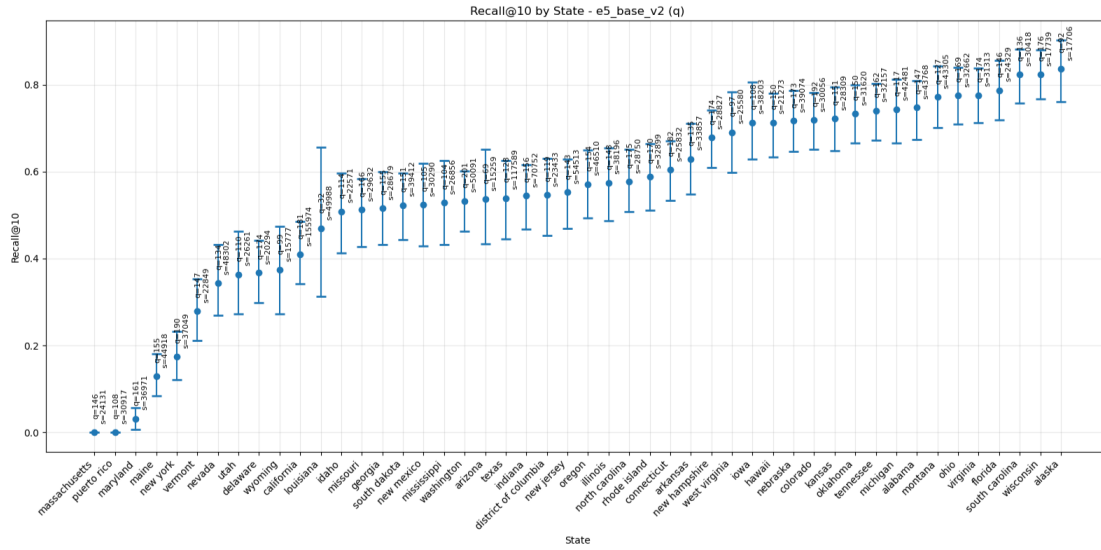


Figure 2: Per-state recall@10 performance with confidence intervals. The best embedding model e5-large-v2 is used on the original queries with Claude 3 Sonnet expansions.

retrieval) establishing upper performance bound, and (3) answering question with top 10 retrieved passages in context (end-to-end performance on retrieve and reason task) (see 4)

Table 2: QA Performance with Qwen 2.5 7B for different contexts

Context	Accuracy [95% CI]
Query only	0.556 [0.543, 0.569]
Query + top-10 recall	0.712 [0.700, 0.722]
Query + golden statute passage	0.771 [0.761, 0.781]

We examined (close to) zero recall for a few states and overall large variability in per state performance, and identified that the main reason is mismatch in golden statute passage text and the text with the same statute index in the statutes corpus, something that can be addressed in the dataset revision. For three lowest performing states the mismatch is 100% while for the other states is 80-0% range with a strong correlation with per-state recall. In total 2000 out of 6853 samples suffer from mismatch. We repeated recall and QA evaluations after removing invalid samples.

Table 3: Recall@10 Performance by Model and Query Type for the Housing Statutes dataset after removing 2000 samples with the statute text mismatch.

Model	Query Type	Recall@10 [95% CI]
e5_base_v2	Query	0.522 [0.506, 0.536]
	Claude 3 Sonnet Expansion	0.572 [0.559, 0.586]
	Query + Claude 3 Sonnet Expansion	0.599 [0.585, 0.613]
	Qwen 2.5 7B Instruct Expansion	0.416 [0.403, 0.430]
	Query + Qwen 2.5 7B Instruct Expansion	0.475 [0.461, 0.490]
e5_large_v2	Query	0.578 [0.565, 0.592]
	Claude 3 Sonnet Expansion	0.611 [0.598, 0.625]
	Query + Claude 3 Sonnet Expansion	0.621 [0.607, 0.634]
	Qwen 2.5 7B Instruct Expansion	0.486 [0.471, 0.499]
	Query + Qwen 2.5 7B Instruct Expansion	0.498 [0.484, 0.511]
Qwen 3 0.6B	Query	0.438 [0.424, 0.452]
	Claude 3 Sonnet Expansion	0.582 [0.569, 0.595]
	Query + Claude 3 Sonnet Expansion	0.590 [0.576, 0.604]
	Qwen 2.5 7B Instruct Expansion	0.484 [0.471, 0.498]
	Query + Qwen 2.5 7B Instruct Expansion	0.533 [0.520, 0.547]

Table 4: QA Performance with Qwen 2.5 7B for different contexts after removing 2000 samples with statute text mismatch.

Context	Accuracy [95% CI]
Query only	0.xxx [0.xxx, 0.xxx]
Query + top-10 recall	0.xxx [0.xxx, 0.xxx]
Query + golden statute passage	0.xxx [0.xxx, 0.xxx]

References

- [AWW⁺23] Akari Asai, Zeqiu Wu, Yizhong Wang, Avirup Sil, and Hannaneh Hajishirzi. Self-rag: Learning to retrieve, generate, and critique through self-reflection, 2023.
- [BBM25] Umar Butler, Abdur-Rahman Butler, and Adrian Lucas Malec. The massive legal embedding benchmark (mleb), 2025.
- [CKB⁺21] Karl Cobbe, Vineet Kosaraju, Mohammad Bavarian, Mark Chen, Heewoo Jun, Lukasz Kaiser, Matthias Plappert, Jerry Tworek, Jacob Hilton, Reiichiro Nakano, Christopher Hesse, and John Schulman. Training verifiers to solve math word problems, 2021.

- [DMSH24] Matthew Dahl, Varun Magesh, Mirac Suzgun, and Daniel E Ho. Large legal fictions: Profiling legal hallucinations in large language models. *Journal of Legal Analysis*, 16(1), 2024.
- [FNM⁺25] Yu Fan, Jingwei Ni, Jakob Merane, Yang Tian, Yoan Hermstrüwer, Yinya Huang, Mubashara Akhtar, Etienne Salimbeni, Florian Geering, Oliver Dreyer, Daniel Brunner, Markus Leippold, Mrinmaya Sachan, Alexander Stremitzer, Christoph Engel, Elliott Ash, and Joel Niklaus. Lexam: Benchmarking legal reasoning on 340 law exams, 2025.
- [Guh24] Neel Guha. legalbench. Hugging Face Datasets, 2024.
- [HGM⁺23] Shibo Hao, Yi Gu, Haodi Ma, Joshua Jiahua Hong, Zhen Wang, Daisy Zhe Wang, and Zhiting Hu. Reasoning with language model is planning with world model, 2023.
- [HWQ⁺24] Abe Bohan Hou, Orion Weller, Guanghui Qin, Eugene Yang, Dawn Lawrie, Nils Holzenberger, Andrew Blair-Stanek, and Benjamin Van Durme. Clerc: A dataset for legal case retrieval and retrieval-augmented analysis generation, 2024.
- [HYG⁺25] Yinghao Hu, Yaoyao Yu, Leilei Gan, Bin Wei, Kun Kuang, and Fei Wu. Evaluating test-time scaling LLMs for legal reasoning: OpenAI o1, DeepSeek-r1, and beyond. In *Findings of the Association for Computational Linguistics: EMNLP 2025*. Association for Computational Linguistics, 2025.
- [HZZC25] Yunhai Hu, Yilun Zhao, Chen Zhao, and Arman Cohan. Mcts-rag: Enhancing retrieval-augmented generation with monte carlo tree search, 2025.
- [JCL⁺24] Jinhao Jiang, Jiayi Chen, Junyi Li, Ruiyang Ren, Shijie Wang, Wayne Xin Zhao, Yang Song, and Tao Zhang. Rag-star: Enhancing deliberative reasoning with retrieval augmented verification and refinement, 2024.
- [JZQ⁺23] Rolf Jagerman, Honglei Zhuang, Zhen Qin, Xuanhui Wang, and Michael Bendersky. Query expansion by prompting large language models, 2023.
- [KZS25] Zhewei Kang, Xuandong Zhao, and Dawn Song. Scalable best-of-n selection for large language models via self-certainty, 2025.
- [LKH⁺25] Jihyung Lee, Daehui Kim, Seonjeong Hwang, Hyoungun Kim, and Gary Lee. Koblex: Open legal question answering with multi-hop reasoning, 2025.
- [NLS⁺24] Mengjia Niu, Hao Li, Jie Shi, Hamed Haddadi, and Fan Mo. Mitigating hallucinations in large language models via self-refinement-enhanced knowledge retrieval, 2024.
- [NWZ⁺24] Cheng Niu, Yuanhao Wu, Juno Zhu, Siliang Xu, KaShun Shum, Randy Zhong, Juntong Song, and Tong Zhang. 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, 2024. Association for Computational Linguistics.
- [oth26] others. Benchmarking legal reasoning on 340 law exams. *arXiv preprint arXiv:2505.12864v5*, 2026.
- [PA24] Nicholas Pipitone and Ghita Houir Alami. Legalbench-rag: A benchmark for retrieval-augmented generation in the legal domain, 2024.
- [QYK⁺24] Ernesto Quevedo, Jorge Yero, Rachel Koerner, Pablo Rivas, and Tomas Cerny. Detecting hallucinations in large language model generation: A token probability approach, 2024.
- [RZG⁺25] Kiran Ramnath, Kang Zhou, Sheng Guan, Soumya Smruti Mishra, Xuan Qi, Zhengyuan Shen, Shuai Wang, Sangmin Woo, Sullam Jeoung, Yawei Wang, Haozhu Wang, Han Ding, Yuzhe Lu, Zhichao Xu, Yun Zhou, Balasubramaniam Srinivasan, Qiaojing Yan, Yueyan Chen, Haibo Ding, Panpan Xu, and Lin Lee Cheong. A systematic survey of automatic prompt optimization techniques. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics, 2025.

- [SHL⁺24] Weijia Shi, Xiaochuang Han, Mike Lewis, Yulia Tsvetkov, Luke Zettlemoyer, and Wentau Yih. Trusting your evidence: Hallucinate less with context-aware decoding. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 2: Short Papers)*, Mexico City, Mexico, June 2024. Association for Computational Linguistics.
- [SMG⁺24] Pranab Sahoo, Prabhash Meharia, Akash Ghosh, Sriparna Saha, Vinija Jain, and Aman Chadha. A comprehensive survey of hallucination in large language, image, video and audio foundation models, 2024.
- [SML⁺25] Ola Shorinwa, Zhiting Mei, Justin Lidard, Allen Z. Ren, and Anirudha Majumdar. A survey on uncertainty quantification of large language models: Taxonomy, open research challenges, and future directions. *ACM Comput. Surv.*, 58(3), 2025.
- [WWN25] Zhichao Wang, Cheng Wan, and Dong Nie. Review of inference-time scaling strategies: Reasoning, search and rag, 2025.
- [WWS⁺23a] Xuezhi Wang, Jason Wei, Dale Schuurmans, Quoc Le, Ed Chi, Sharan Narang, Aakanksha Chowdhery, and Denny Zhou. Self-consistency improves chain of thought reasoning in language models, 2023.
- [WWS⁺23b] Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Brian Ichter, Fei Xia, Ed Chi, Quoc Le, and Denny Zhou. Chain-of-thought prompting elicits reasoning in large language models, 2023.
- [YYZ⁺23] Shunyu Yao, Dian Yu, Jeffrey Zhao, Izhak Shafran, Thomas L. Griffiths, Yuan Cao, and Karthik Narasimhan. Tree of thoughts: Deliberate problem solving with large language models, 2023.
- [YZF24] Tian Yu, Shaolei Zhang, and Yang Feng. Auto-rag: Autonomous retrieval-augmented generation for large language models, 2024.
- [YZY⁺23] Shunyu Yao, Jeffrey Zhao, Dian Yu, Nan Du, Izhak Shafran, Karthik Narasimhan, and Yuan Cao. React: Synergizing reasoning and acting in language models, 2023.
- [ZGA⁺25] Lucia Zheng, Neel Guha, Javokhir Arifov, Sarah Zhang, Michal Skreta, Christopher D. Manning, Peter Henderson, and Daniel E. Ho. A reasoning-focused legal retrieval benchmark. In *Proceedings of the Symposium on Computer Science and Law on ZZZ*, CSLAW ’25. ACM, 2025.
- [ZLL⁺25] Yanzhao Zhang, Mingxin Li, Dingkun Long, Xin Zhang, Huan Lin, Baosong Yang, Pengjun Xie, An Yang, Dayiheng Liu, Junyang Lin, Fei Huang, and Jingren Zhou. Qwen3 embedding: Advancing text embedding and reranking through foundation models, 2025.
- [ZYDX25] Kepu Zhang, Weijie Yu, Sunhao Dai, and Jun Xu. Citalaw: Enhancing llm with citations in legal domain, 2025.
- [SM22] Maciej Świechowski, Konrad Godlewski, Bartosz Sawicki, and Jacek Mańdziuk. Monte carlo tree search: a review of recent modifications and applications. *Artificial Intelligence Review*, 56(3), 2022.