ETH zürich

ETH AI CENTER

263-3300-10L Data Science Lab: Advanced Legal Robot Researcher Agent-based RAG for Complex Legal Questions

Ozgur Temmuz Celik*
Swiss Federal Institute of Technology Zurich
(ETH Zurich), Switzerland

Maximilian Scheidl*
Swiss Federal Institute of Technology Zurich
(ETH Zurich), Switzerland

Adrian Ott
Ernst & Young AG, Switzerland

Prof. Ryan Cotterell
ETH AI Center, ETH Zurich, Switzerland

Abstract

This paper investigates the design, implementation, and evaluation of a Retrieval-Augmented Generation (RAG)-based system tailored for the legal domain. The proposed architecture addresses the complexity inherent in legal question-answering by decomposing questions into sub-components, leveraging a multi-stage retrieval pipeline to enhance document relevance and applying advanced prompting techniques for the final answer generation. To systematically evaluate the system's performance, we introduce a novel evaluation framework that assesses key metrics such as clarity, accuracy, and coverage of main points. Furthermore, various prompting techniques, namely chain-of-thought prompting, least-to-most prompting, and self-refinement prompting, are evaluated for their effectiveness in the final analysis. Our results demonstrate the effectiveness of our evaluation strategy in reliably assessing system performance and aligning with expert judgments. Among the prompting techniques, Self-Refinement Prompting proved to be the most effective, delivering the highest clarity, accuracy, coverage and overall score. Lastly, we highlight opportunities for further improvements in retrieval methods and prompting strategies to optimize the system's performance and robustness.

Keywords

Retrieval-Augmented Generation, legal question-answering, artificial intelligence, natural language processing, prompting techniques, document retrieval, legal technology

1 Introduction

In recent years, artificial intelligence (AI) and natural language processing (NLP) have transformed numerous industries by automating repetitive tasks and improving decision-making processes. The legal field, characterized by its reliance on complex textual data and nuanced reasoning, has started adopting these technologies to streamline workflows, reduce costs, and enhance accessibility to legal knowledge.

Among emerging AI paradigms, Retrieval-Augmented Generation (RAG) has garnered attention for its ability to combine document retrieval with generative modeling. RAG systems excel in scenarios requiring domain-specific knowledge retrieval and contextualized reasoning, making them well-suited for complex applications such as legal question-answering. However, applying RAG architectures to the legal domain presents unique challenges, including the need for precise retrieval, handling multi-faceted legal questions, and ensuring the factual accuracy of generated outputs.

This work presents a novel RAG-based system designed for legal question-answering tasks. The architecture incorporates a subquestion decomposition strategy, enabling it to handle intricate legal queries by dividing them into smaller components and retrieving relevant context for each subquestion. Additionally, we propose a custom evaluation framework tailored to assess the performance of RAG systems in this domain. This framework evaluates key metrics such as clarity, accuracy, and the coverage of main points, ensuring a systematic approach to measuring performance. We further investigate the efficacy of various prompting strategies, including chain-of-thought prompting, least-to-most prompting, and self-evaluation prompting, to determine their impact on the final analysis.

Through rigorous experimentation, we assess the system's performance and provide insights into its strengths and limitations. The findings underscore the potential of RAG systems in transforming legal workflows and highlight the importance of robust evaluation frameworks in advancing AI-driven solutions for the legal domain.

2 Related Work

The adoption of Retrieval-Augmented Generation (RAG) [7] architectures for complex question-answering tasks has gained significant attention in recent years. These systems combine a retrieval module with a generative language model, enabling them to address tasks requiring domain-specific knowledge while maintaining the flexibility of generative reasoning. A significant advantage of RAGs over LLMs is their improved factual accuracy, especially when dealing with rapidly changing information [2]. The ability to add and remove information from a database is also important given challenges of model unlearning [13] and the increased legal attention

*Equal contribution.

263-3300-10L Data Science Lab, December 19, 2024, Zurich, Switzerland

this topic has received [4]. These facts make RAGs ideal for legal question answering systems.

Using RAGs to aid in legal work is also gaining ground. For example, a recent work [12] has suggested using case-based reasoning for RAG systems for legal question answering. Another study [8] has introduced a dataset to evaluate RAG systems on long-form legal question answering and tested it on an end-to-end retrieve-then-read type RAG system.

As we can see, the evaluation of RAGs is a topic in its own right, and multiple approaches have been proposed to evaluate them. A common approach involves evaluating the answer quality with a simple LLM call by asking the model to compare the predicted answer with the ground truth. On the other hand, there are readily available libraries that perform more complex LLM calls, such as DeepEval [6] and GroUSE [9], both specifically designed for LLMs. These libraries natively accept the question, context, predicted answer, and ground truth to generate scores in predefined categories, for which they make LLM calls with partially modifiable prompts. Ragas [3] also uses three predefined categories: faithfulness, answer relevance, and context relevance. It uses an LLM-call technique similar to DeepEval and GroUSE for faithfulness and context relevance, while for answer relevance, it generates synthetic questions based on the predicted answer and then computes their cosine similarity to the original question in the embedding space. This enables Ragas to evaluate without needing ground-truth answers. Ares [10] evaluates RAG systems on the same three metrics as Ragas but uses LLM judges (classifiers) trained with synthetic data to generate predictions for each one of those metrics and then uses 150 hand-annotated samples to generate confidence intervals with prediction-powered inference.

3 Evaluation

3.1 Dataset

The dataset used in the project is provided by EY team and consists of 44 complex legal questions coming from textbooks. Each of these questions includes a textbook-derived ground-truth answer, an AI-generated response, and an expert rating of that response. The dataset is provided to us in CSV format. Throughout our tests, we realized that some samples were performing much worse than expected, and upon further inspection and talks with EY legal team, we found that those samples were follow-up questions to a previous question and hence lack necessary context. We corrected the missing context problem for those questions and used the corrected dataset. All the results presented in this paper are from this corrected dataset.

3.2 Framework

We want our evaluation system to be aligned well with the experts and also for a given question and ground truth answer, produce high scores for the positive answers, rephrased version of ground truth answer, and low scores for negative answer, ground truth answer to a completely unrelated questions.

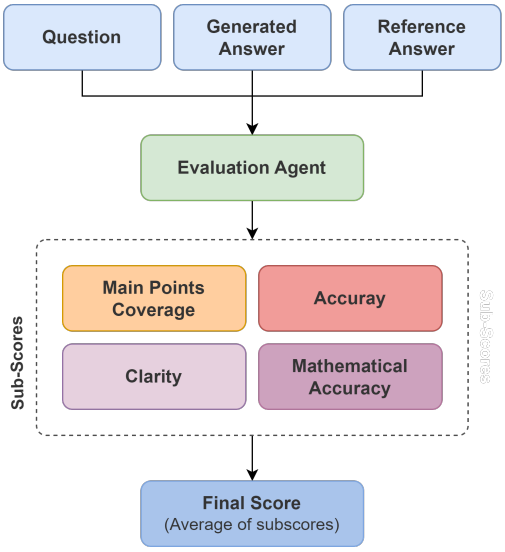

Figure 2: Our Evaluation Framework

We initially tried to use the evoluation systems mentioned in the related work, namely DeepEval, GroUSE, Ragas, and Ares. DeepEval and Grouse ratings did not align well with our expert evaluations. After consulting with the EY legal team, we determined that Ragas's metrics were not well suited to legal questions. Although Ares's confidence intervals were appreciated, it faced similar issues as Ragas and was also constrained by an insufficient sample size. Moreover, our primary objective is to assess the quality of final answers in a dataset where ground-truth answers are available. While some approaches discussed do not depend on ground-truth answers, we do not impose such a restriction on ourselves. Therefore, we decided to develop our own evaluation system. After consulting with the EY legal team, we chose four evaluation metrics: main points, accuracy, clarity, and, when applicable, mathematical accuracy. We rely on separate LLM calls for each metric because different metrics may require different inputs. By providing tailored inputs for each metric, we achieve better results than we would by making a single call and asking the LLM to focus only on certain inputs for certain metrics.

Main Points Score. For this criteria, we pass the LLM the question, predicted answer and the ground truth answer. We prompt the LLM to first identify main points (key elements and concepts) ground truth answer adresses in the question and then we ask LLM how well those main points are covered in the predicted answer. We are not interested in the accuracy of those main points in predicted answer and just prompt LLM to see if they are mentioned in the predicted answer. Overall, we prompt LLM to first identify and list those main points, evaluate the predicted answer on how well it covers them by highlighting key areas of agreement and discrepancy, and then score it from 1 to 5. We realized explicitly listing and evaluating predicted answer on each one of those main points increase the performance and stability.

Accuracy Score. For this criteria, we pass the LLM question, predicted answer and ground truth answer. We first ask LLM to identify

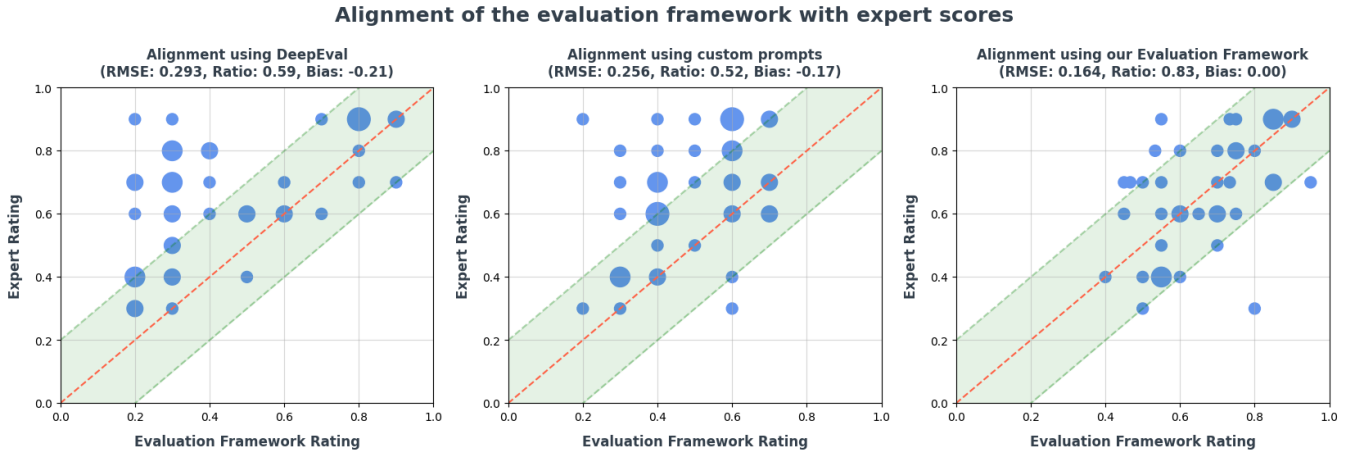

Figure 1: Score Alignment

and list facts and information found in the ground truth answer, and then for each one of those facts, we ask it to evaluate the predicted answer by looking how accurately those facts are represented in predicted answer. We then ask LLM to output a score between 1 and 5. It should be noted that this metric focuses on the correctly understanding and applying the legal facts rather than accuracy of mathematical calculations. Latter are measured separately.

Clarity Score. This criteria solely focuses on how well structured the given answer is and how well it presents the information it has. For this reason, we only pass the predicted answer to an LLM and do not evaluate the it on its correctness. We prompt an LLM to evaluate the predicted answers clarity by highlighting points where it is not clear, and then score it between 1 and 5.

Mathematical Accuracy Score. Many questions in our dataset require mathematical calculations, for example to determine the tax owed. Our discussions with EY legal team revealed that those mathematical calculations, whenever present, are highly important. For this reason we separated the accuracy of mathematical calculations from the accuracy of the representation of laws, but the way we prompt the LLM is quite similar. We provide the LLM question, predicted answer, and the ground truth answer to LLM. We then ask it to identify and list mathematical calculations in ground truth answer, then evaluate the predicted answer on each one them, and finally provide a score between 1 and 5. But additionally, we ask it to return a score of 0 if there are no calculations in the ground the answer.

After getting subscores for those metrics we take the average of them. If the mathematical accuracy score is not 0, then it is also included in the average, otherwise, it is not.

3.3 Answer alignment

One of our goals is to make sure our evaluation method produces ratings that align well with the expert ratings from EY. Those ratings are between 0.1 to 1.0 so we scale predicted ratings. Following our discussions with the EY legal team, our expectation is to make sure that almost all the predicted ratings fall in between ±0.2 of expert ratings.

In **??**, we can compare the DeepEval based evaluation, initial naive prompting evaluation system and our final evaluation system. To see how well an evaluation system aligns with the expert evaluations, we look at three metrics: RMSE, ratio of samples fall within between ±0.2 of expert rating, and the bias compared to expert ratings. DeepEval had the most RMSE, which is the measure of variance of difference between our evaluation and expert rating, and the highest bias. We can see that even a naive prompting outperformed DeepEval in RMSE and bias. On the other hand DeepEval had slightly higher ratio of samples falling in between ±0.2 of expert rating. On the other hand, our final evaluation system has almost halved the RMSE of DeepEval, increased the ratio 20% in absolute terms, and reduced the bias to 0. Such strong alignment with our evaluation system and expert ratings make it possible for us to test the effects of the changes we made on RAG system without needing experts to check the generated results.

3.4 Positive and Negative Answer Comparison

Training with synthetic data papers, like Ares [10], commonly use synthetic positive and negative samples where the model is expected to perform output positive or negative ratings. We wanted to apply a similar method to check our evaluation system. For the positive samples, we prompt an LLM to rephrase the ground truth answer and give that rephrased ground truth answer as the predicted answer. In this case we expect high score across all the subscores. For the negative samples, we simply provide the ground truth answer of another question as the predicted answer. We expect the clarity score to be high while other subscores are low.

Now we will compare two versions of our evaluation strategy. First one is the previous version where we used a single LLM call to get the ratings for all subscores, and the second one is the current method where we use separate LLM calls. In Table 1, we can see that our current technique, separate prompts, rate the positive samples high and the negative samples quite low, except for their clarity score. On the other hand, the previous method, while doing well for

the positive samples, do not rate the negative samples low enough. Since the negative answers we provide are randomly sampled from list other ground truth answers, they should be almost always completely irrelevant leading to very low scores for all but clarity subscore. This is indeed what we observe for our current evaluation method.

Table 1: Positive And Negative Answers

	Main Points	Accuracy	Clarity	Math Accuracy
CP	4.98	4.82	4.84	4.96
CN	2.43	2.97	4.36	2.03
SP	4.95	4.98	4.09	4.92
SN	1.04	1.04	3.77	1.06

C for previous combined prompt technique, S for current separate prompt technique, P for positive answers, and N for negative answers

4 Methodology

4.1 Architecure and Methods

Our RAG architecture. The diagram illustrates a high-level overview of our RAG architecture, which processes complex questions by breaking them into subquestions. These subquestions undergo a structured pipeline starting with a case breakdown, where the query is divided into subquestions and filters. For each subquestion, the system retrieves relevant document chunks using HyDE queries and hybrid search, followed by an expansion step that gathers contextual information around the retrieved chunks. The chunks are then reranked using a specialized model to prioritize the most relevant content. Finally, the system synthesizes the information into a final answer, incorporating calculations and conclusions to address the original query comprehensively.

HyDE Queries. Hypothetical Document Embeddings (HyDE) [5] are introduced to bridge the gap between queries and documents in the embedding space. Retrieval methods can struggle due to the semantic gap between short and focused queries and long and detailed documents. HyDE works by creating a hypothetical document that contains the answer to the query. The common approach involves searching for the query and a hypothetical document separately. In contrast, we took a different route: while the questions are mostly in English, we generate the hypothetical answer to it in German and append this answer to the question. After that we do a single hybrid search, both dense and sparse, and retrieve the chunks. This has the added benefit of bridging the language gap between the queries and documents since most of the queries are in English whereas almost all the documents are in German. Thanks to that we can now use sparse search methods like bm25 more effectively.

Context Enhancement. A common method of context enhancement is using contextual chunk headers where each chunk is prepended with its chunk header which can be document title, section title etc. Realizing that the retrieved information for subquestions sometimes lack this context we follow a similar logic and append each

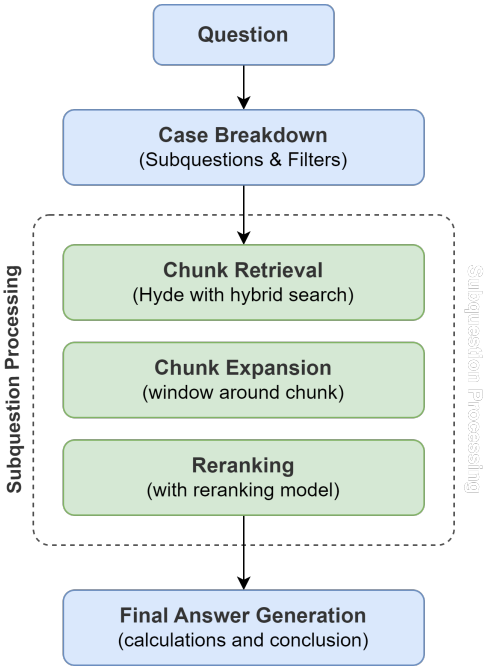

Figure 3: Our RAG Architecture

retrieved block (at the subquestion level) with the corresponding subquestion. This follows the template "For the question {subquestion}, the retrieved information is {retrieved_information}"

Question partitioning. In our RAG architecture, we address complex legal questions by splitting them into multiple subquestions. This approach allows us to break down intricate queries into manageable components, each of which can be addressed more effectively. For each subquestion, we retrieve relevant documents separately, ensuring that the context provided to the model is both precise and tailored to the specific aspect of the problem being analyzed.

4.2 Final Analysis prompting techniques

To improve the answer quality of the RAG system, we investigate three different prompting techniques for solving the legal questions at hand. By incorporating well-known and widely used prompting frameworks, we aim to ensure the robustness of our techniques. In this section, the three proposed techniques are presented and analyzed in detail.

Chain-Of-Thought Prompting. Our initial and most simple final analysis technique utilizes Chain-Of-Thought[11] prompting to generate an answer. It involves passing the retrieved context and the initial query directly to the language model, augmented by a detailed and structured system prompt. The system prompt establishes a professional Swiss Tax Law AI assistant framework with clear guidelines for analysis, ensuring a systematic approach to identifying applicable laws, conducting case analysis, and generating conclusions. Since this method only requires one LMM call it is the fastest of the presented techniques in this section.

Least-To-Most Prompting. Least-to-Most Prompting [14] employs a structured decomposition strategy, breaking down a complex legal question into a sequence of simpler, interconnected subproblems. In the initial step, the language model is tasked with identifying up to four distinct but related subproblems, limiting the number to ensure manageable processing times. If the question's complexity allows, fewer subproblems may be generated. Each subproblem is then addressed sequentially, with the solutions from prior steps incorporated as context for the subsequent ones. Once all subproblems are resolved, a final answer is synthesized, akin to the Chain-Of-Thought approach, but with an enriched context that includes the resolved subproblems.

Self-Refinement prompting. Self-Refinement Prompting [1] introduces an iterative review mechanism, enabling the language model to critically assess and refine its responses. Initially, the model generates an answer using the Chain-Of-Thought technique. It then evaluates this answer against three key criteria: coverage of the main points of the question, accuracy of presented facts and information, and clarity of the answer. For each criterion, the model assigns a score from 1 to 5. If the average score falls below a predefined threshold, the model can refine its answer, iterating up to three times to avoid unnecessary processing loops. Once the average score exceeds the threshold or the maximum number of refinements is reached, the answer is finalized and returned.

5 Experimental Setup

The database was provided to us by EY along with the backbone of the retrieval system and methods that were desired to be included. As the embedding model, we use text-embedding-3-large by OpenAI. For our database, we used Azure Database. OpenAI's GPT-4o mini as the planner agent, and OpenAI's GPT-4o as the analysis and evaluation agents. The same GPT-4o model was also used in DeepEval and GroUSE. Consequently, the chain-of-thought prompting, least-to-most prompting, and self-evaluation prompting techniques we experimented with for the analysis agent were all tested using GPT-4o as well.

6 Results

6.1 Final Analysis propmting

To evaluate the effectiveness of our proposed final analysis prompting techniques, we tested each of them against our test data using GPT-4o. Table 2 presents the results of these evaluations, comparing the techniques across our four key metrics: coverage of main points, accuracy, clarity, and mathematical accuracy.

Table 2: Evaluation of Final Analysis Prompting Techniques

	Chain-of-thought	Least-to-most	Self-Refine
Main Points	2.91	3.18	3.16
Accuracy	2.86	2.89	3.05
Clarity	4.95	4.75	4.98
Math Accuracy	2.59	2.57	2.84
Overall Score	3.49	3.47	**3.63**

Among the tested approaches, Self-Evaluation Prompting achieved the highest overall score of 3.63, demonstrating its effectiveness in producing comprehensive and accurate answers. Its iterative refinement mechanism allowed it to score the highest in accuracy (3.05) and mathematical accuracy (2.84), while also achieving near-perfect clarity (4.98). This indicates that incorporating a self-review process improves the system's ability to identify errors and enhance the quality of responses.

Chain-of-thought Prompting, the simplest and fastest method, achieved an overall score of 3.49, excelling in clarity (4.95) but lagging in other metrics. Its reliance on a single model pass makes it efficient but less robust for complex queries, where the lack of iterative refinement or problem decomposition limits its ability to improve coverage and accuracy.

Least-to-Most Prompting, with an overall score of 3.47, showed strength in covering main points (3.18), outperforming both Chain-of-thought and Self-Evaluation Prompting in this metric. By breaking down complex queries into manageable subproblems, it provided a structured framework for analysis. However, its slightly lower scores in clarity (4.75) and mathematical accuracy (2.57) suggest that while decomposition helps with thoroughness, it can introduce minor inconsistencies in presentation and calculation.

6.2 Model Comparison

Cost is a critical consideration in designing RAG systems, especially when balancing budget constraints with performance requirements. In our analysis, we compared the performance of our system using GPT-4o and the more cost-effective GPT-4o-mini as the models for final analysis prompting. This evaluation aimed to quantify the trade-offs between performance and cost in practical deployment scenarios.

Table 3: Evaluation of our RAG system using GPT-4o and GPT-4o-mini for the final analysis prompting

	Self-Refine with 4o	Self-Refine 4o-mini
Main Points	3.16	2.45
Accuracy	3.05	2.45
Clarity	4.98	4.95
Math Accuracy	2.84	2.18
Overall Score	**3.63**	**3.16**

The results, summarized in Table 3, demonstrate a noticeable performance advantage when using GPT-4o over GPT-4o-mini. While both models exhibit high levels of clarity in their responses, GPT-4o consistently outperforms GPT-4o-mini across all other evaluation metrics, including the ability to capture main points, accuracy, and mathematical reasoning. For example, the overall score for GPT-4o reached 3.63 compared to GPT-4o-mini's 3.16, underscoring the superior analytical capabilities of the former.

7 Discussion

This study demonstrates the feasibility and effectiveness of applying Retrieval-Augmented Generation (RAG) architectures to

complex legal question-answering tasks. By leveraging a subquestion decomposition approach, the proposed system addresses the challenges associated with retrieving and synthesizing information for intricate legal queries. We also show the effectiveness of RAG systems for legal questions even when the database is a non-English language and the queries themselves are in English. Experimental results highlight the strengths of various prompting techniques, with Self-refinement prompting emerging as particularly effective.

Despite these advances, the study identifies areas requiring further improvement. Enhancing the retrieval mechanism to ensure higher precision and relevance of retrieved documents remains a key priority. Furthermore, the integration of advanced prompting frameworks and the exploration of novel methods such as hierarchical prompting could further improve performance in legal contexts. Addressing these limitations will be essential for scaling the system to handle more complex and diverse legal tasks. Additionally, evaluation system relies on the ground truth answers limiting its applicability to more general scenarios. Furthermore, evaluation system only evaluates the end result of the RAG system. A more granular evaluation system where we can evaluate different modules like retriever quality can increase the effectiveness of the evaluation.

The findings of this study contribute to the growing body of research on AI-driven solutions in the legal domain and provide a foundation for future work. By bridging the gap between retrieval and generation, the proposed RAG architecture offers a promising pathway for enhancing the accessibility and efficiency of legal services, ultimately paving the way for broader adoption of AI technologies in the legal sector.

8 Future Work

While our current implementation offers valuable insights into the performance of various prompting techniques, there are several areas for future improvement and exploration we would like to outline:

Optimizing Retrieval Mechanisms. We believe there is significant potential in improving the retrieval mechanism and document structure. Doing so would ensure that the final analysis model receives the most relevant and accurate information, thereby improving the overall performance of the system.

Expanding Evaluation Dataset. Currently, the test set contains 44 legal problems. It would be reasonable to extend this dataset further to ensure more accurate and stable evaluation.

Enhancing the Final Analysis Stage. Future research could investigate alternative methods or hybrid approaches to refine the final analysis stage. This could further enhance the quality, accuracy, and reliability of generated responses.

Acknowledgments

We would like to thank the EY team for providing us with such an interesting challenge. A special thanks goes to Dimitar Dimitrov for the fantastic collaboration and great support throughout the project. We also want to thank the Data Science Lab team for their organization, as well as our academic coach for their guidance and encouragement.

References

[1] [n. d.].
[2] 2024. https://www.edps.europa.eu/data-protection/technology-monitoring/techsonar/retrieval-augmented-generation-rag_en
[3] Shahul Es, Jithin James, Luis Espinosa-Anke, and Steven Schockaert. 2023. RAGAS: Automated Evaluation of Retrieval Augmented Generation. arXiv:2309.15217 [cs.CL] https://arxiv.org/abs/2309.15217
[4] European Commission. 2016. Regulation (EU) 2016/679 of the European Parliament and of the Council of 27 April 2016 on the protection of natural persons with regard to the processing of personal data and on the free movement of such data, and repealing Directive 95/46/EC (General Data Protection Regulation) (Text with EEA relevance). https://eur-lex.europa.eu/eli/reg/2016/679/oj
[5] Luyu Gao, Xueguang Ma, Jimmy Lin, and Jamie Callan. 2022. Precise Zero-Shot Dense Retrieval without Relevance Labels. arXiv:2212.10496 [cs.IR] https://arxiv.org/abs/2212.10496
[6] Jeffrey Ip, jwongster2, Kritin Vongthongsri, Anindyadeep, Vasilije, Pratyush K. Patnaik, agokrani, lplcor, Vytenis Šliogeris, Jan F., fschuh, Jonathan Bennion, Andrea Romano, Simon Podhajsky, Fabian Greavu, Jonas, Philip Nuzhnyi, Andrés, Ananya Raval, João Felipe Pizzolotto Bini, Martino Mensio, Andy, pedroallenrevez, César García, oftenfrequent, nabeel chhatri, John Lemmon, Yudhiesh Ravindranath, Rohinish, and Nikita Parfenov. 2024. *confident-ai/deepeval*. https://github.com/confident-ai/deepeval
[7] Patrick Lewis, Ethan Perez, Aleksandra Piktus, Fabio Petroni, Vladimir Karpukhin, Naman Goyal, Heinrich Küttler, Mike Lewis, Wen tau Yih, Tim Rocktäschel, Sebastian Riedel, and Douwe Kiela. 2021. Retrieval-Augmented Generation for Knowledge-Intensive NLP Tasks. arXiv:2005.11401 [cs.CL] https://arxiv.org/abs/2005.11401
[8] Antoine Louis, Gijs van Dijck, and Gerasimos Spanakis. 2023. Interpretable Long-Form Legal Question Answering with Retrieval-Augmented Large Language Models. arXiv:2309.17050 [cs.CL] https://arxiv.org/abs/2309.17050
[9] Sacha Muller, António Loison, Bilel Omrani, and Gautier Viaud. 2024. GroUSE: A Benchmark to Evaluate Evaluators in Grounded Question Answering. arXiv:2409.06595 [cs.CL] https://arxiv.org/abs/2409.06595
[10] Jon Saad-Falcon, Omar Khattab, Christopher Potts, and Matei Zaharia. 2024. ARES: An Automated Evaluation Framework for Retrieval-Augmented Generation Systems. arXiv:2311.09476 [cs.CL] https://arxiv.org/abs/2311.09476
[11] Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Brian Ichter, Fei Xia, Ed Chi, Quoc Le, and Denny Zhou. 2023. Chain-of-Thought Prompting Elicits Reasoning in Large Language Models. arXiv:2201.11903 [cs.CL] https://arxiv.org/abs/2201.11903
[12] Nirmalie Wiratunga, Ramitha Abeyratne, Lasal Jayawardena, Kyle Martin, Stewart Massie, Ikechukwu Nkisi-Orji, Ruvan Weerasinghe, Anne Liret, and Bruno Fleisch. 2024. CBR-RAG: Case-Based Reasoning for Retrieval Augmented Generation in LLMs for Legal Question Answering. arXiv:2404.04302 [cs.CL] https://arxiv.org/abs/2404.04302
[13] Yuanshun Yao, Xiaojun Xu, and Yang Liu. 2024. Large Language Model Unlearning. arXiv:2310.10683 [cs.CL] https://arxiv.org/abs/2310.10683
[14] Denny Zhou, Nathanael Schärli, Le Hou, Jason Wei, Nathan Scales, Xuezhi Wang, Dale Schuurmans, Claire Cui, Olivier Bousquet, Quoc Le, and Ed Chi. 2023. Least-to-Most Prompting Enables Complex Reasoning in Large Language Models. arXiv:2205.10625 [cs.AI] https://arxiv.org/abs/2205.10625