A Novel Multi-Step Prompt Approach for LLM-based Q&As on Banking Supervisory Regulation

Daniele Licari[1,2,*,†], Canio Benedetto[1,†], Praveen Bushipaka[2], Alessandro De Gregorio[1,†], Marco De Leonardis[1,†] and Tommaso Cucinotta[2]

[1]Banca d'Italia, Via Nazionale, 91, Rome, 00184, Italy

[2]Scuola Superiore Sant'Anna, P.zza dei Martiri della Libertà, 33, Pisa, 56100, Italy

Abstract

This paper investigates the use of large language models (LLMs) in analyzing and answering questions related to banking supervisory regulation concerning reporting obligations. We introduce a multi-step prompt construction method that enhances the context provided to the LLM, resulting in more precise and informative answers. This multi-step approach is compared with standard "zero-shot" and "few-shot" approaches, which lacks context enrichment. To assess the quality of the generated responses, we utilize an LLM evaluator. Our findings indicate that the multi-step approach significantly outperforms the zero-shot method, producing more comprehensive and accurate responses.

Keywords

Regulatory Q&A, Banking Supervisory Reporting Regulation, Artificial Intelligence, GenAI, GPT-4o, RAG, LLM Evaluator

1. Introduction

The advent of generative AI (GenAI), and specifically of large language models (LLMs), offers significant opportunities, among others, in the legal and financial sector, facilitating the implementation of innovative solutions across various domains of activities [1, 2, 3, 4, 5]. One of the most promising applications is the business case for supporting the navigation and analysis of complex regulatory documents [6, 7, 8, 9], which can be particularly valuable for compliance officers, legal teams, and other professionals working in financial institutions who need to have a clear and timely understanding of the regulations and the consequently derived obligations.

Supervisory authorities could benefit from a tool that streamlines the consultation of complex legislation, providing swift responses to entities and enhancing efficiency [10]. While LLMs offer advantages for this purpose, they also pose risks like bias and inaccuracies [11]. Therefore, it is essential to establish strong verification procedures and retain human supervision to counter these risks. The complexity of regulatory documents, with their dense network of cross-referenced texts/cats and specialized content, necessitates careful analysis to retrieve the needed information ensuring at the same time effective risk management and limit the burden of such manual compliance.

This study introduces a novel methodology to automate and expedite the "question & answer" (Q&A) process in regulatory compliance, leveraging advanced large language models (LLMs) to provide accurate and timely responses to inquiries about the European Banking Authority's (EBA) reporting regulations. Our multi-step approach aligns with Retrieval-Augmented Generation (RAG) principles, enhancing context retrieval and generative capabilities through mechanisms like explicit extraction of Capital Requirements Regulation (CRR) references, implicit reference analysis, and a dedicated cross-encoder for precise regulatory text retrieval. This methodology ensures tailored response generation suited to the complex regulatory compliance context, where precise and comprehensive answers are crucial.

Our work finds particular applications within the domain of EBA regulatory reporting because it is characterized by a large and complex set of interrelated documents, including delegated and implementing acts, technical standards, guidelines, and recommendations, which cover various aspects of financial entities. Such complexity makes the business case both challenging and rewarding.

In this work, we focus on Regulation (EU) N.2013/575, also called Capital Requirements Regulation (CRR) https://eur-lex.europa.eu/legal-content/en/ALL/?uri=

CLiC-it 2024: Tenth Italian Conference on Computational Linguistics, Dec 04 — 06, 2024, Pisa, Italy

[*] Corresponding author.

[†] The views and opinions expressed in this paper are those of the authors and do not necessarily reflect the official policy or position of the Bank of Italy.

daniele.licari@bancaditalia.it (D. Licari); canio.benedetto@bancaditalia.it (C. Benedetto); praveen.bushipaka@santannapisa.it (P. Bushipaka); alessandro.degregorio@bancaditalia.it (A. De Gregorio); marco.deleonardis@bancaditalia.it (M. De Leonardis); tommaso.cucinotta@santannapisa.it (T. Cucinotta)

0000-0002-2963-9233 (D. Licari); 0000-0002-8446-9468 (C. Benedetto); 0009-0009-7753-8662 (P. Bushipaka); 0000-0001-7577-3655 (A. De Gregorio); 0009-0004-6523-186X (M. De Leonardis); 0000-0002-0362-0657 (T. Cucinotta)

celex%3A32013R0575, specifically on the topic of Liquidity Risk as a first use case to evaluate the potential benefit of enriched context for an accurate response generation. The main reason for this choice is that this topic is supported by a relatively limited number of regulatory documents, so it was a good starting point since the regulation is not readily available in the form of a structured dataset and its pre-processing is usually a time-consuming task.

We used the actual EBA Q&As dataset [12] as the foundation for developing a system capable of generating automated responses to questions formulated by analysts on EBA reporting requirements and rules. By harnessing the capabilities of LLMs we aim to create a tool that can deliver accurate and contextually relevant answers to any inquiry on the content of the CRR.

Recent studies highlight the potential of LLMs for qualitative assessment [13, 14, 15, 16]. For this reason, in this work we also propose the use of an "LLM Evaluator" to automate the validation process.

The structure of this paper is the following. Chapter 2 introduces the methodology and provides a detailed description of the approach adopted in this study; it explains the dataset utilized and the normative retrieval techniques employed to identify the regulatory documents necessary to address the EBA's Q&As. Chapter 3 presents the LLM Evaluator and the evaluation criteria. Chapter 4 reports experimental results and results and presents the main outcomes of the study. Chapter 5 discusses challenges as well as potential areas for future developments.

2. Methodology

This research employs a multi-step methodology to construct a comprehensive prompt for the GPT-4 omni (GPT-4o) language model [17], enabling it to answer EBA-related questions effectively. This step-wise approach focuses on enriching the context provided by the user's question. First, it identifies relevant EBA regulations (specifically CRR references) within the inquiry. Second, it incorporates response examples to guide the LLM's output format ensuring alignment with EBA regulations. This enriched context is then leveraged by a powerful LLM to generate more accurate and informative responses (details in Appendix B.1).

2.1. Dataset Construction

To develop and then evaluate our LLM-based Q&A system, firstly we extracted a subset from the EBA's Single-rule-book-qa online resource [12], comprising "question-and-answer" pairs submitted to the EBA between 2013 and 2020. In particular, we focused on the following variables: question ID, question, submission date, status, topic, legal act, article [within that act], background information,final answer, submission date and status (details in Table 4, Appendix 4) Secondly, we implemented a two-step filtering process aimed at ensuring model efficacy: by excluding non-English entries, and by focusing on CRR-related questions within the same timeframe. This resulted in a final dataset of 1597 CRR-related questions and answers, which was then split into training (50%), validation (10%), and test sets (40%) for robust evaluation (token number distribution in Figure 1 in Appendix A). The distribution of samples for the dataset is summarized in Table 1.

Table 1
Sample distribution across training, validation, and test sets for CRR-related Q&A and the subset of only Liquidity Risk Q&A.

Set	CRR-related Q&A	Liquidity Risk Q&A
Training	798	58
Validation	162	12
Test	637	46

2.2. Context Enrichment

The context enrichment process is a three-step approach designed to identify, within the data set, the most relevant CRR references to provide an appropriate content to formulate the answer to the inquiry. The first step simply involves extracting explicit CRR references, if directly mentioned in the question (Article in tab 4). The second step leverages on the capabilities of the GPT-4o (prompt in Appendix C.1) to analyse the "question" and the "background information" to identify other CRR references that are not explicitly stated by the user. The last step of the process utilizes our CRR Ranker model, a cross-encoder architecture that has been trained to identify and retrieve pertinent references from the Capital Requirements Regulation in response to specific inquiries. This 3-steps comprehensive approach ensures a broader and potentially more accurate understanding of the the inquiry and the specific legal act(s) related to the CRR that the Q&A tool deems applicable.

2.2.1. CRR Ranker Training

With regard to the context enrichment, i.e. the CRR Ranker Training, we employed a specifically trained cross-encoder model [18] to identify relevant CRR references for enriching inquiry context. We used a dedicated "question-article" pair dataset derived from our EBA Q&A Train Dataset, excluding questions related to CRR Article 99 https://www.eba.europa.eu/regulation-and-policy/

single-rulebook/interactive-single-rulebook/14212 due to their frequent lack of topical relevance. Each data point consisted of a question (user query and background information), an associated CRR article, and a binary label indicating relevance (1 for relevant, 0 for not applicable).

We constructed the training dataset by selecting positive and negative samples. Positive samples comprised question-article pairs where the article explicitly addressed the user's query. Additionally, we included pairs formed by questions and implicit CRR references extracted from the user's text, context information, and official response using GPT-4o (used prompt in Appendix C.1).

Negative training samples were mined by using the BAAI bge-large-en-v1.5 pre-trained language model [19]. For the CRR Ranker Training we employed a two-phase process for negative sample selection: first, all CRR articles were encoded using the bge-large-en-v1.5 model, and cosine similarity was utilized to rank them relative to the user's question; second, a set of 20 negative examples was randomly chosen from a pre-defined ranking interval (250-300). The choice of 20 negative samples provides a good balance between computational efficiency and the availability of enough training data. This approach aimed to balance the representation of relevant and irrelevant information within the training data, ensuring the model learns to distinguish between the user's query and potentially related but ultimately off-topic CRR articles [20].

The final dataset comprised 12,533 unique "question-article" pairs with positive and negative labels. This data was split into training (10,179 pairs) and development (2,354 pairs) sets for model fine-tuning. This fine-tuning aimed to learn robust semantic representations for questions and CRR articles, enabling the model to effectively identify relevant CRR references for enriching user query context.

We selected the BAAI BGE Reranker v2 m3 model [18] as the basis for our cross-encoder, owing to its task-specific aptness and its demonstrated superior performance relative to the BGE Reranker Large [19], as reported in Section 4. We adopted the Cross-Entropy Binary Classification loss function, following the approach suggested in the BGE Rerank Git repository [21]. To promote stable convergence, we incorporated a warmup schedule (with a number of steps $0.1 \times$ len(train_data) $\times$ num_epochs step) that gradually increases the learning rate during the initial phase of training. The entire fine-tuning process was conducted over 4 epochs. We employed an evaluation interval of 800 steps during training and saved the model that achieved the highest F1 score on the development set.

Finally, we evaluated the model's retrieval ability of CRR items for a given user question on EBA Q&A Test Dataset. This evaluation employed recall metrics at various retrieval cutoffs, including recall@5, recall@10, recall@20, and recall@30 (results in Section 4).

2.3. Examples Enrichment

To improve the model's understanding of the desired response format, tone, and content, we adopted a few-shot prompting approach [22]. This involved extracting five relevant examples from the EBA Q&A Train Dataset with the same topic as the user question we want to answer. These examples served as demonstrations for the model, showcasing the ideal structure, language style, and level of detail expected in the final responses. Notably, the selection process ensured heterogeneity within the chosen topic, meaning the examples covered various aspects to promote a broader understanding. Limiting the number of examples to five struck a balance between providing diverse demonstrations and maintaining cost-efficiency during inference, as the LLM's input token length has limitations.

2.4. Answer Generation

Figure 2 in Appendix B.1 details how we construct a comprehensive prompt that enhances GPT-4o's ability to effectively answer user questions. The final prompt in Appendix C.2 integrates the enriched context (extracted CRR references) and the example enrichment (demonstrations of desired response format, tone, and content). This comprehensive prompt is fed to GPT-4o through the OpenAI API, enabling it to generate a well-reasoned and informative response that adheres to the EBA's regulatory framework and professional tone.

2.5. Comparison with RAG Principles

Our multi-step prompt approach aligns with the core principles of Retrieval-Augmented Generation (RAG) while incorporating tailored enhancements that improve context enrichment for regulatory Q&A tasks. Like RAG, our method integrates information retrieval with language generation, but it adds specialized steps to enhance context enrichment. These include explicit extraction of CRR references, implicit analysis using LLM capabilities, and precise retrieval through a dedicated cross-encoder. Compared to standard RAG, which often relies on single-stage retrieval, our structured multi-step process adds a higher level of granularity, including example enrichment through few-shot prompts. This ensures not only factual accuracy but also alignment with domain-specific language standards, ultimately improving response quality for complex regulatory inquiries. Overall, our approach extends the RAG principles to generate tailored, contextually enriched answers, which is particularly beneficial for the intricate requirements of regulatory compliance.

3. LLM Evaluator

In our pipeline, we employ an LLM Evaluator to assess the quality of generated responses, defined in Section 2, compared to the EBA's answers already provided. Employing an LLM Evaluator offers significant advantages in terms of cost-effectiveness and efficiency compared to traditional human evaluation/comparison methods. Recent research highlights the potential of LLMs for large-scale natural language evaluation tasks [23, 24, 25].

The evaluation process uses a scale from one to four, based on two evaluation criteria: correctness and completeness. A generated response is considered correct if its content aligns with the information presented in the official answer. Additionally, a response is deemed complete if it incorporates all relevant regulatory references provided in the official answer. The following scoring rubric outlines the evaluation criteria:

- **Score 1:** The *generated answer* is completely incorrect and incomplete compared to the *official answer.*
- **Score 2:** The *generated answer* is incorrect but either complete or partially complete compared to the *official answer.* It contains some useful information found in the *official answer*, but the main statement is incorrect.
- **Score 3:** The *generated answer* is correct but only partially complete. The main statement matches the *official answer*, but some information from the *official answer* is missing.
- **Score 4:** The *generated answer* is fully correct and complete. It is essentially a rephrased version of the *official answer* with no significant differences.

To preliminary validate the effectiveness of our LLM evaluator, we conducted an experiment using a synthetic dataset. This dataset was carefully designed to test various aspects of language generation and was evaluated by both a human expert and the LLM. The alignment between the human expert's assessments and those of the LLM was then analyzed. The complete details of the final prompt used for LLM evaluator are provided in Appendix C.3.

The dataset comprises 60 Q&A pairs, balanced across the four score categories. For each category, two pairs were excluded as they were used as examples for the prompt for the LLM evaluator, resulting in a final dataset of 52 Q&A pairs to measure the alignment between the human and LLM evaluator. Using GPT-4o, we obtained a Kendall-tau coefficient of 0.77, with a p-value of $6 \cdot 10^{-11}$. These results justified the adoption of the LLM evaluator over a human one, especially for tasks involving prompt optimization and evaluation. The figure in Appendix B.2 illustrates the complete process of evaluating agreement between the LLM evaluator and the human expert.

4. Experiments and Results

This section describes the results obtained by measuring retrieval effectiveness and answer quality. Retrieval performance is measured by the number of relevant regulations retrieved (recall) using different encoder models. Answer quality is then evaluated by a separate LLM, which scores each generated response based on factors like relevance and adherence to EBA legal acts. We compare the multi-step prompt approach with a few-shot and zero-shot one focusing on a single topic within the EBA Q&A framework, specifically Liquidity Risk. Finally, we test our Multi-Step pipeline with other LLM models, such as Google Gemini Flash 1.5 and Llama 3.1 70B.

4.1. CRR Retrieval

We employed "recall" as the primary metric to assess the effectiveness of bi and cross encoder models in retrieving relevant CRR articles based on the information submitted with the inquiry. "Recall" signifies the proportion of truly relevant CRR articles retrieved from the dataset compared to all the pertinent actual articles [26]. In the context of legal information retrieval, prioritizing the retrieval of all crucial regulatory information for the inquiry makes the recall a particularly relevant metric.

Our primary objective was to identify a model that delivers exceptional retrieval accuracy while maintaining computational efficiency. This potentially excluded models with an extremely large number of parameters, as they can be computationally expensive to run.

We conducted a performance comparison between our fine-tuned CRR Ranker and several pre-trained models:

- Bi-encoders: all-MiniLM-L6-v2 [27], gte-large-en-v1.5 [28], and bge-large-en-v1.5 [19].
- Cross-encoders: bge-reranker-large [19], bge-reranker-v2-m3 [29, 18].

The detailed results (presented in table 2) show the achieved recall scores on EBA Q&As Test Dataset for each model. Our fine-tuned CRR Ranker significantly outperformed all other models, achieving a more than 20% improvement compared to the best pre-trained model (bge-large-en-v1.5).

4.2. Answer Generation

Here we compare the performance of our multi-step approach with a zero-shot one for answering EBA liquidity

Table 2
Recall scores on EBA Q&As Test Dataset

Models	r@5	r@10	r@20	r@30
all-MiniLM	0.37	0.46	0.55	0.59
gte-large	0.39	0.48	0.57	0.63
bge-large	0.41	0.52	0.62	0.67
bge-reranker-large	0.17	0.23	0.31	0.38
bge-reranker-v2-m3	0.24	0.31	0.39	0.44
CRR Ranker (ours)	**0.51**	**0.67**	**0.81**	**0.86**

risk inquiries, using our LLM as the evaluation system (Figure in Appendix B.3). To this end, we utilized a subset of 46 Q&As from our EBA Q&A Test dataset specifically focused on liquidity risk.

We tested:

- **Zero-Shot Approach:** for each question, a standard prompt was provided to the LLM. It encompassed both the specific query and any relevant contextual information they provided.
- **Few-Shot Approach:** for each question, a few examples were provided along with the query to guide the LLM in generating responses.
- **Multi-Step Approach:** for each question, we created prompts following our established multi-step approach, incorporating context enrichment and example enrichment (as detailed in previous sections).

The LLM Evaluator assessed each response based on its correctness and completeness relative to the official EBA response. As described in Section 3, the LLM Evaluator assigned an overall score on a scale of 1 (completely incorrect and incomplete) to 4 (fully correct and comprehensive).

Table 3 summarizes the evaluation results for responses generated by the different approaches. The "multi-step" approach consistently achieved higher counts in the high-quality rating categories compared to both the "zero-shot" and "few-shot" ones. This demonstrates that the multi-step approach significantly outperformed the other methods in terms of response quality. The LLM evaluator awarded the multi-step approach an average score of 2.7, representing a 12.5% improvement over the zero-shot and few-shot approaches, which both received an average score of 2.4. Notably, a larger portion of the responses generated by our multi-step approach received scores of 3 or higher, indicating correct answers. In contrast, only 2 out of 46 responses generated by the multi-step approach were rated as completely incorrect (score 1), compared to 6 such responses for the zero-shot approach and 11 for the few-shot approach. These findings suggest that the context enrichment in the multi-step prompts effectively guides the primary LLM toward generating more comprehensive and informative responses that accurately reflect the EBA regulations.

Table 3
Evaluation results for responses generated by zero-shot, few-shot and multi-step

Rating	zero-shot	few-shot	**multi-step (gpt4o)**
1	6	12	**2**
2	18	11	**14**
3	19	16	**26**
4	3	**7**	4

4.2.1. Other LLMs

In this section, we extend our analysis of the multi-step pipeline by incorporating evaluations using additional large language models (LLMs), specifically Google Gemini Flash 1.5 and Llama 3.1 70B. Google Gemini Flash 1.5 is widely recognized for its high-speed processing capabilities and efficiency in response generation, making it a suitable benchmark for comparative performance analysis. Conversely, Llama 3.1 70B is noted for its robustness in handling complex queries while maintaining moderate computational demands, providing an interesting contrast in terms of performance and resource efficiency.

Our experimental results indicate that the average evaluation score achieved by Google Gemini Flash 1.5 was 2.0, whereas Llama 3.1 70B attained an average score of 2.2. Notably, these scores did not surpass the performance of the GPT-4o zero-shot approach, which underscores the advanced capabilities of GPT-4o in addressing the complexities of regulatory compliance inquiries. This observation highlights the inherent strength of GPT-4o in generating accurate and contextually relevant responses, outperforming the other models under similar conditions.

Future research will focus on an in-depth analysis of these models with a view toward optimizing each step of the multi-step pipeline in a model-specific manner. By tailoring our methodology to align with the distinctive strengths and limitations of each model, we aim to further enhance the overall accuracy and reliability of the generated responses.

5. Challenges and Advancements

Our work has highlighted several key challenges that are worth discussing. One of the primary issues concerns the limited size of our test dataset. This constraint arose because we focused on the single topic of Liquidity Risk. However, to achieve robust human alignment and ensure the system addresses diverse user inquiries across EBA topics, future efforts should prioritize dataset expansion and human evaluation integration.

Another topic for reflection is that the study emphasizes the need to retrieve relevant CRR articles. Future research could investigate methods to further refine the generated responses by incorporating legal reasoning and argumentation capabilities into the LLM [30, 31], and the most relevant Q&As as examples for few-shot prompting [6].

It is also crucial to underscore the importance of optimizing prompts for this kind of application, and we plan to address this moving forward. Our future research endeavors will focus on investigating automatic prompt engineering techniques [32] leveraging the LLM Evaluator as a metric to optimize. These techniques aim to tailor and optimize prompts based on the specific topic of inquiries, enhancing overall performance.

Moreover, currently we have utilized only one model, GPT-4o, but we intend to extend our testing to include other models that have demonstrated similar performance levels in the field of open question answering [33]. This will help us identify the most effective model for our application with an unbiased evaluation [34].

Similarly, in the context of LLM evaluators, we also intend to explore additional models, including open-source options [35, 36], that have shown strong performance in assessing the quality of responses from various LLMs. This approach is expected to increase the correlation between human and LLM evaluations, thereby enhancing the system's overall accuracy and reliability. The scientific community is very active in this area to better understand the limitations of the different types of models considered as evaluators [37].

By addressing the identified limitations through increased human involvement, expanded data coverage, and domain-specific evaluation methods, we believe it is possible to enhance the system's effectiveness and generalizability across a wide range of regulatory domains.

6. Conclusion

This study explored a novel approach for generating automated responses to inquiries on the Regulation (EU) N.2013/575, specifically on the liquidity risk subject. We proposed a multi-step prompt construction method that enriches the context to be provided to LLMs, enabling them to generate more accurate and informative answers. An LLM Evaluator, which demonstrated strong agreement with human experts, was employed to compare our multi-step approach with standard zero-shot and few-shot methods that lack context enrichment. The quality of the generated responses was assessed, and our findings indicate that the multi-step approach significantly outperforms both the zero-shot and few-shot methods, resulting in responses that are more comprehensive and accurate in relation to the EBA regulation. These results suggest that the multi-step prompt construction is a promising approach for enhancing LLM performance in legal information retrieval tasks, particularly within domains with complex regulatory frameworks like regulatory reporting. Even at this early stage, the tool has demonstrated its ability to make the work of the human analyst more efficient. Future research directions include exploring the use of different LLM architectures and investigating alternative methods for incorporating human feedback into the prompt construction process. Lastly, exploring the generalization of this approach to other regulatory domains would be valuable.

Acknowledgments

The authors would like to express their sincere gratitude to Vincenzo Capone, Pamela Maggiori, Daniele Bovi, Fabio Zambuto, Francesca Monacelli, and Roberto Sabbatini (Bank of Italy) for their insightful comments and stimulating discussions on an earlier draft of this paper. Their feedback greatly enhanced the clarity and focus of our work.
They would also like to thank the anonymous reviewers for their invaluable suggestions and constructive feedback.

References

[1] S. Wu, O. Irsoy, S. Lu, V. Dabravolski, M. Dredze, S. Gehrmann, P. Kambadur, D. Rosenberg, G. Mann, BloombergGPT: A Large Language Model for Finance, 2023. URL: http://arxiv.org/abs/2303.17564, arXiv:2303.17564 [cs, q-fin].

[2] J. Lai, W. Gan, J. Wu, Z. Qi, P. S. Yu, Large Language Models in Law: A Survey, 2023. URL: http://arxiv.org/abs/2312.03718. doi:10.48550/arXiv.2312.03718, arXiv:2312.03718 [cs].

[3] C. Biancotti, C. Camassa, Loquacity and Visible Emotion: ChatGPT as a Policy Advisor, 2023. URL: https://papers.ssrn.com/abstract=4533699. doi:10.2139/ssrn.4533699.

[4] J. J. Horton, Large Language Models as Simulated Economic Agents: What Can We Learn from Homo Silicus?, 2023. URL: https://arxiv.org/abs/2301.07543v1.

[5] P. Homoki, Z. Ződi, Large language models and their possible uses in law, Hungarian Journal of Legal Studies 64 (2024) 435–455. URL: https://akjournals.com/view/journals/2052/64/3/article-p435.xml. doi:10.1556/2052.2023.00475, publisher: Akadémiai Kiadó Section: Hungarian Journal of Legal Studies.

[6] N. Wiratunga, R. Abeyratne, L. Jayawardena, K. Martin, S. Massie, I. Nkisi-Orji, R. Weerasinghe, A. Liret, B. Fleisch, Cbr-rag: Case-based reasoning for retrieval augmented generation in llms for legal question answering, 2024. URL: https://arxiv.org/abs/2404.04302. arXiv:2404.04302.

[7] A. Louis, G. van Dijck, G. Spanakis, Interpretable Long-Form Legal Question Answering with Retrieval-Augmented Large Language Models, 2023. URL: http://arxiv.org/abs/2309.17050. doi:10.48550/arXiv.2309.17050, arXiv:2309.17050 [cs].

[8] W. Zhang, H. Shen, T. Lei, Q. Wang, D. Peng, X. Wang, GLQA: A Generation-based Method for Legal Question Answering, in: 2023 International Joint Conference on Neural Networks (IJCNN), 2023, pp. 1–8. URL: https://ieeexplore.ieee.org/document/10191483?denied=. doi:10.1109/IJCNN54540.2023.10191483, iSSN: 2161-4407.

[9] A. Abdallah, B. Piryani, A. Jatowt, Exploring the state of the art in legal QA systems, Journal of Big Data 10 (2023) 127. URL: https://doi.org/10.1186/s40537-023-00802-8. doi:10.1186/s40537-023-00802-8.

[10] J. Prenio, Peering through the hype - assessing suptech tools' transition from experimentation to supervision (2024). URL: https://www.bis.org/fsi/publ/insights58.htm.

[11] L. Huang, W. Yu, W. Ma, W. Zhong, Z. Feng, H. Wang, Q. Chen, W. Peng, X. Feng, B. Qin, T. Liu, A survey on hallucination in large language models: Principles, taxonomy, challenges, and open questions, 2023. URL: https://arxiv.org/abs/2311.05232. arXiv:2311.05232.

[12] Single Rulebook Q&A | European Banking Authority, 2013-2024. URL: https://www.eba.europa.eu/single-rule-book-qa.

[13] S. Ye, D. Kim, S. Kim, H. Hwang, S. Kim, Y. Jo, J. Thorne, J. Kim, M. Seo, FLASK: Fine-grained Language Model Evaluation based on Alignment Skill Sets, 2024. URL: http://arxiv.org/abs/2307.10928. doi:10.48550/arXiv.2307.10928, arXiv:2307.10928 [cs].

[14] L. Zheng, W.-L. Chiang, Y. Sheng, S. Zhuang, Z. Wu, Y. Zhuang, Z. Lin, Z. Li, D. Li, E. P. Xing, H. Zhang, J. E. Gonzalez, I. Stoica, Judging LLM-as-a-Judge with MT-Bench and Chatbot Arena, 2023. URL: http://arxiv.org/abs/2306.05685. doi:10.48550/arXiv.2306.05685, arXiv:2306.05685 [cs].

[15] Y. Liu, D. Iter, Y. Xu, S. Wang, R. Xu, C. Zhu, G-Eval: NLG Evaluation using Gpt-4 with Better Human Alignment, in: H. Bouamor, J. Pino, K. Bali (Eds.), Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing, Association for Computational Linguistics, Singapore, 2023, pp. 2511–2522. URL: https://aclanthology.org/2023.emnlp-main.153. doi:10.18653/v1/2023.emnlp-main.153.

[16] C.-M. Chan, W. Chen, Y. Su, J. Yu, W. Xue, S. Zhang, J. Fu, Z. Liu, ChatEval: Towards Better LLM-based Evaluators through Multi-Agent Debate, 2023. URL: http://arxiv.org/abs/2308.07201. doi:10.48550/arXiv.2308.07201, arXiv:2308.07201 [cs].

[17] OpenAI, J. Achiam, S. Adler, S. Agarwal, L. Ahmad, I. Akkaya, F. L. Aleman, D. Almeida, J. Altenschmidt, S. Altman, S. Anadkat, R. Avila, I. Babuschkin, S. Balaji, V. Balcom, P. Baltescu, H. Bao, M. Bavarian, J. Belgum, I. Bello, J. Berdine, G. Bernadett-Shapiro, C. Berner, L. Bogdonoff, O. Boiko, M. Boyd, A.-L. Brakman, G. Brockman, T. Brooks, M. Brundage, K. Button, T. Cai, R. Campbell, A. Cann, B. Carey, C. Carlson, R. Carmichael, B. Chan, C. Chang, F. Chantzis, D. Chen, S. Chen, R. Chen, J. Chen, M. Chen, B. Chess, C. Cho, C. Chu, H. W. Chung, D. Cummings, J. Currier, Y. Dai, C. Decareaux, T. Degry, N. Deutsch, D. Deville, A. Dhar, D. Dohan, S. Dowling, S. Dunning, A. Ecoffet, A. Eleti, T. Eloundou, D. Farhi, L. Fedus, N. Felix, S. P. Fishman, J. Forte, I. Fulford, L. Gao, E. Georges, C. Gibson, V. Goel, T. Gogineni, G. Goh, R. Gontijo-Lopes, J. Gordon, M. Grafstein, S. Gray, R. Greene,

J. Gross, S. S. Gu, Y. Guo, C. Hallacy, J. Han, J. Harris, Y. He, M. Heaton, J. Heidecke, C. Hesse, A. Hickey, W. Hickey, P. Hoeschele, B. Houghton, K. Hsu, S. Hu, X. Hu, J. Huizinga, S. Jain, S. Jain, J. Jang, A. Jiang, R. Jiang, H. Jin, D. Jin, S. Jomoto, B. Jonn, H. Jun, T. Kaftan, L. Kaiser, A. Kamali, I. Kanitscheider, N. S. Keskar, T. Khan, L. Kilpatrick, J. W. Kim, C. Kim, Y. Kim, J. H. Kirchner, J. Kiros, M. Knight, D. Kokotajlo, L. Kondraciuk, A. Kondrich, A. Konstantinidis, K. Kosic, G. Krueger, V. Kuo, M. Lampe, I. Lan, T. Lee, J. Leike, J. Leung, D. Levy, C. M. Li, R. Lim, M. Lin, S. Lin, M. Litwin, T. Lopez, R. Lowe, P. Lue, A. Makanju, K. Malfacini, S. Manning, T. Markov, Y. Markovski, B. Martin, K. Mayer, A. Mayne, B. McGrew, S. M. McKinney, C. McLeavey, P. McMillan, J. McNeil, D. Medina, A. Mehta, J. Menick, L. Metz, A. Mishchenko, P. Mishkin, V. Monaco, E. Morikawa, D. Mossing, T. Mu, M. Murati, O. Murk, D. Mély, A. Nair, R. Nakano, R. Nayak, A. Neelakantan, R. Ngo, H. Noh, L. Ouyang, C. O'Keefe, J. Pachocki, A. Paino, J. Palermo, A. Pantuliano, G. Parascandolo, J. Parish, E. Parparita, A. Passos, M. Pavlov, A. Peng, A. Perelman, F. d. A. B. Peres, M. Petrov, H. P. d. O. Pinto, Michael, Pokorny, M. Pokrass, V. H. Pong, T. Powell, A. Power, B. Power, E. Proehl, R. Puri, A. Radford, J. Rae, A. Ramesh, C. Raymond, F. Real, K. Rimbach, C. Ross, B. Rotsted, H. Roussez, N. Ryder, M. Saltarelli, T. Sanders, S. Santurkar, G. Sastry, H. Schmidt, D. Schnurr, J. Schulman, D. Selsam, K. Sheppard, T. Sherbakov, J. Shieh, S. Shoker, P. Shyam, S. Sidor, E. Sigler, M. Simens, J. Sitkin, K. Slama, I. Sohl, B. Sokolowsky, Y. Song, N. Staudacher, F. P. Such, N. Summers, I. Sutskever, J. Tang, N. Tezak, M. B. Thompson, P. Tillet, A. Tootoonchian, E. Tseng, P. Tuggle, N. Turley, J. Tworek, J. F. C. Uribe, A. Vallone, A. Vijayvergiya, C. Voss, C. Wainwright, J. J. Wang, A. Wang, B. Wang, J. Ward, J. Wei, C. J. Weinmann, A. Welihinda, P. Welinder, J. Weng, L. Weng, M. Wiethoff, D. Willner, C. Winter, S. Wolrich, H. Wong, L. Workman, S. Wu, J. Wu, M. Wu, K. Xiao, T. Xu, S. Yoo, K. Yu, Q. Yuan, W. Zaremba, R. Zellers, C. Zhang, M. Zhang, S. Zhao, T. Zheng, J. Zhuang, W. Zhuk, B. Zoph, GPT-4 Technical Report, 2024. URL: http://arxiv.org/abs/2303.08774. doi:10.48550/arXiv.2303.08774, arXiv:2303.08774 [cs].

[18] J. Chen, S. Xiao, P. Zhang, K. Luo, D. Lian, Z. Liu, Bge m3-embedding: Multi-lingual, multi-functionality, multi-granularity text embeddings through self-knowledge distillation, 2024. arXiv:2402.03216.

[19] S. Xiao, Z. Liu, P. Zhang, N. Muennighoff, C-pack: Packaged resources to advance general chinese embedding, 2023. arXiv:2309.07597.

[20] H. Xuan, A. Stylianou, X. Liu, R. Pless, Hard negative examples are hard, but useful, 2021. URL: https://arxiv.org/abs/2007.12749. arXiv:2007.12749.

[21] S. Xiao, Z. Liu, P. Zhang, N. Muennighoff, FlagEmbedding/FlagEmbedding/reranker at master · FlagOpen/FlagEmbedding, 2024. URL: https://github.com/FlagOpen/FlagEmbedding/tree/master/FlagEmbedding/reranker.

[22] T. B. Brown, B. Mann, N. Ryder, M. Subbiah, J. Kaplan, P. Dhariwal, A. Neelakantan, P. Shyam, G. Sastry, A. Askell, S. Agarwal, A. Herbert-Voss, G. Krueger, T. Henighan, R. Child, A. Ramesh, D. M. Ziegler, J. Wu, C. Winter, C. Hesse, M. Chen, E. Sigler, M. Litwin, S. Gray, B. Chess, J. Clark, C. Berner, S. McCandlish, A. Radford, I. Sutskever, D. Amodei, Language models are few-shot learners, 2020. URL: https://arxiv.org/abs/2005.14165. arXiv:2005.14165.

[23] Y. Liu, D. Iter, Y. Xu, S. Wang, R. Xu, C. Zhu, G-eval: Nlg evaluation using gpt-4 with better human alignment, 2023. URL: https://arxiv.org/abs/2303.16634. arXiv:2303.16634.

[24] Y. Dubois, X. Li, R. Taori, T. Zhang, I. Gulrajani, J. Ba, C. Guestrin, P. Liang, T. B. Hashimoto, Alpacafarm: A simulation framework for methods that learn from human feedback, 2024. URL: https://arxiv.org/abs/2305.14387. arXiv:2305.14387.

[25] J. Fu, S.-K. Ng, Z. Jiang, P. Liu, Gptscore: Evaluate as you desire, 2023. URL: https://arxiv.org/abs/2302.04166. arXiv:2302.04166.

[26] C. D. Manning, P. Raghavan, H. Schütze, Introduction to Information Retrieval, Cambridge University Press, USA, 2008.

[27] P. S. H. Lewis, Y. Wu, L. Liu, P. Minervini, H. Küttler, A. Piktus, P. Stenetorp, S. Riedel, PAQ: 65 million probably-asked questions and what you can do with them, CoRR abs/2102.07033 (2021). URL: https://arxiv.org/abs/2102.07033. arXiv:2102.07033.

[28] Z. Li, X. Zhang, Y. Zhang, D. Long, P. Xie, M. Zhang, Towards general text embeddings with multi-stage contrastive learning, arXiv preprint arXiv:2308.03281 (2023).

[29] C. Li, Z. Liu, S. Xiao, Y. Shao, Making large language models a better foundation for dense retrieval, 2023. arXiv:2312.15503.

[30] F. Yu, L. Quartey, F. Schilder, Exploring the effectiveness of prompt engineering for legal reasoning tasks, in: A. Rogers, J. Boyd-Graber, N. Okazaki (Eds.), Findings of the Association for Computational Linguistics: ACL 2023, Association for Computational Linguistics, Toronto, Canada, 2023, pp. 13582–13596. URL: https://aclanthology.org/2023.findings-acl.858. doi:10.

18653/v1/2023.findings-acl.858.

[31] Y. an Lu, H. yu Kao, 0x.yuan at semeval-2024 task 5: Enhancing legal argument reasoning with structured prompts, in: International Workshop on Semantic Evaluation, 2024. URL: https://api.semanticscholar.org/CorpusID:270765544.

[32] Q. Ye, M. Axmed, R. Pryzant, F. Khani, Prompt engineering a prompt engineer, 2024. URL: https://arxiv.org/abs/2311.05661. arXiv:2311.05661.

[33] Z. Huang, Z. Wang, S. Xia, P. Liu, Olympicarena medal ranks: Who is the most intelligent ai so far?, 2024. URL: https://arxiv.org/abs/2406.16772. arXiv:2406.16772.

[34] A. Panickssery, S. R. Bowman, S. Feng, Llm evaluators recognize and favor their own generations, 2024. URL: https://arxiv.org/abs/2404.13076. arXiv:2404.13076.

[35] S. Kim, J. Suk, S. Longpre, B. Y. Lin, J. Shin, S. Welleck, G. Neubig, M. Lee, K. Lee, M. Seo, Prometheus 2: An open source language model specialized in evaluating other language models, 2024. URL: https://arxiv.org/abs/2405.01535. arXiv:2405.01535.

[36] S. Kim, J. Suk, J. Y. Cho, S. Longpre, C. Kim, D. Yoon, G. Son, Y. Cho, S. Shafayat, J. Baek, S. H. Park, H. Hwang, J. Jo, H. Cho, H. Shin, S. Lee, H. Oh, N. Lee, N. Ho, S. J. Joo, M. Ko, Y. Lee, H. Chae, J. Shin, J. Jang, S. Ye, B. Y. Lin, S. Welleck, G. Neubig, M. Lee, K. Lee, M. Seo, The biggen bench: A principled benchmark for fine-grained evaluation of language models with language models, 2024. URL: https://arxiv.org/abs/2406.05761. arXiv:2406.05761.

[37] H. Huang, Y. Qu, H. Zhou, J. Liu, M. Yang, B. Xu, T. Zhao, On the limitations of fine-tuned judge models for llm evaluation, 2024. URL: https://arxiv.org/abs/2403.02839. arXiv:2403.02839.

A. Dataset

Table 4
EBA Q&As dataset. For this research, we focused on the fields highlighted in yellow.

Variable Name	Description
Question ID	The unique identifier for each question.
Topic	The general topic or category under which the question falls.
Subject matter	The specific subject matter of the question.
Legal act	The specific legal act to which the question relates. (e.g., CRR)
Article	The specific article of the legal to which the question relates.
COM Delegated or Implementing Acts/RTS/ITS/GLs/Recommendations	Other legislation, standards, guidelines or recommendations to which the question relates.
Article/Paragraph	The specific article or paragraph within the above-mentioned
Question	The actual question asked.
Background on the question	Any additional information or context provided by the question submitter.
Final answer	The official answer provided to the question.
Submission date	The date when the question was submitted.
Final publishing date	The date when the final answer to the question was published.
Status	The current status of the question (e.g. Final, rejected, etc.).
Type of submitter	The type of entity that submitted the question (e.g. Credit institution, investment firm, etc.).
Answer prepared by	The entity that prepared the answer to the question.

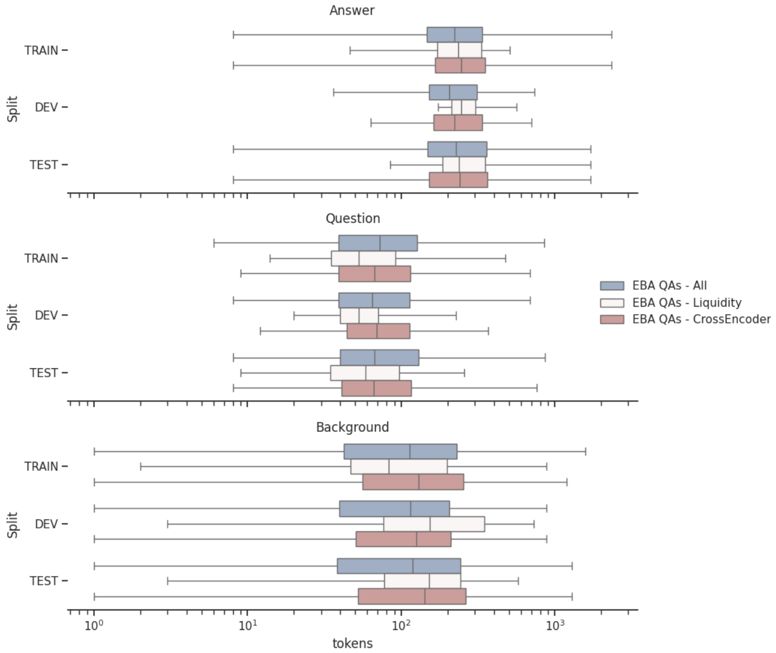

Figure 1: Distribution of tokens among Questions, Background, and Answers in datasets and splits

B. Multi-Step Generation and Evalutation

B.1. Multi-Step Approach for Answer Generation

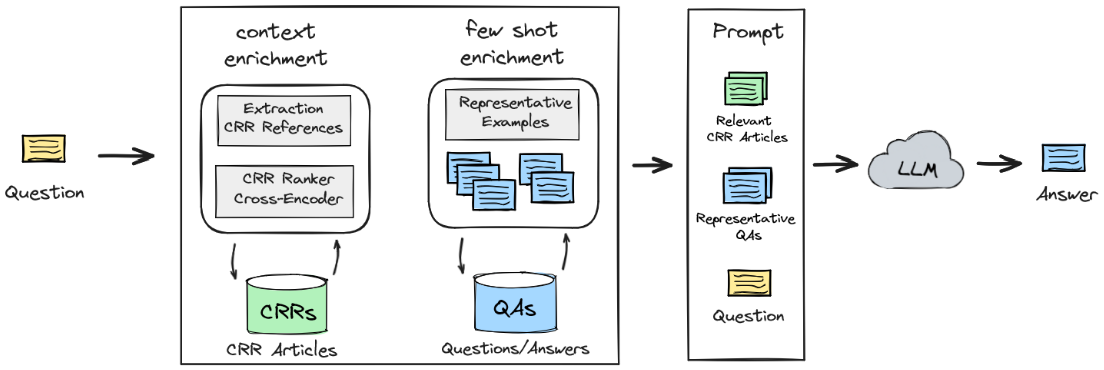

Figure 2: Multi-Step Approach for Answer Generation

B.2. LLM evaluator Alignment

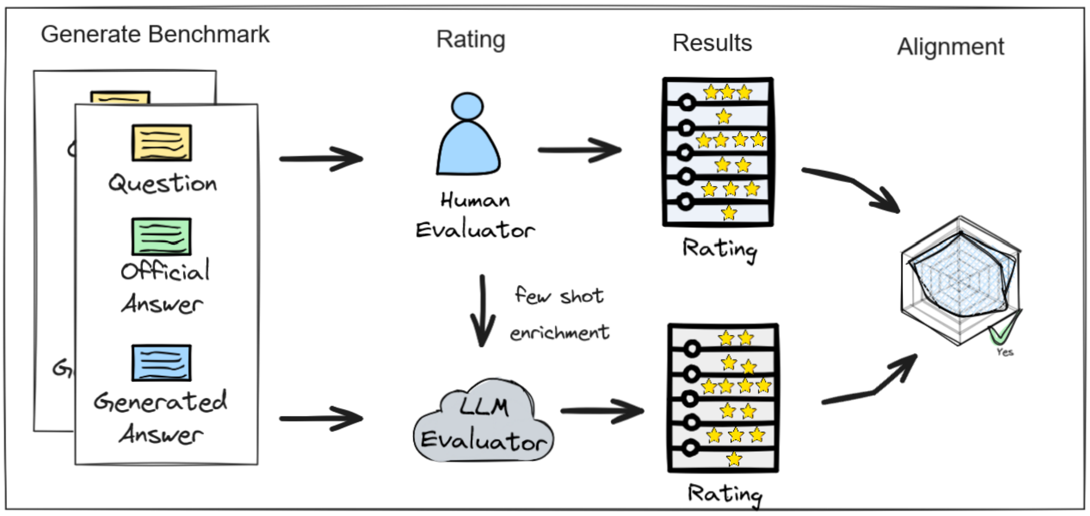

Figure 3: Evaluating Alignment between the LLM evaluator and the human expert

B.3. Multi-Step vs. Zero-Shot

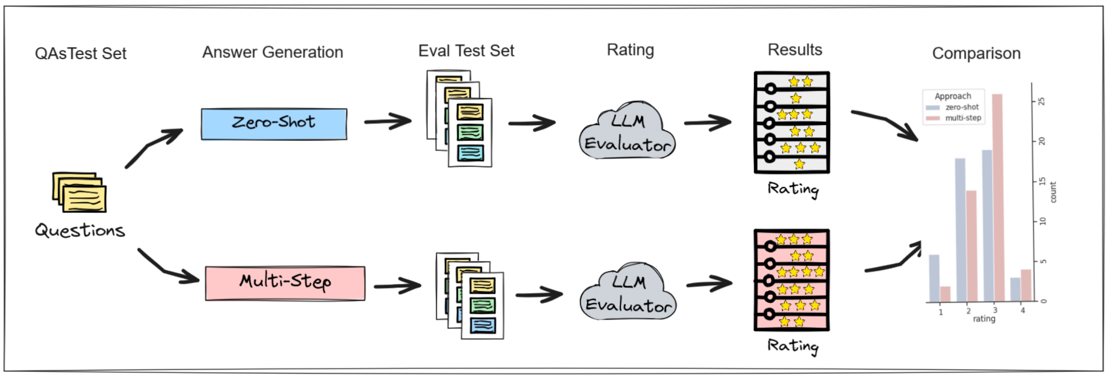

Figure 4: Multi-Step vs. Zero-Shot Approach for EBA Liquidity Risk Inquiries

C. Prompt template

C.1. Extracting Law References

Gpt4-omni Prompt

#task
Extract from the text (#text) any reference to regulatory documents contained in it and insert them into a list (e.g. ["regulatory document name": ["article 1","article 2",...]]). I will provide you an example (#text (example)) and the expected output (#output (example)):

#text (example) "In accordance with Article 425 (1) of Regulation (EU) No. 575/2013 (CRR) institutions may exempt contractual liquidity inflows from borrowers and bond investors arising from mortgage lending funded by covered bonds eligible for preferential treatment as set out in Article 129b (4-6) of CRR or by bonds as referred to in Article 52(4) of Directive 2009/65/EC from the 75% inflow cap."

#output (example) "["Regulation (EU) No. 575/2013 (CRR)": ["425","129b"], "Directive 2009/65/EC" : ["52"]]"
#text
> *text_to_extract*

#output (list only)

This prompt was used to extract any reference to regulatory documents from the provided text_to_extract) (placeholder to input text)

C.2. Answer Generation

Gpt4-omni Prompt

```
" #system
You are a virtual assistant for the European Banking Authority (EBA), handling user inquiries related to Liquidity Risk regulations. The user's query specifically pertains to Regulation (EU) No. 575/2013 (CRR) or Delegated Regulation (EU) No. 2015/61 (LCR DA)."""

#task
Answer the question based on the instructions below.
1. Analyze the User's Question (#question):
- Identify the central topic and relevant keywords related to Liquidity Risk and the specified EBA regulations.
2. Leverage the Provided Context (#context):
- Incorporate the context (including CRR articles and additional information) to tailor the answer to the user's specific scenario.
3. Liquidity Risk Topic:
- Reference relevant articles from provided context (#context) that address the specific aspect of Liquidity Risk raised in the question. 4. Desired Answer (#answer):
- Use only the information provided in the context and examples (if provided) to answer the question.
- Craft a well-reasoned and informative response that covers all aspects of the user's query.
- Clearly articulate the regulatory implications while considering the provided context.
- Maintain a professional and informative tone suitable for the EBA.

 #examples:
Example 1: > example_1
Example 2: > example_2
Example 3: > example_3
Example 4: > example_4
Example 5: > example_5

#question:
> question

#context:
> context
 > enhanced_context

#answer:
```

This prompt was used to generate answer given a question and context. #examples section (placeholder to include 5 examples) and enhanced_context (placeholder to include CRR articles), highlighted in yellow, were used only for multi-step approach.

C.3. LLM as Evaluator

Gpt4-omni Prompt

I will provide you with two answers to a question. One is the #official answer, which serves as the benchmark. The other is the #generated answer, which needs to be evaluated against the #official answer. You must compare the answers step by step.

Consider the following definitions for this evaluation:

- Correctness: A #generated answer is correct if its content aligns with that of the #official answer.
- Completeness: A #generated answer is complete if it includes all the information present in the #official answer.

Your task is to act as an evaluator and rate the #generated answer according to the following scale:

RATING 1: The #generated answer is completely incorrect and incomplete compared to the #official answer.
RATING 2: The #generated answer is incorrect but either complete or partially complete compared to the #official answer. It contains some useful information found in the #official answer but the main statement is incorrect.
RATING 3: The #generated answer is correct but only partially complete. The main statement matches the #official answer, but some information from the #official answer is missing.
RATING 4: The #generated answer is fully correct and complete. It is essentially a rephrased version of the #official answer with no significant differences.
Please provide a single numerical rating (1-4) followed by a brief explanation for your rating

<EXAMPLE 1>
...
<EXAMPLE 8>

Compute the score in the following case:

#question
> question

#background
> background

#official answer
> answer

#generated answer
generated answer

Output:

This prompt was used to compare an AI-generated answer (#generated answer) to an official one (#official answer), rating its correctness, completeness, and providing an explanation.