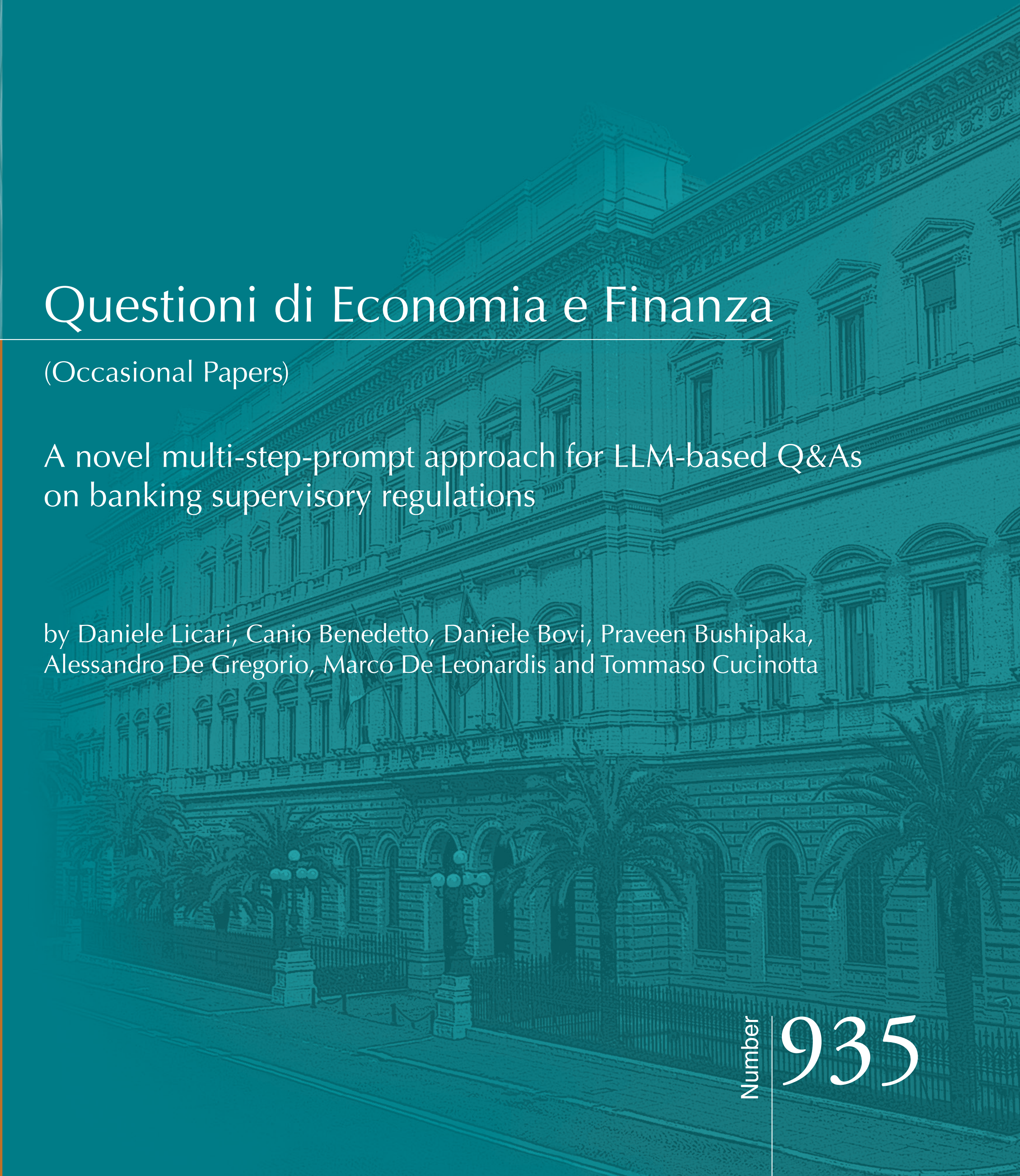
BANCA D'ITALIA
EUROSISTEMA



Questioni di Economia e Finanza

(Occasional Papers)

A novel multi-step-prompt approach for LLM-based Q&As on banking supervisory regulations

by Daniele Licari, Canio Benedetto, Daniele Bovi, Praveen Bushipaka, Alessandro De Gregorio, Marco De Leonardis and Tommaso Cucinotta

Number 935

April 2025

BANCA D'ITALIA
EUROSISTEMA

Questioni di Economia e Finanza

(Occasional Papers)

A novel multi-step-prompt approach for LLM-based Q&As on banking supervisory regulations

by Daniele Licari, Canio Benedetto, Daniele Bovi, Praveen Bushipaka, Alessandro De Gregorio, Marco De Leonardis and Tommaso Cucinotta

Number 935 – April 2025

The series Occasional Papers *presents studies and documents on issues pertaining to the institutional tasks of the Bank of Italy and the Eurosystem. The* Occasional Papers *appear alongside the* Working Papers *series which are specifically aimed at providing original contributions to economic research.*

The Occasional Papers *include studies conducted within the Bank of Italy, sometimes in cooperation with the Eurosystem or other institutions. The views expressed in the studies are those of the authors and do not involve the responsibility of the institutions to which they belong.*

The series is available online at www.bancaditalia.it .

ISSN 1972-6643 (online)

Designed by the Printing and Publishing Division of the Bank of Italy

A NOVEL MULTI-STEP-PROMPT APPROACH FOR LLM-BASED Q&AS ON BANKING SUPERVISORY REGULATIONS

by Daniele Licari*, Canio Benedetto*, Daniele Bovi*, Praveen Bushipaka**,
Alessandro De Gregorio*, Marco De Leonardis* and Tommaso Cucinotta**

Abstract

This paper investigates the use of large language models (LLMs) in analysing and answering questions related to banking supervisory regulations. We propose a multi-step-prompt approach that enriches the context provided to the LLM with relevant articles from the Capital Requirements Regulation (CRR). We compare our method against standard 'zero-shot' prompting, where the LLM answers are solely based on its pre-trained knowledge, and a standard 'few-shot' prompting, where the LLM is given only a limited number of examples of questions and answers to draw on each time. To assess the quality of the answers returned by the LLM, we also build an 'LLM evaluator' which, for each question, compares the correctness and completeness of the answer resulting from our multi-step prompt approach and from the two standard prompting methods with the official answer made available by the European Banking Authority (EBA), which is taken as a benchmark. Our findings on inquiries concerning Liquidity Risk rules indicate that our multi-step approach significantly improves the quality of LLM-generated answers, offering the analyst a valuable starting point to formulate appropriate answers to particularly complex questions.

JEL Classification: C45, C63, G21, G28.

Keywords: regulatory Q&A, banking supervisory regulation, Artificial Intelligence, GenAI, GPT-4o, RAG, LLM evaluator.

DOI: 10.32057/0.QEF.2025.935

* Bank of Italy.

** Sant'Anna School of Advanced Studies.

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

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

Abstract

This paper investigates the use of large language models (LLMs) in analysing and answering questions related to banking supervisory regulations. We propose a multi-step- prompt approach that enriches the context provided to the LLM with relevant articles from the Capital Requirements Regulation (CRR). We compare our method against standard 'zero-shot' prompting, where the LLM answers are solely based on its pre-trained knowledge, and a standard 'few-shot' prompting, where the LLM is given only a limited number of examples of questions and answers to draw on each time. To assess the quality of the answers returned by the LLM, we also build an 'LLM evaluator' which, for each question, compares the correctness and completeness of the answer resulting from our multi-step prompt approach and from the two standard prompting methods with the official answer made available by the European Banking Authority (EBA), which is taken as a benchmark. Our findings on inquiries concerning Liquidity Risk rules indicate that our multi-step approach significantly improves the quality of LLM-generated answers, offering the analyst a valuable starting point to formulate appropriate answers to particularly complex questions.

JEL Classification: C45, C63, G21, G28

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

* Bank of Italy

** Sant'Anna School of Advanced Studies

1. Introduction[1]

The advent of generative AI (GenAI), and specifically of large language models (LLMs), offers significant opportunities to implement innovative solutions to improve the efficiency and effectiveness of various domains of activities (Wu et al., 2023; Lai et al., 2023; Biancotti and Camassa, 2023; Horton, 2023; Homoki and Ződi, 2024).

One of the most promising applications is the support to the navigation and analysis of complex regulatory documents (Wiratunga et al., 2024; Louis et al., 2023; Zhang et al., 2023; Abdallah et al., 2023), which can be particularly valuable for compliance officers, legal teams, and other professionals, who need to have a clear and timely understanding of the regulations and the consequent obligations. This is, for example, the case of the EU harmonised banking supervisory reporting obligations, i.e. the set of rules defined by the European Banking Authority (EBA) about the information that the banking system is obliged to produce and transmit to the national competent authorities (NCAs). It is an articulated set of rules including the Capital Requirements Regulation[2] (CRR) and several delegated and implementing acts, technical standards, guidelines, and recommendations. In our work, however, for the sake of simplicity, we will only consider the CRR.

The complexity of such regulatory documents, with their dense network of cross-referenced texts and specialised content, is a challenge also for the same NCAs whose experts always necessitate a careful and deep analysis across the various acts to retrieve the needed information (Prenio, 2024). While LLMs offer advantages for this purpose, they also pose risks like bias and inaccuracies (Huang et al., 2023). Therefore, it is essential to establish strong verification procedures and retain human supervision to limit these risks.

This study introduces a novel methodology to automate and expedite the "question & answer" (Q&A) process in regulatory compliance, leveraging advanced LLMs to provide accurate, complete and timely responses to inquiries[3] usually presented by the banking sector about the EBA's supervisory regulations. Our multi-step approach aligns with Retrieval-Augmented Generation principles (Lewis et al., 2020) through the implementation of mechanisms like explicit extraction of CRR references, implicit reference analysis, and a dedicated cross-encoder for precise regulatory text retrieval. Our work finds application within the domain of EBA regulatory rules, that are characterised by a large and complex set of interrelated documents, although in this first research we will limit the focus the CRR.

Compared to standard inquiries of LLM engines, our approach improves the suitability of the response generation for the regulatory compliance context where precise and comprehensive answers are crucial, but the complex articulation of the legal acts makes the cross analysis very complicated and time consuming for the human expert. The business case is therefore both challenging and rewarding.

[1] The authors would like to express their sincere gratitude to Manuela Biagi, Vincenzo Capone, Francesca Monacelli, Roberto Sabbatini and Fabio Zambuto (Bank of Italy) for their insightful comments and stimulating discussions which greatly enhanced the clarity and focus of our work. A special thanks goes to Pamela Maggiori (Bank of Italy), whose careful analysis has significantly contributed to the development of the LLM evaluator.
The authors would also like to thank the anonymous reviewers of the CLiC-it 2024 Conference and the participants of the 4th IFC Workshop on Data Science in Central Banking for their invaluable suggestions and constructive feedback.
The views expressed in the paper are those of the authors and do not necessarily reflect those of the Bank of Italy.

[2] Regulation (EU) 2013/575: https://eur-lex.europa.eu/legal-content/en/ALL/?uri=celex%3A32013R0575

[3] For the purpose of this paper, we equally use the words "question" and "inquiry" to mean a request for clarification on the supervisory regulations presented by a domain expert.

Of all the possible topics covered in the CRR, in this work, we focus on inquiries related to 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 regulated by a relatively limited number of documents, predominantly within the CRR. This manageable scope facilitated the initial development and evaluation of our methodology, especially considering the time-intensive nature of pre-processing regulatory information.

We retrieve actual questions presented by banks and official EBA answers on the topic of Liquidity Risk from the "EBA Q&As database"[4] as the foundation for developing a novel system capable of generating automated accurate and complete (i.e. contextually relevant) responses to inquiries. It is important to emphasise that the questions contained in the "EBA Q&A database" are characterized to be all particularly complex and challenging by nature, as they have proven to be difficult issues to answer for the industry experts who submitted them, requiring an authentic interpretation of the regulations by EBA specialists. This means that we aim to build an LLM-based Q&A tool able to provide the results of high-quality analyses of supervisory regulations. That said, given the relevance in the field of legal acts interpretation of producing a correct and complete answer, in formal and practical terms we consider the output of the tool as a draft text for refinement for the experts and not a replacement of their analysis. The ambition of course is to obtain a draft text that needs as little adjustment as possible. Lastly, to assess the quality of the results obtained with the tool, we leverage recent studies on the potential of LLMs for qualitative assessment (Ye et al., 2024; Zheng et al., 2023; Liu et al., 2023; Chan et al., 2023) and, 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 description of the approach adopted; it explains the dataset used and the retrieval techniques employed to identify the relevant regulatory documents necessary to address the questions on Liquidity Risk. Chapter 3 presents the LLM Evaluator and the evaluation criteria. Chapter 4 reports the experimental results and presents the main outcomes of the study. Chapter 5 discusses the challenges as well as the potential areas for future developments.

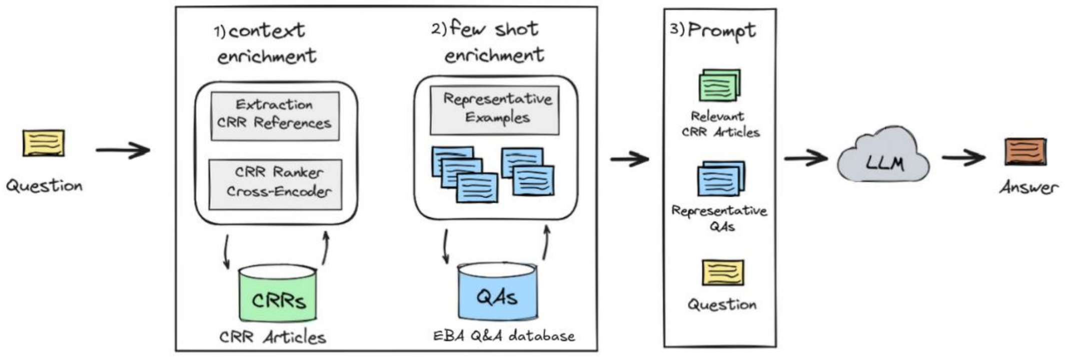
2. Methodology

This research employs a multi-step methodology to construct a comprehensive prompt for the GPT-4 omni 2024-11-20 version (GPT-4o) language model (OpenAI et al., 2024), to appropriately enrich questions on EBA's banking supervisory rules concerning Liquidity Risk, as contained in the EBA Q&A database, in order to obtain an answer resembling the official EBA answer (also contained in the EBA Q&A database) as much as possible.

Our stepwise approach focuses on enriching the context provided with the question in three steps: first, it identifies the relevant CRR references specified in the inquiry; second, it incorporates examples of other Q&As to ensure that the LLM's output format is aligned with the EBA's (e.g. lexicon, style, structure of the answer); third, this enriched context is then leveraged by GPT-4o to generate the answer (details in Figure 1).

[4] Stakeholders can submit questions to the EBA on the practical application or implementation of the banking, payment services, AML/CFT and other legislation that falls within the EBA's remit. This includes the associated delegated and implementing acts, RTS, ITS, guidelines and recommendations.
EBA Single Rulebook Q&A: https://www.eba.europa.eu/single-rule-book-qa

Figure 1: Multi-Step Approach for Answer Generation



2.1. Dataset construction

As a preliminary step, we extract from the EBA Q&A database, all the pairs of question and answer submitted between 2013 and 2020 (to consider Q&As referring to a more stable regulatory framework, before the introduction of CRR2). We focused on the following variables: question ID, question, submission date, status, topic, legal act, article, background information, final answer, submission date and status (details in Table 1). Given our specific focus on CRR articles, we excluded variables such as 'COM Delegated or Implementing Acts/RTS/ITS/GLs/Recommendations' and the related and more granular 'Article/Paragraph'. While 'Subject matter' offers a concise headline for the question-answer, we prioritised the richer semantic context provided by 'question' and 'background information' for our retrieval and answer generation process. Also 'Final publishing date', 'Type of submitter', and 'Answer prepared by' were deemed less relevant for this purpose and were therefore excluded."

Table 1

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.

Secondly, we implemented a two-step filtering process aimed at strengthening the model efficacy: by excluding non-English questions, and by focusing on CRR-related questions.

This resulted in a final dataset of 1,597 CRR-related pairs of question and answer, which was then split into training (50%), validation (10%), and test set (40%) for evaluation.

The distribution of samples for the dataset is summarized in Table 2.

Table 2
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	(percentage)
Training	798	58	50%
Validation	162	12	10%
Test	637	46	40%
Total	1597	116	100%

2.2. Context Enrichment

The context enrichment process is a three-step approach designed to identify, within the dataset built as indicated in Paragraph 2.1, the most relevant CRR references to provide an appropriate background to formulate the answer to the inquiry. The first step simply involves extracting explicit CRR references, if directly mentioned in the question (content of the variable "Article" in Table 1). The second step leverages on the capabilities of GPT-4o (see "Prompt 1" in Appendix) to analyse the content of the variables "Question" and "Background information" to identify other CRR references that are not explicitly stated in the question. The last step of the process makes use of our "CRR Ranker" model, i.e. a cross-encoder architecture that has been trained to identify and retrieve pertinent references from the CRR in response to specific inquiries. This 3-steps comprehensive approach provides a broader and more accurate context to what is included in the original question, thus improving the possibility for the model to "understand" the question and identify the specific CRR references deemed to be relevant.

2.2.1. CRR Ranker training

With regard to the last step of the context enrichment described in the paragraph above, we employ a specifically trained cross-encoder model (Chen et al., 2024) to identify the most relevant CRR references to be considered. We use a dedicated dataset of "question-article" pairs derived from our EBA Q&A Train Database, from which we exclude questions related to CRR Article 99 being this irrelevant for the purpose of our study. Each data point consists of a question (including the related background information - see Table 1), the associated CRR article, and a binary label indicating relevance.

We constructed the training dataset by selecting positive samples, consisting of pairs of questions and articles where the article is deemed relevant to the corresponding question, and negative samples, consisting of pairs where the article is unrelated to the question:

- Positive samples comprise question-article pairs where the article is explicitly linked to the question (see variable Article in Table 1). Additionally, we include pairs formed by questions and other CRR references implicitly contained in the text, context information, and EBA's final answer, that we extract using GPT-4o (see "Prompt 1" in Appendix);
- Negative training samples are mined using the BAAI bge-large-en-v1.5 pre-trained language model (Xiao et al., 2023) employing a two-phase process. First, all CRR articles are encoded using the bge-large-en-v1.5 model, and cosine similarity is used to rank them relative to a question; second, a subset of 20 negative examples was randomly chosen from a pre-defined ranking interval[5] (250-300). The choice of 20 negative samples provides a good balance between computational efficiency and the availability of a sufficient number of training data. This approach aimed to balance the representation of relevant and irrelevant information within the training data, ensuring the model learns to select articles that are actually related to the question and discard those that, although semantically similar, are ultimately off-topic (Xuan et al., 2021).

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

We select the *BAAI BGE Reranker v2 m3* model (Chen et al., 2024) as the basis for our cross-encoder, owing to its task-specific aptness and its demonstrated superior performance relative to the BGE Reranker Large (Xiao et al., 2023), as reported in Section 4. We adopt the Cross-Entropy Binary Classification loss function, following the approach suggested in the BGE Rerank Git repository (Xiao et al., 2024). To promote stable convergence, we incorporate a warmup schedule (with a number of *steps 0.1 × len(train_data) × 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 evaluate the model's ability to retrieve CRR articles for a given question present in the Q&A Test Database. This evaluation employes recall metrics at various retrieval cutoffs, including recall@5, recall@10, recall@20, and recall@30 (see the results in Section 4).

2.3. Enriching the prompt with examples

To improve the model's understanding of the desired format to be used to formulate the answer, we adopt a few-shot prompting approach (Brown et al., 2020). This involves extracting 5 relevant examples of "question and answer" pairs from the EBA Q&A Train Database concerning other similar issues within the topic of liquidity risk. These examples serve as demonstrations for the tool, showcasing the ideal structure, language style, and level of detail expected in the final answers. Notably, the selection process ensures variety within the chosen topic, meaning the examples cover various aspects to promote a broader understanding of the expected communication standards. Limiting the number of examples to five strikes a balance

[5] Of the overall ~500 articles, we chose the 250-300 range to select negative examples that are not obviously irrelevant but also not highly similar to the question. This "middle ground" provides challenging negative examples that help the model learn subtle differences between relevant and irrelevant content.

between providing diverse demonstrations and maintaining cost-efficiency during inference, as the LLM's input token length is limited.

2.4. Improving the prompt to enhance answer generation

To generate accurate and contextually appropriate answers, we employ an advanced prompt engineering strategy leveraging GPT-4o's capabilities as detailed in Figure 1. Our approach integrates three techniques to enhance answer generation within a regulatory compliance framework (see "Prompt 2" in Appendix).

Firstly, we implement a role-based contextualization method that aligns the model to specific regulatory requirements and professional communication standards. Secondly, we use step-by-step reasoning to decompose complex queries into logical, hierarchical components. Finally, we employ few-shot prompting by carefully selecting and integrating examples from our training dataset as a contextual guide (as detailed in Paragraph 2.3).

3. LLM Evaluator

To assess the quality of the answers returned by the tool (following the steps illustrated in Section 2), we developed an LLM Evaluator that compares the answers produced by the tool with the official ones of the EBA "Q&A" database. Recent research highlights that employing an LLM Evaluator offers significant advantages in terms of cost-effectiveness and efficiency compared to the traditional human evaluation method when dealing with large-scale natural language evaluation tasks (Liu, Iter et al., 2023; Dubois et al., 2024; Fu et al., 2023).

The evaluation process uses a scale from 1 to 4, based on two evaluation criteria: correctness and completeness. A generated answer 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.

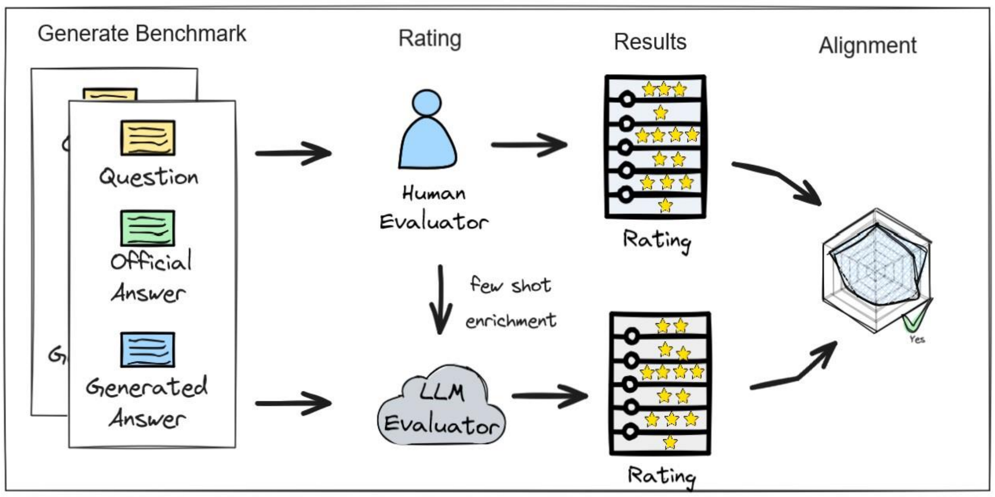
The following scoring rubric outlines the evaluation criteria:

- **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 that of the LLM was then analysed (see "Prompt 3" in Appendix for the complete details of the final prompt used for the LLM evaluator).

The synthetic 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 the LLM Evaluator. Using GPT-4o, we obtained a Kendall-tau coefficient of 0.86, with a p-value of 6.23×10^{-11}. These results justified the adoption of the LLM Evaluator, especially for tasks involving prompt optimization. Figure 3 illustrates the complete process of evaluating the alignment between the LLM Evaluator and the human expert.

Figure 3: Evaluating the alignment 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. The first is measured by the number of relevant articles retrieved (i.e. the recall) using different encoder models, the second is then evaluated by the LLM Evaluator. We compare the multi-step prompt approach with a few-shot and a zero-shot one focusing on a single topic within the EBA Q&A framework, specifically Liquidity Risk. Finally, we benchmark our multi-step pipeline using other LLM models, such as Google Gemini Flash 1.5 and Llama 3.1 70B.

4.1. Measuring the relevance of the CRR articles retrieved by the tool

We employed the “recall” as the primary metric to assess the performance of bi- and cross-encoder models in retrieving relevant CRR articles based on the information submitted with the question. The recall at N (r@N) is defined as the proportion of relevant CRR articles over the first N retrieved by the model (Manning et al., 2008). In the context of legal information retrieval, prioritizing the retrieval of all regulatory information pertinent to the question makes the recall a particularly relevant metric. Our primary objective was to identify a model that delivers high 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 an effectiveness comparison between our fine-tuned CRR Ranker and several other pre-trained models:

- Bi-encoders: *all-MiniLM-L6-v2* (Lewis et al., 2021), *gte-large-en-v1.5* (Li et al., 2023), and *bge-large-en-v1.5* (Xiao et al., 2023).
- Cross-encoders: *bge-reranker-large* (Xiao et al., 2023), *bge-reranker-v2-m3* (Li, Zhang et al., 2023; Chen et al., 2024).

The detailed results on the EBA Q&As Test Database are presented in Table 3. Our fine-tuned CRR Ranker significantly outperformed all other models, achieving a more than 20% improvement compared to the best pre-trained model tested (*bge-large-en-v1.5*).

Table 3
Recall scores on EBA Q&As Test Dataset

Model	r@5	r@10	r@20	r@30
bge-reranker-large	0.17	0.23	0.31	0.38
bge-reranker-v2-m3	0.24	0.31	0.39	0.44
all-MiniLM	0.37	0.46	0.55	0.59
gte-large	0.39	0.48	0.57	0.63
bge-large-en-v1.5	0.41	0.52	0.62	0.67
CRR Ranker (ours)	**0.51**	**0.67**	**0.81**	**0.86**

4.2. Comparing our multi-step approach with standard ones

This paragraph presents the comparison of our multi-step approach against a zero-shot baseline for answering EBA liquidity risk questions. Our evaluation focuses on a subset of 46 liquidity risk-specific Q&As taken from the EBA Q&A Test database. We compared the answers produced by the multi-step approach described in the Section 2 (where the LLM receives a prompt containing the query, the enriched context and the examples) with the ones produced by a:

- **Zero-Shot approach:** the LLM receives a standard prompt containing just the query;
- **Few-Shot approach:** the LLM receives a prompt containing the query and some examples.

The quality of the answers produced by the three approaches is assessed by our LLM Evaluator, which scores each answer based on correctness and completeness relative to the official EBA one, using a scale of 1 (fully incorrect and incomplete) to 4 (fully correct and comprehensive), as described in Section 3.

To optimize the multi-step approach, we conduct experiments focusing on:

- **Optimizing the top-k parameter:** we investigate the impact of the number of retrieved articles (top-k) on answer quality. Using the LLM Evaluator, we experiment with different values of k to identify the optimal number for achieving the best performance.
- **LLM Example filtering and re-ranking:** we explore incorporating LLM-based filtering and re-ranking techniques (Lee and Roh, 2024; Chang et al., 2024) to further refine the retrieved information and the example selection.

Table 4 summarizes the evaluation results for answers generated by different approaches. Our findings demonstrate the significant superiority of the multi-step approach over “zero-shot” and “few-shot” methods in terms of answer correctness. Specifically, the "Multi Step - top 5" method, which retrieves the top 5 most relevant articles using the CRR Ranker, achieves the highest average LLM score (2.63 ± 0.90) and the highest percentage of correct answers (56.5%). This represents a significant improvement over the “zero-shot” and “few-shot” approaches, with average scores increasing by 14.4% and 3.5% respectively (from 2.30 and 2.54 to 2.63); more importantly, the percentage of correct answers increases by 23.8% and 30% respectively (from 45.7 and 43.5 to 56.5).

This underscores the importance of combining advanced retrieval techniques with a well-designed answer generation prompt.

Table 4 Evaluation results for responses generated by zero-shot, few-shot and multi-step.

Approach with GPT4o	AVG Score	# Correct (score>2)	% Correct
Zero Shot	2.30 (±0.89)	21	45.7
Few Shot	2.54 (±1.00)	20	43.5
Multi Step - top 5	**2.63 (±0.90)**	**26**	**56.5**
Multi Step - top 10	2.50(±0.94)	23	50
Multi Step - top 20	2.48(±0.86)	23	50
+ LLM-Reranker top 5	2.50(±0.89)	23	50
+ LLM Examples filter	2.48(±0.96)	24	53.3

It is important to highlight that the potential negative impact of incorrect information in this critical domain makes “correctness” the primary dimension for an immediate usage of the response. On the other hand, we argue that the “completeness” of the response serves as a key indicator of the tool’s overall usefulness, as it ensures the provision of a comprehensive set of relevant references that support the analyst’s assessment. In addition, given the way the evaluation scale is defined, achieving a score of 4 is extremely difficult, as it requires the tool’s response to be nearly identical the one provided by the EBA or its paraphrase. This is very challenging, as there may well be different equally valid variations to the same official answer, each varying depending on the specific sensitivity and expertise of the analyst producing the answer.

Our results also reveal a crucial caveat: enriching the context with too much information can introduce irrelevant or weak references, negatively impacting the quality of the answer. This emphasises the need to balance context enrichment with noise reduction. That is why, while it is fundamental to retrieve relevant information, ensuring its quality is equally important to prevent hindering the LLM's reasoning and answer generation capabilities. In this regard, employing an LLM reranker to select the most relevant articles (see “Prompt 4” in Appendix) and filtering out irrelevant examples (see “Prompt 5” in Appendix) improves the correctness of the answer compared to the "Multi Step - top 20" baseline.

These findings suggest that the context enrichment of our multi-step prompts effectively guides the LLM toward generating more comprehensive and informative answers.

4.2.1. Other LLMs

Our multi-step pipeline is also tested using other large language models (LLMs), specifically *Google Gemini Flash 1.5* and *Llama 3.1 70B*. The first is widely recognized for its high-speed processing capabilities and efficiency in response generation, making it a suitable benchmark for comparative performance analysis. Conversely, the second 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* achieved 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 the inquiries in our dataset.

Future research will focus optimizing each step of the multi-step pipeline in a model-specific manner to further enhance the overall accuracy and reliability of the generated answers.

5. Challenges and Advancements

In our work we were confronted with several challenges. One of the main issues was the limited size of our test dataset since, as already mentioned, we focused on the single topic of Liquidity Risk, to limit the very time-consuming preprocessing phase of regulatory documents.

However, to achieve a robust alignment with the results obtained by the human expert and ensure that the tool addresses diverse inquiries across other EBA Q&A topics, future efforts should prioritize dataset expansion and human evaluation integration. The study also emphasizes the need to retrieve relevant CRR articles to properly enrich the context but warned against the risk of gathering too much information.

Future research could also investigate methods to further refine the format of the generated answers by incorporating legal reasoning and argumentation capabilities into the LLM (Yu et al., 2023; Lu and Kao, 2024), and Q&As selected case by case as relevant examples for few-shot prompting (Wiratunga et al., 2024).

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 will focus on investigating automatic prompt engineering techniques (Ye et al., 2024). Moreover, we intend to extend our study to test other models that have demonstrated similar performance levels as GPT-4o in the field of open question answering (Huang et al., 2024). This will help us identify the most effective model for our application (Panickssery et al., 2024).

Similarly, in the context of LLM evaluators, we also intend to explore additional models, including open-source options (Kim, Suk et al., 2024; Kim, Suk, Cho et al., 2024), that have shown strong performance in assessing the quality of outputs of other LLMs. This approach could furtherly increase the correlation between human and LLM evaluations, thereby enhancing the tool'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 (Huang, Qu et al., 2024).

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 tool's effectiveness and generalizability across a wide range of regulatory domains.

6. Conclusion

This study explores a novel approach for generating answers to questions on the Regulation (EU) 2013/575, specifically on the liquidity risk topic. We propose a multi-step prompt construction method that enriches the context provided to an LLM, enabling it to generate more accurate and informative answers. An LLM Evaluator, which demonstrated strong agreement with human experts, is employed to compare our multi-step approach with standard zero-shot and few-shot methods that lack context enrichment. The quality of the answers returned by the LLM is assessed, and our findings indicate that the multi-step approach significantly outperforms both the other methods, resulting in more comprehensive and accurate answers. 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. Specifically, it is worth noting that the CRR Ranker enables the retrieval of most relevant documents in a fast and accurate manner, significantly aiding in the process of addressing a question. Even at this early stage, the tool has demonstrated its ability to make the work of the human analyst more efficient. We believe that answers with a quality score of at least 2 already provide a useful starting point for addressing complex regulatory questions. The results obtained with our tool provide a solid basis for the human expert (who should always stay “in the loop” when dealing with sensitive issues such as legal requirements), reducing the time needed to analyse the issue behind the questions.

The development of a more precise metric to quantify usefulness would require the involvement of additional experts in supervisory regulation, which would allow a more accurate assessment of the correlation between the score and perceived usefulness in improving process efficiency. Specifically, this could be achieved through a dedicated internal survey among the domain experts to quantitatively assess the “perceived usefulness” of the tool, for instance, by requesting to evaluate, for each pair of question and answer the estimated reduction in the time required to analyse the question and produce the answer with and without the tool.

Finally, our research highlights the advantages of using public cloud infrastructures for LLM research and development. This approach enables efficient testing of diverse LLMs, resource and cost optimization, proving particularly suitable for projects utilizing public data. On the other hand, even when less desirable in terms of cost, powerful on-premises solutions remain preferable, if not the only ones, for managing sensitive data.

References

Abdallah, A., 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.

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

Brown, T. B., 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.

Chan, C.-M., 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].

Chang, C.-Y., Z. Jiang, V. Rakesh, M. Pan, C.-C. M. Yeh, G. Wang, M. Hu, Z. Xu, Y. Zheng, M. Das, N. Zou, Main-rag: Multi-agent filtering retrieval-augmented generation, 2024. URL: https://arxiv.org/abs/2501.00332. arXiv:2501.00332.

Chen, J., 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.

Dubois, Y., 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.

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

Homoki, P., 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.

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

Huang, H., 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.

Huang, L., 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.

Huang, Z., 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.

Kim, S., 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.

Kim, S., 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.

Lai, J., 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].

Lee, J., M. Roh, Multi-reranker: Maximizing performance of retrieval-augmented generation in the financerag challenge, 2024. URL: https://arxiv.org/abs/2411.16732. arXiv:2411.16732.

Lewis, P., E. Perez, A. Piktus, F. Petroni, V. Karpukhin, N. Goyal, H. Küttler, M. Lewis, W. Yih, T. Rocktäschel, S. Riedel, D. Kiela, Retrieval-augmented generation for knowledge-intensive NLP tasks, 2020. In Proceedings of the 34th International Conference on Neural Information Processing Systems (Vancouver, BC, Canada). URL: https://proceedings.neurips.cc/paper_files/paper/2020/file/6b493230205f780e1bc26945df7481e5-Paper.pdf.

Lewis, P. S. H., 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.

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

Li, Z., 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).

Liu, Y., 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.

Liu, Y., 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.

Louis, A., 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].

Lu, Y. an, 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.

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

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].

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

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

Wu, S., 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].

Wiratunga, N., 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.

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

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

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

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

Ye, S., 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].

Yu, F., 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.

Zhang, W., 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.

Zheng, L., 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].

Appendix

Prompt 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)

Prompt 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.

Prompt 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.

Prompt 4 – Rerank

Gpt4-omni Prompt

You are RankGPT, an intelligent assistant that can rank passages based on their relevancy to the query.

I will provide you with num passages, each indicated by number identifier [].

Rank the passages based on their relevance to query: {query}. Search
Query: {query}

Rank the {num} passages above based on their relevance to the search query.

The passages should be listed in descending order using identifiers. The most relevant passages should be listed first.
The output format should be [] > [] > [] > [] > ..., e.g., [1] > [2] > [3] > [4] > ...

Only response the ranking results, do not say any word or explain.

Prompt 5 Examples Filter Prompt

Gpt4-omni Prompt

You are a virtual assistant for the European Banking Authority (EBA), responsible for analyzing inquiries related to Liquidity Risk regulations under Regulation (EU) No. 575/2013 (CRR) and Delegated Regulation (EU) No. 2015/61 (LCR DA).

Your task is to filter out irrelevant examples provided by the user.

Follow these instructions to determine which examples are not useful for addressing the user’s specific question.

1. Understand the user’s question (#question) by identifying its core topic, keywords, and references to relevant regulations or concepts.
2. Analyze the provided context (#context), including the operational details and CRR articles referenced, toclarify the regulatory framework applicable to the question.
3. Review the examples (#examples), which are numbered from 1 to 5 and contain separate Q&A entries,each with its own context and answer.
4. Evaluate the relevance of each example by checking if it directly contributes to answering the user’squestion based on: - Relevance to the regulatory topic or specific articles mentioned in the question. - Applicability of the example’s context to the user’s scenario. - Alignment with the CRR or LCR DA framework relevant to the question.
5. For each example, determine if it is irrelevant and briefly justify why it does not provide useful informationfor the specific question.

6 Output a list of the relevant examples by their number (do not provide any short justification but only the list of number).

#question:
question

#context:
context

#examples
examples