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

Daniele Licari[1], Canio Benedetto[1], Daniele Bovi[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 - (name.surname@bancaditalia.it)

[2]Scuola Superiore Sant'Anna, P.zza dei Martiri della Libertà, 33, Pisa, 56100, Italy – (name.surname@santannapisa.it)

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

Outline

Introduction

1. Motivation
2. EBA Single Rulebook Q&A
3. Dataset
4. Challenges in Q&A System Dev
5. Methodology Overview

Methodology

1. 3-Step CRR Articles Retrieval
2. Answer Generation Process
3. LLM Evaluator
4. Results and Analysis
5. Conclusion and Future Work

Motivation

- **Challenges in Regulatory Compliance**
  Regulatory documents are complex, with dense cross-references and specialized content, making manual analysis time-consuming and error-prone.

- **Need for Efficient Solutions**
  Compliance professionals and supervisory authorities require tools to streamline the navigation and interpretation of these regulations.

- **Explore the potential of LLMs**
  Large Language Models (LLMs) can serve as powerful assistants, helping users understand regulatory frameworks through appropriate tools while simplifying and streamlining work-related tasks as **Question & Answer** (Q&A) systems.

- **Proposed Methodology**
  This study uses a **Retrieval-Augmented Generation (RAG)** like framework using LLMs to automate and enhance the Q&A process for regulatory compliance using the **European Banking Authority (EBA) Single Rulebook Q&A**.

European Banking Authority (EBA) Single Rulebook Q&A

- 1,597 Q&A pairs covering regulatory topics in European banking (2013-2020).
- Provides clarification on the application of complex banking regulations.
- Supports compliance professionals in understanding legal obligations.
- Each Q&A contains extensive information. For the development of our system, the following fields proved particularly useful:
  - **Background of the question**: any additional information or context provided by the question submitter.
  - **Question**: the actual question being asked.
  - **Answer**: the official answer provided to the question.
  - **Submission date**: the date the question was submitted, used to determine the regulatory documents valid at that time.
  - Specific references provided by the user, including **Article**, **Paragraph**, **Subparagraph**, **COM Delegated or Implementing Acts/RTS/ITS/GLs/Recommendations**, or other relevant legislation, standards, guidelines, or recommendations related to the question.

Legal act	Regulation (EU) No 575/2013 (CRR)
Topic	Transparency and Pillar 3
Article	449a
Paragraph	-
Subparagraph	-
COM Delegated or Implementing Acts/RTS/ITS/GLs/Recommendations	Regulation (EU) 2022/2453 - ITS on ESG disclosures
Article/Paragraph	1
Date of submission	30/01/2023
Published as Final Q&A	25/08/2023
Disclose name of institution / entity	No
Type of submitter	Competent authority
Subject matter	ESG P3 - Reg 2453/22 - Scope of application
Question	What is the scope of application of Regulation (EU) 2022/2453?
Background on the question	Article 449a CRR states that: "large institutions which have issued securities that are admitted to trading on a regulated market of any Member State, as defined in point (21) of Article 4(1) of Directive 2014/65/EU, shall disclose information on ESG risks". In the EBA/ITS/2022/01 we found the following clarification: "Institutions should disclose the information at the highest level of consolidation consideration in the EU, as regulated in Article 13 CRR." (page 52) Due to the above, there is different understanding of the scope of application of the ITS on Pillar 3 ESG disclosures, both between banks as well as the regulator. Shall all large listed institutions in a jurisdiction disclose the P3 ESG information (this means disclosure on subconsolidated basis for some banks) or only large listed institutions which are the EU highest parent institutions (this significantly reduces the number of banks subject to the Reg. 2022/2453 in PL).
Final answer	As clarified by the EBA/ITS/2022/01, institutions should disclose the information at the highest level of consolidation in the EU, as regulated in Article 13 CRR. For the detailed information on scope of application of Article 449a CRR please refer to QA2022_6652.
Link	https://www.eba.europa.eu/single-rule-book-qa/qna/view/publicId/2023_6708

Q&A Dataset

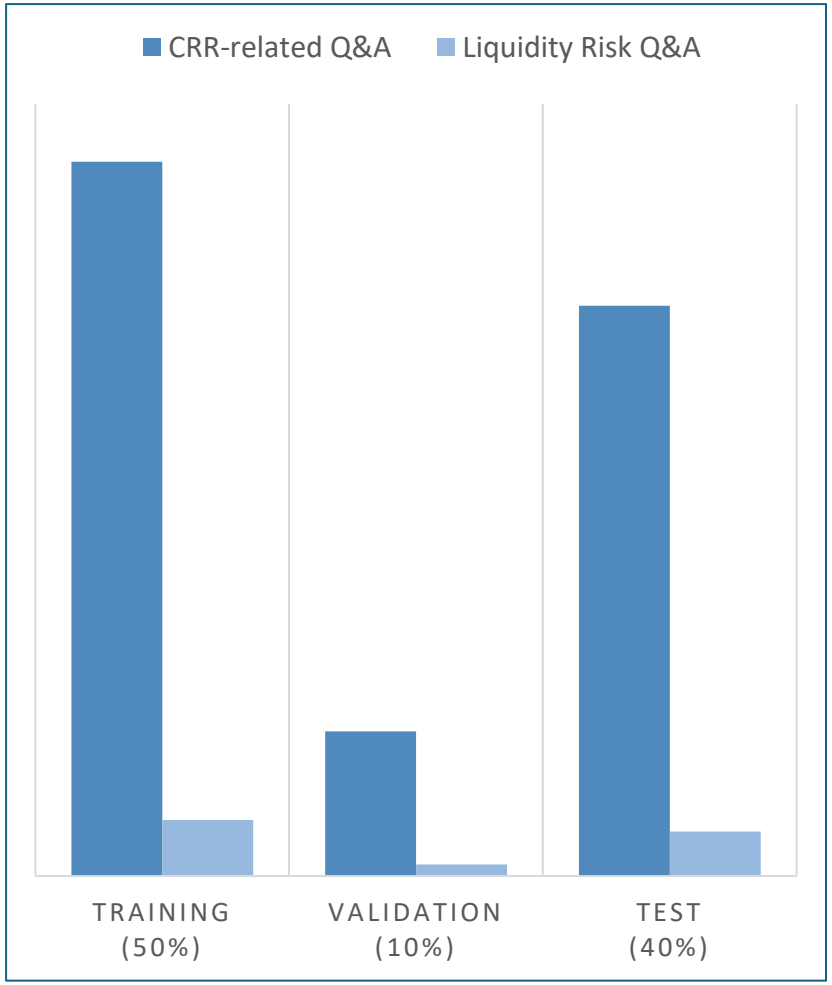
- Focus Topic: **Liquidity Risk (**112 Q&A).

- Dataset split:

  - **Training** Set: used to train and optimize the RAG phase.
  - **Validation** Set utilized for hyperparameter optimization during RAG development.
  - **Test** Set: used to evaluate the final model performance.

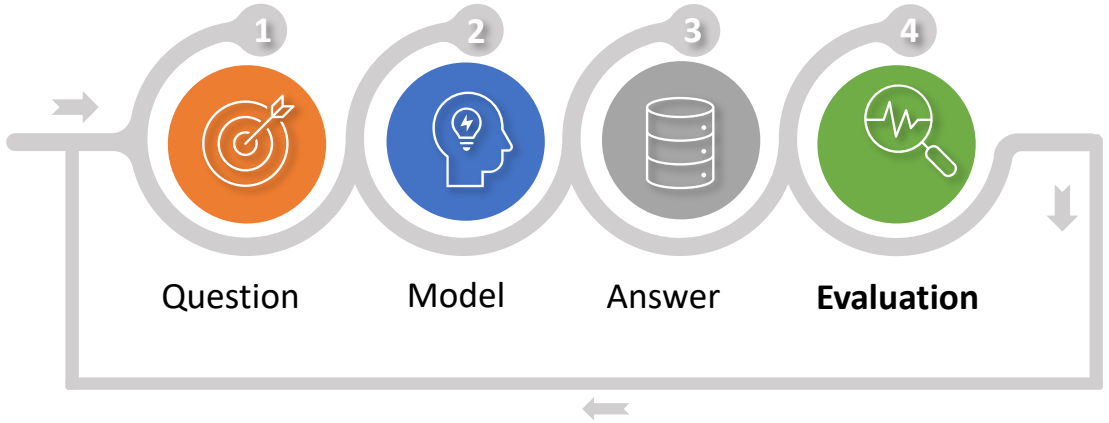
- Additional relevant documents: **Capital Requirements Regulation (CRR)** EU No. 575/2013: core banking framework counting around 500 articles.

- Planned additional documents: relevant materials to be included, such as
  - **Liquidity Coverage Requirement (LCR) Delegated Regulation**
  - **Implementing Technical Standards on Supervisory Reporting**



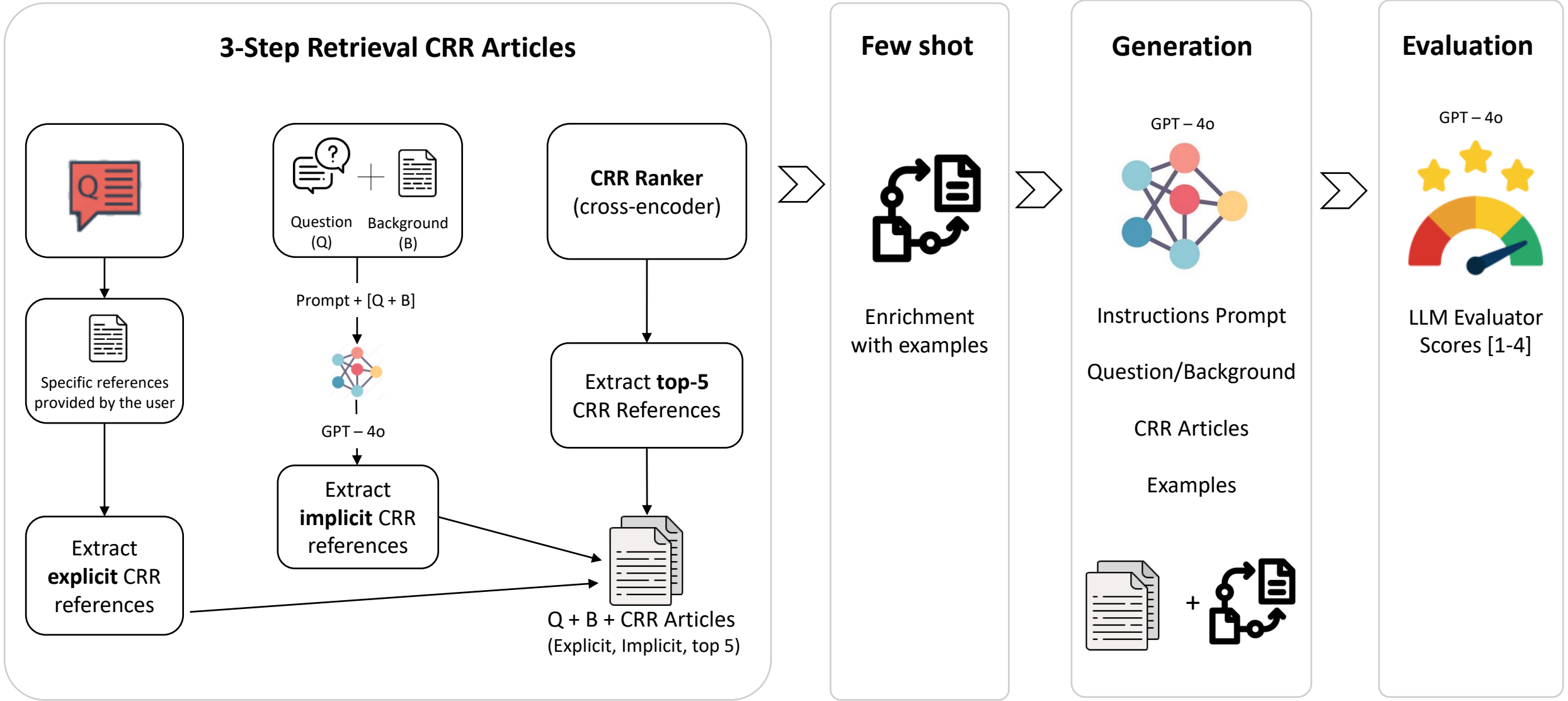
Challenges in Q&A System Development

- **Development Challenge**: Testing and developing a Q&A system requires repeated testing to identify the best approach.



- **Evaluation Phase**: Accurate evaluation is essential but often very **time-consuming** and **tedious**.
- **LLMs as evaluators**: LLMs act as automated evaluators, enabling **scalable**, **interpretable**, and **cost-efficient** assessments. By automating repetitive tasks, they save resources while ensuring consistent quality evaluation.
- **Challenges to Consider**: LLM judgments must align with human reasoning, which is ensured through an initial calibration and verification phase.

Methodology Overview



Outline

Introduction

1. Motivation
2. Challenges in Q&A System Dev
3. EBA Single Rulebook Q&A
4. Dataset
5. Methodology Overview

Methodology

1. 3-Step CRR Articles Retrieval
2. Answer Generation Process
3. LLM Evaluator
4. Results and Analysis
5. Conclusion and Future Work

3-Step Retrieval CRR Articles: Context Enrichment

Question ID	2016_2731
Legal act	Regulation (EU) No 575/2013 (CRR)
Article	411
Question	To be eligible for Article 153(4) CRR, do firms need to have externally verified annual sales numbers, and if so how frequently should firms assess eligibility?
Background	This question is on the application of a definition in the context of the LCR.

Q&A

1

A: 411

Q + B

2

LLM-Based Extraction of CRR References

#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)

A: 153

CRR Articles

A: [153, 411]

3

Ranking Model

CRR Ranker

Top K

Article 153, 411

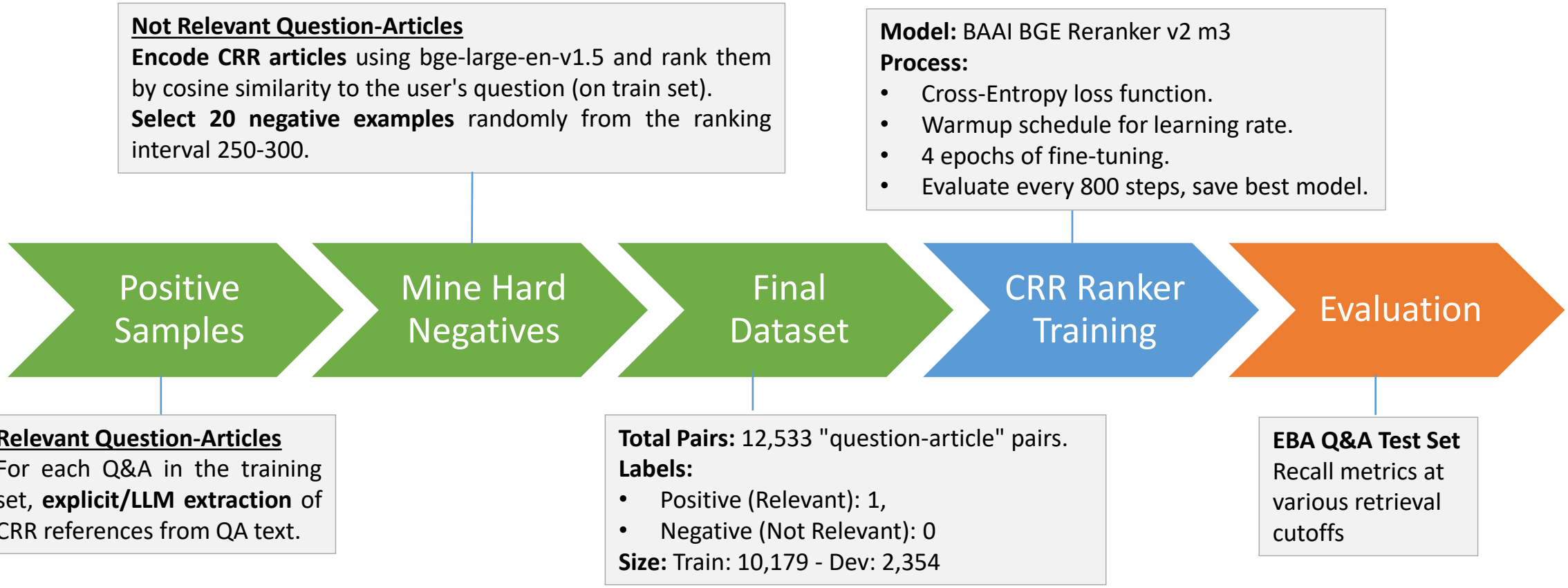
Relevant CRR Articles

Three-Step Approach:

1. **Explicit Extraction**: Directly extract CRR references mentioned in the Q&A.
2. **LLM Extraction**: identify other relevant CRR references from the *Question* and *Background* information using GPT-4o .
3. **Ranking Model**: retrieve the **top *K*** relevant CRR articles based on their similarity to the *Question* and *Background*

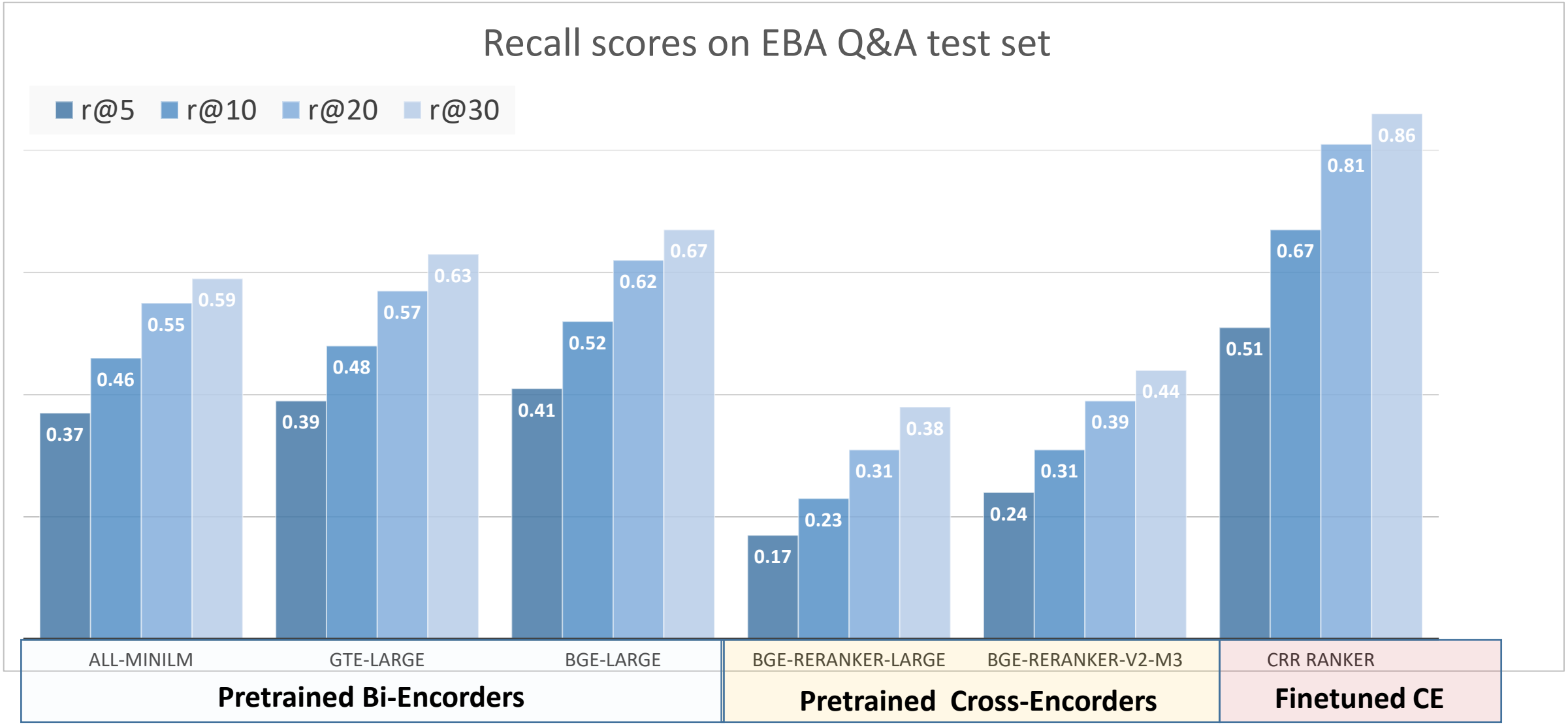
CRR Ranker Training

CRR Ranker model was trained **to retrieve pertinent articles from the CRR** in response to specific inquiries



H. Xuan et al., "Hard negative examples are hard, but useful," 2021. arXiv:2007.12749.
Chen et al., "Bge m3-embedding: Multi-lingual, multi-functionality, multi-granularity text embeddings through self-knowledge distillation," 2024. arXiv:2402.03216.

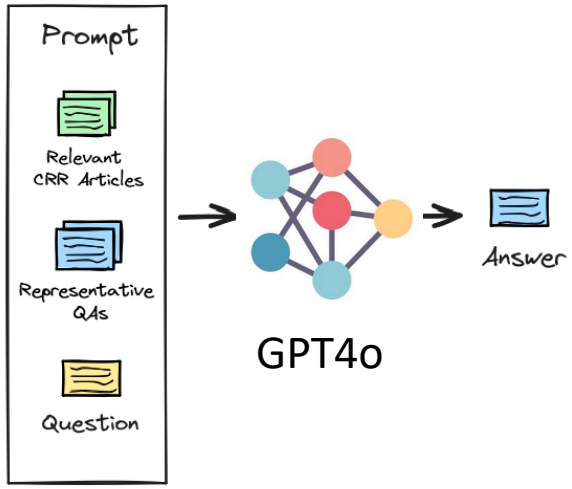
Results: CRR Retrieval



* Weiwei Sun,et al (2023). Is ChatGPT Good at Search? Investigating Large Language Models as Re-Ranking Agent

Answer Generation

- GPT 4o* for Generation.
- Role-based CoT few–shot approach includes 5 examples from the train set.
- Provides the desired tone, structure, language style, and level of detail expected.



Prompt for Answer Generation

" #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 ID:
1. 2013_192
2. 2018_3730
3. 2013_301
4. 2019_4705
5. 2014_783

#question:
> *question*

#context:
> *context*
> *enhanced_context*

Retrieved CRR Articles from our 3-Step Approach

#answer:

Role-based

Task
Step by Step Reasoning

Few-shot Guidance

Question

Context Enrichment

*2024-11-20 version

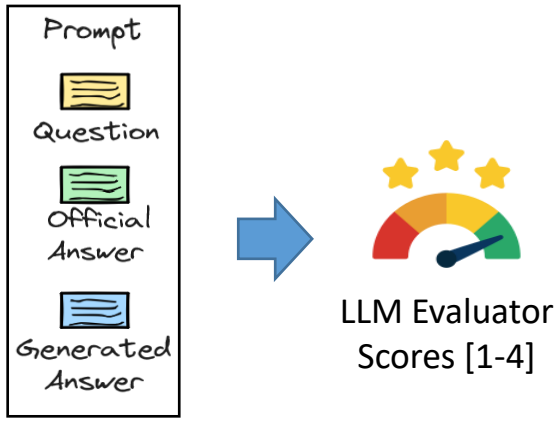
LLM Evaluator

Evaluation Scale (1-4):

- **Correctness**: Alignment with the official answer.
- **Completeness**: Inclusion of all relevant regulatory references.

Scores:

1. Completely incorrect.
2. Incorrect but complete or partially complete.
3. Correct but partially complete
4. Fully correct and complete.



Prompt for Answer Evaluation

I will provide you with two answers to a question: the #official answer (benchmark) and the #generated answer (to be evaluated). Compare them step by step based on:

Correctness: Does the #generated answer align with the #official answer?
Completeness: Does the #generated answer include all relevant information from the #official answer?
Rating Scale (1-4):

1. Completely incorrect and incomplete.
2. Incorrect but complete or partially complete.
3. Correct but partially complete.
4. Fully correct and complete.

Provide a numerical rating (1-4) followed by a brief explanation. Format your output as follows: Output: {score} Motivation: {motivation}

Examples
<EXAMPLE 1> ... <EXAMPLE 8>

Compute the score:
Question:
> question
> background
Official Answer:
> answer
Generated Answer:
> generated answer
Output:

Task

Criteria

Few-shot

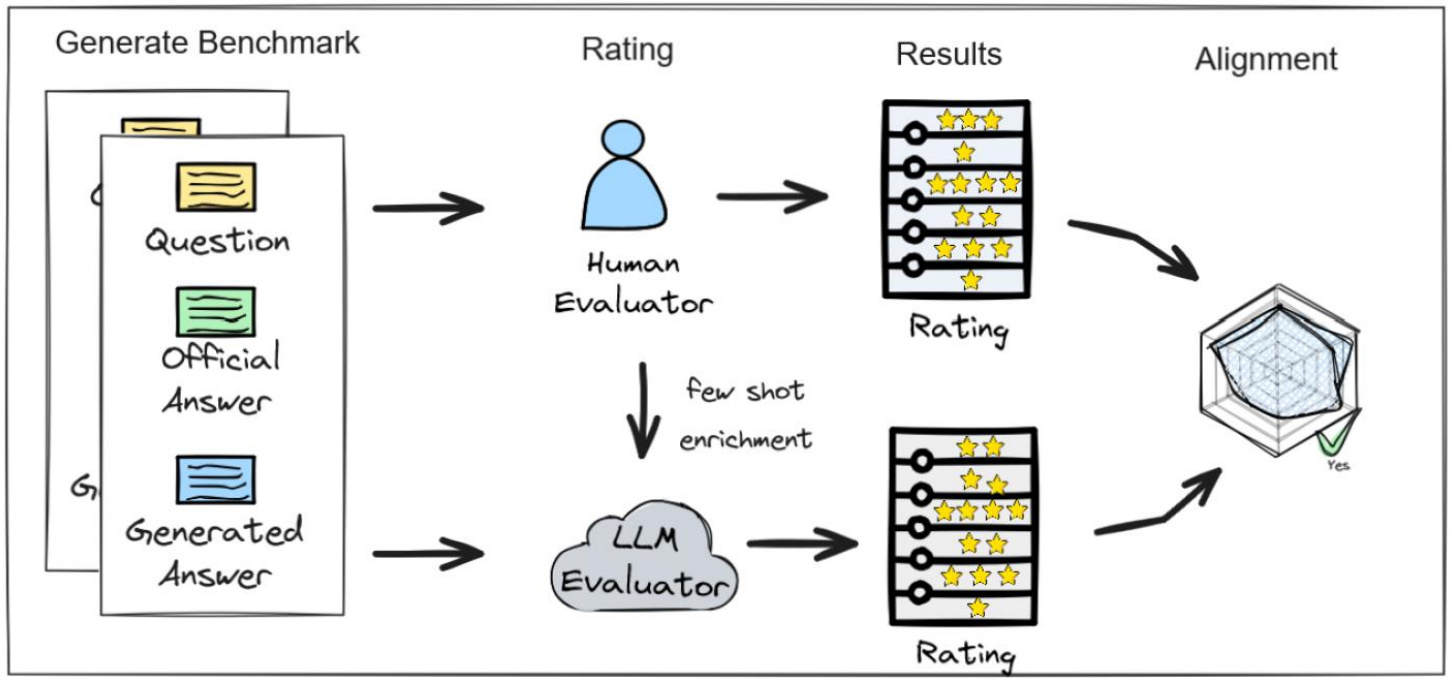
Question

Official Answer

Generated Answer

LLM Evaluator: Alignment with Human Expert

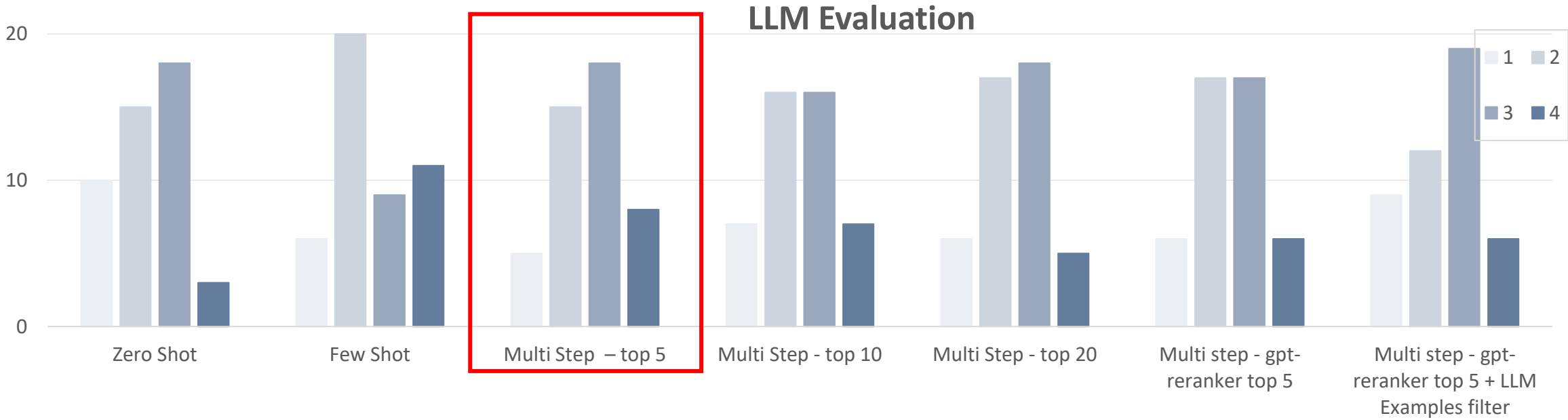
- Preliminary verification on a synthetic human-validated dataset.
- 60 Q&A pairs on Liquidity Risk topic balanced across the four categories.
- With GPT -4o**, Kendall-tau coefficient: 0.86** ,and p-value : 6.23^{-11}



Model	T	p-value
GPT4o (May '24)	0.77	6.10e-11
GPT4o (Nov '24)	**0.86**	6.23e-15
O1-mini (Sep '24)	0.64	9.44e-11

Kendall-tau: Agreement between the LLM Evaluator and human experts

Results: Answer Generation on Liquidity Risk Test Set (46 Q&As)



Approach with GPT4o (Nov'24)	AVG LLM 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

Multi-Step: Uses LLMs for extracting and the fine-tuned model for retrieving **top K** CRR articles.

Conclusion

Contributions:

- **Multi-Step Prompt Construction**: Enhances context for LLMs, improving answer precision and informativenes.
- **Context Enrichment**: Uses explicit and implicit CRR references, LLM capabilities, and a cross-encoder for precise retrieval.
- **LLM Evaluator**: Automates validation, ensuring response quality in terms of correctness and completeness.
- **Dataset Development**: Creates a comprehensive dataset from EBA's Single Rulebook Q&A for training and evaluation.
- **Performance Improvement**: Multi-step approach outperforms zero-shot and few-shot methods, providing better responses.

Future Directions:

- Increase the dataset size and generalization to other domains.
- Self-reflection and human feedback integration
- Explores different LLM architectures and fine tuning.

Thank you for your attention!

daniele.licari@bancaditalia.it
alessandro.degregorio@bancaditalia.it

ReRANK prompt

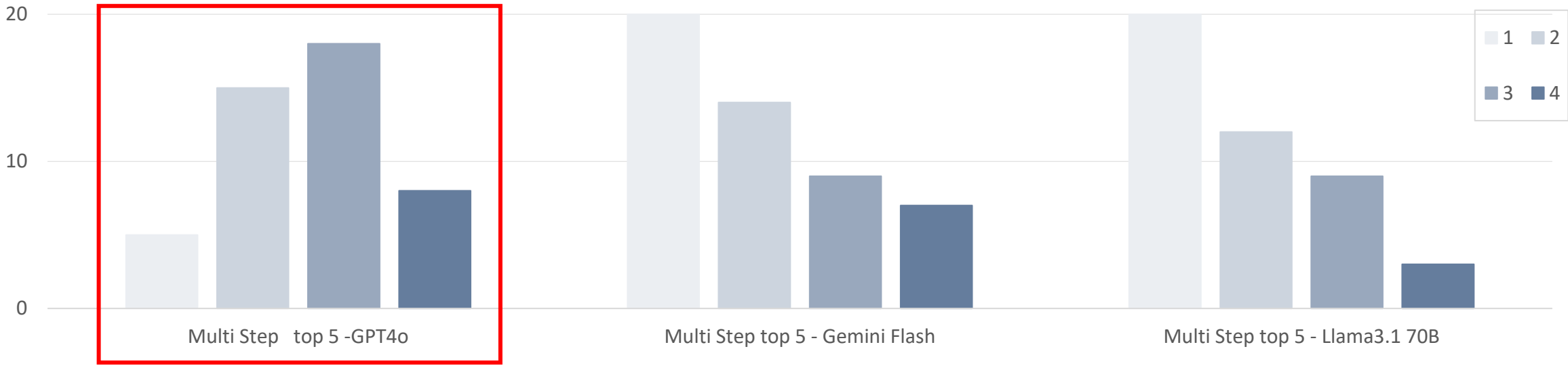
You are RankGPT, an intelligent assistant that can rank passages based on their relevancy to the query.
I will provide you with {num} passages, each indicated by number identifier []. \nRank the passages based on their relevance to query: {query}.

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.

Examples Filter 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, to clarify 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's question 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 information for 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}

Multi Step Evaluation with other LLMs



Approach	Model	AVG Score	# Correct (score>2)	% Correct
Multi Step – top 5	**GPT4o 2024-11-20**	**2.63 (±0.90)**	**26**	**56.5**
	Gemini Flash 1.5	1.76(±0.84)	8	17.3
	Llama 3.1 70B	1.85(±0.96)	12	26