2024 IEEE International Conference on Big Data (BigData) | 979-8-3503-6248-0/24/$31.00 ©2024 IEEE | DOI: 10.1109/BigData62323.2024.10825431

Advancing Risk and Quality Assurance: A RAG Chatbot for Improved Regulatory Compliance

Lars Hillebrand*†¶, Armin Berger*†‡, Daniel Uedelhoven†¶, David Berghaus†¶,
Ulrich Warning§, Tim Dilmaghani§, Bernd Kliem§, Thomas Schmid§, Rüdiger Loitz§, Rafet Sifa†‡

†*Fraunhofer IAIS*, Sankt Augustin, Germany
‡*University of Bonn*, Bonn, Germany
§*PricewaterhouseCoopers GmbH*, Düsseldorf, Germany
¶*Lamarr Institute*, Germany

Abstract—**Risk and Quality (R&Q) assurance in highly regulated industries requires constant navigation of complex regulatory frameworks, with employees handling numerous daily queries demanding accurate policy interpretation. Traditional methods relying on specialized experts create operational bottlenecks and limit scalability. We present a novel Retrieval Augmented Generation (RAG) system leveraging Large Language Models (LLMs), hybrid search and relevance boosting to enhance R&Q query processing. Evaluated on 124 expert-annotated real-world queries, our actively deployed system demonstrates substantial improvements over traditional RAG approaches. Additionally, we perform an extensive hyperparameter analysis to compare and evaluate multiple configuration setups, delivering valuable insights to practitioners.**

Index Terms—**Large Language Models, Retrieval Augmented Generation, Legal, Compliance**

I. Introduction

Compliance with Risk Management & Quality Standards is fundamental in regulated industries like auditing, finance, and legal services, where non-compliance can lead to significant legal penalties and financial losses. Employees face the challenge of navigating complex regulations and policies, handling numerous internal queries daily that require nuanced interpretation of trusted sources.

Large Language Models (LLMs) such as GPT [1], [3] are becoming increasingly important in legal work [5]. We propose a specialized ChatBot powered by GPT with an optimized Retrieval Augmented Generation (RAG) pipeline. Our system (see Figure 2) integrates hybrid search and relevance boosting to enhance retrieval accuracy and response quality.

We evaluate our solution using a handcrafted dataset with expert-annotated answers and develop a custom evaluation framework using `DeepEval`[1]. The framework achieves a 0.70 correlation coefficient with expert assessments. The ChatBot has been successfully deployed within the R&Q department of PricewaterhouseCoopers GmbH. The key contributions of this work are:

Example Query and R&Q Chatbot Response

Employee query:
"How far does the prohibition on the provision of tax advisory services extend in the case of EU PIEs?"

Workflow:
```
Preparing and embedding query
Running retrieval using hybrid search
Re-ranking text- and vector search results
Generating final answer with references
```

Final answer:
"To answer the question of how far the prohibition on providing tax advisory services to EU Public Interest Entities (PIEs) extends, we need to analyze the relevant information from the provided documents. [...] In summary, the prohibition on providing tax advisory services to EU-PIEs in Germany is very extensive and encompasses almost all tax advisory services listed on the blacklist. However, there are specific exceptions and national regulations that may allow certain tax advisory services under certain conditions."

Fig. 1. Example user question, model workflow and generated answer.

- **Development of a RAG ChatBot for R&Q standards:** We introduce a specialized ChatBot combining advanced AI capabilities with RAG.
- **Establishment of a Robust Evaluation Framework:** We devise an automated chatbot evaluation method corroborated by expert assessments.
- **Insights into Hyperparameter Optimization:** We identify how core hyperparameters affect system performance.

II. Methodology

This section presents our system designed to enhance compliance with Risk Management & Quality standards through ML-driven solutions. The system comprises: (1) an ingestion pipeline for document processing and indexing; (2) a RAG chatbot leveraging LLMs; and (3) an automated evaluation framework.

A. Ingestion Pipeline and Knowledge Base Construction

Our ingestion pipeline uses `Unstructured`[2] to parse documents into a structured data model. Documents are chunked with overlap for context continuity, with internal document

* Both authors contributed equally to this research.

[1] https://github.com/confident-ai/deepeval.

[2] https://github.com/Unstructured-IO/unstructured.

979-8-3503-6248-0/24/$31.00 ©2024 IEEE

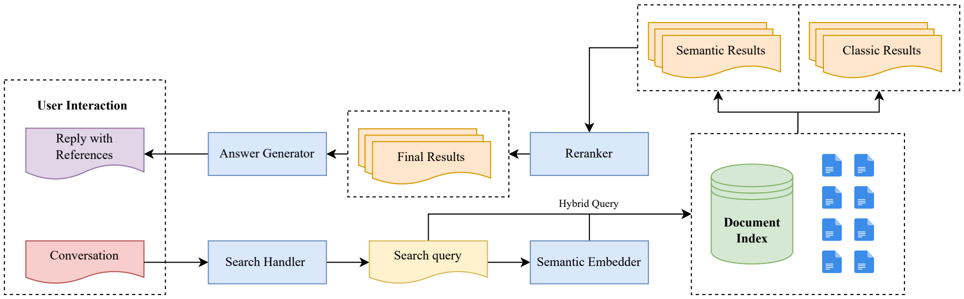

Fig. 2. Architecture of the Retrieval Augmented Generation chatbot system, demonstrating the workflow for query resolution.

chunks receiving a 2× boosting factor. Embeddings are generated using OpenAI's `ada-002` and `3-large` [4] endpoints and indexed via Azure AI Search.

B. Retrieval Augmented Generation Chatbot

Our RAG chatbot system interprets user queries, retrieves relevant information from the knowledge base, and generates contextually appropriate responses. We employ a hybrid search strategy that combines vector similarity search and full-text search using TF-IDF-based BM25 algorithms. The results from both searches are re-ranked using reciprocal rank fusion to enhance retrieval effectiveness. Moreover, we utilize the above described relevance boosting to ensure that the chatbot provides answers based on the most trusted and relevant information.

C. Automated Evaluation Framework

We establish an automated evaluation framework using `DeepEval`[3] and the G-Eval scoring method [2]. The evaluation focuses on correctness, completeness, relevance, and adherence to R&Q standards. We leverage GPT-4o as the LLM backbone for the evaluation and define the metric range between 0 (worst) and 5 (best). The following evaluation steps are performed to create the final score per sample.

Evaluation Steps

Answer Evaluation Steps

- Check if the facts in 'Actual Output' contradict any facts in 'Expected Output'.
- DO NOT punish long and detailed answers if the 'Actual Output' is perfectly correct. Generally, more details in the 'Actual Output' are encouraged.
- If the 'Actual Output' misses details compared to the 'Expected Output' you should slightly penalize omission of detail.

Context Evaluation Steps

- Summarize the expected 'Context' and note the most important points.
- Compare the summary with the 'Retrieval Context' and check if the most important points are present.
- If the 'Retrieval Context' is missing important points compared to the 'Context' you should penalize the response.
- If the 'Retrieval Context' contains irrelevant information, you should very slightly penalize the response.
- If the 'Retrieval Context' contains contradictory information, you should heavily penalize the response.

To validate reliability, we compare LLM-based scores with manual evaluations from domain experts across 124 responses, achieving a Pearson correlation coefficient of $r = 0.70$. While acknowledging potential LLM biases [6], this correlation supports the use of automated evaluations as proxies for expert judgment.

III. Experiments

We conduct experiments on an expert-curated dataset to provide insights for implementing LLM-based chatbots in production environments. We present our dataset, experimental setup, and discuss our findings.

A. Data

Our dataset[4] comprises 124 R&Q question-answer pairs created by domain experts. An illustrative question example is highlighted in Figure 1. Of these questions, 110 used internal sources, with the remainder drawing from external data. Thirteen experts contributed to the dataset creation, with oversight from three senior R&Q specialists. Multiple review rounds ensured quality control through random sampling and qualitative assessment.

B. Model Configurations

Our ablation studies examine three key areas: (1) ingestion parameters, (2) retrieval parameters, and (3) model parameters, measuring their impact on system performance. Table I presents the complete configuration space, with bold values indicating our baseline setup. All configurations use an LLM temperature value of 0 to increase answer robustness. Through systematic evaluation of ingestion and retrieval parameters, we identify the optimal configuration achieving the highest correctness scores. While initially using `ada-002` for embeddings, we discovered that `3-large` yields superior performance during our retrieval optimization process, leading to its adoption in subsequent experiments. The final optimized configuration is then used as the foundation to assess different LLM backbones (see Table III).

C. Prompt Design

Our prompt design includes a template that ensures consistent and accurate responses from the LLM. We utilize dynamic language detection using the `langdetect`[5] library

[3] https://github.com/confident-ai/deepeval.

[4] Dataset and Python code are currently unpublishable due to ongoing industrial project constraints.

[5] https://github.com/Mimino666/langdetect.

TABLE I
HYPERPARAMETER CONFIGURATIONS. BOLD VALUES INDICATE THE BASELINE SETUP USED FOR ABLATION STUDIES.

Module	Hyperparameter	Configurations
Ingestion	Max Chunk Size	256, **512**, 1024, 2048
	Min Chunk Overlap	32, **64**, 128, 256
	Markdown Conversion	Yes, **No**
Search	Top-k	5, **10**, 20
	Search Type	Text, **Hybrid**, Vector
	Relevance boosting	Yes, **No**
	Embedding model	ada-002, 3-large
ChatBot	LLM-Backbone (GPT)	4o-mini, **4o**, 3.5-Turbo, 4-Turbo

to automatically adjust the language of the response to match the user's query. The prompt instructs the model to cite sources appropriately, and avoid hallucinations by stating when information is not present in the provided context.

Prompt Template

```
{user_question}

<instruction>
Write your answer in {language}. If you cannot answer the question based on the provided context, state that the information is not present, don't invent or hallucinate an answer and don't reference any sources. After each fact you state, provide the corresponding document name and chunk id from the appended sources in brackets and separated by "/". For example: "Apple was founded in 1976." ∗(apple.docx/1)∗
Don't combine sources but list each individual source separately if a fact contains multiple sources. E.g. ∗(apple.docx/1)∗, ∗(apple.docx/2)∗, etc.
You must comply with the following sources format: ∗(<document_name_as_str>/<chunk_id_as_int>)∗
Before answering the question, lay out your full thought process and dissect the user question and its implications.
</instruction>

<document_context>
{retrieved_chunks}
</document_context>
```

D. Results

Our experiments reveal that the optimal configuration, Baseline$_{\text{3-large}}$ with relevance boosting enabled, achieves the highest correctness scores for both answers and context, as detailed in Table II. Hybrid search consistently outperforms individual methods, and relevance boosting further improves the prioritization of internal documents.

Using this optimal configuration, named R&Q-Chatbot, we conduct a comprehensive evaluation across different LLM backbones. Table III shows that while all models deliver reasonable answers, GPT-4o demonstrates the best performance. For robust analysis, each model configuration was evaluated using 5 independent runs, reporting mean and standard deviation for our G-Eval correctness metric.

TABLE II
DETAILED ABLATION STUDY TO EVALUATE MULTIPLE MODEL CONFIGURATIONS. WE REPORT MEAN AND STANDARD DEVIATION VALUES OF 5 INDEPENDENT RUNS (BEST SCORES IN BOLD) FOR BOTH, ANSWER AND CONTEXT CORRECTNESS (SCALE: 0-5).

Model Configuration	**G-Eval Correctness Score**	
	Answer ↑	Context ↑
Baseline$_{\text{ada-002}}$	**3.72** (±.026)	2.80 (±.028)
Chunking: 256/64	3.61 (±.048)	2.75 (±.041)
Chunking: 512/32	3.69 (±.042)	2.84 (±.043)
Chunking: 512/128	3.69 (±.053)	**2.88** (±.046)
Chunking: 1024/128	3.67 (±.045)	2.74 (±.049)
Chunking: 1024/256	3.66 (±.039)	2.83 (±.045)
Chunking: 2048/256	3.56 (±.050)	2.29 (±.015)
+Markdown	3.65 (±.050)	2.77 (±.030)
Baseline$_{\text{3-large}}$	3.76 (±.030)	**2.91** (±.031)
Vector Search	3.72 (±.030)	**2.91** (±.032)
Text Search	3.60 (±.030)	2.62 (±.026)
Top-k: 5	3.72 (±.033)	2.77 (±.027)
Top-k: 20	3.72 (±.016)	2.90 (±.048)
+Relevance Boosting	**3.79** (±.037)	2.90 (±.018)

Chunking: 512/64 (Max Chunk Size = 512, Min Chunk Overlap = 64)

TABLE III
RESULTS OF THE BEST ARCHITECTURAL SETUP FOR DIFFERENT LLM BACKBONES (SCALE: 0-5 AND BEST SCORES IN BOLD).

Model Configuration	**G-Eval Correctness Score**	
	Answer ↑	Context ↑
GPT-4o (R&Q-Chatbot)	**3.79** (±.037)	**2.90** (±.018)
GPT-4-Turbo	3.69 (±.047)	2.84 (±.048)
GPT-4o-mini	3.63 (±.053)	2.79 (±.037)
GPT-3.5-Turbo	3.27 (±.012)	2.53 (±.077)

IV. CONCLUSION AND FUTURE WORK

In this work, we introduced a novel RAG chatbot system tailored for R&Q assurance in highly regulated industries. Our system effectively leverages LLMs with optimized retrieval strategies including hybrid search and relevance boosting to improve query processing and compliance adherence. The evaluation demonstrates significant performance gains over baseline approaches, validating the efficacy of our system.

Future research will focus on extending the chatbot to a dynamic multi-agent system capable of intelligent query dissection, clarifying questions, and multi-hop reasoning to further enhance its conversational capabilities.

V. ACKNOWLEDGMENT

This research has been funded by the Federal Ministry of Education and Research of Germany and the state of North-Rhine Westphalia as part of the Lamarr-Institute for Machine Learning and Artificial Intelligence, LAMARR22B.

REFERENCES

[1] T. Brown, B. Mann, N. Ryder, *et al.*, "Language models are few-shot learners," in *Proc. NeurIPS*, 2020.

[2] Y. Liu, D. Iter, Y. Xu, S. Wang, R. Xu, and C. Zhu, "G-eval: Nlg evaluation using gpt-4 with better human alignment," *arXiv:2303.16634*, 2023.

[3] OpenAI, "Gpt-4 technical report," *arXiv:2303.08774*, 2023.

[4] OpenAI, *New embedding models and api updates*, https://openai.com/index/new-embedding-models-and-api-updates/, 2024.

[5] I. Rodgers, J. Armour, and M. Sako, "How technology is (or is not) transforming law firms," *Annual Review of Law and Social Science*, vol. 19, no. 1, pp. 299–317, 2023.

[6] L. Zheng, W.-L. Chiang, Y. Sheng, *et al.*, "Judging llm-as-a-judge with mt-bench and chatbot arena," *Advances in Neural Information Processing Systems*, 2023.