KnowRAG: A Zero-Shot Diagnostic Analysis of Knowledge Base Coverage in Scientific Retrieval-Augmented Generation

Assmaa MOUTAOUKKIL*, Ali EL MEZOUARY, Kaoutar BOUMALEK
IRF-SIC Laboratory-Department of Computer Science-Faculty of Science, Ibn Zohr University, Agadir, Morocco

***Abstract*—The "hallucination" problem in Large Language Models (LLMs) remains an unresolved hurdle for scientific researchers who require precise, grounded evidence. While Retrieval-Augmented Generation (RAG) aims to mitigate these errors, standard systems are often unoptimized for the structural complexities of scientific papers. We introduce KnowRAG, a zero-shot RAG pipeline specifically designed for scientific applications. Using a novel "LLM-as-a-Judge" diagnostic framework, we evaluated KnowRAG against a standalone GPT-3.5-Turbo baseline across four specialized Q&A Test Sets. Our results demonstrate that KnowRAG significantly improves factual accuracy over the baseline. More importantly, diagnostic analysis reveals that the vast majority of errors (over 46%) stem from Knowledge Base Coverage (knowledge gaps), while generation failures remain negligible at 4%. These findings suggest that retrieval and generation capabilities are no longer the primary bottlenecks in the scientific domain. Instead, this diagnostic analysis advocates for a paradigm shift from model-centric research toward expert data engineering as the definitive path to trustworthy AI. By repurposing the LLM-as-a-Judge framework as a diagnostic instrument rather than a mere performance metric, we move RAG evaluation beyond aggregate scoring toward actionable, evidence-based systemic diagnosis.**

***Keywords*—Large Language Models; Retrieval-Augmented Generation; evaluation; scientific writing; information retrieval; knowledge base; data engineering; GenAI**

I. Introduction

For a scientific researcher, the ideal assistant is a collaborator capable of synthesizing recent, multi-disciplinary literature with absolute citation fidelity. This is one of the reasons many people are getting excited about the possibilities of using Large Language Models (LLMs) [1]–[3], as LLMs have used these massive amounts of human knowledge, to completely transformed AI interaction [4]–[7] and information retrieval [2], [3], [8]–[10], and they now exhibit a very high level of capability in supply/output complex written language [2]. However, this very strength masks a profound weakness for academic research: their tendency to "hallucinate", generating seemingly convincing and persuasive texts that are subtly or blatantly inaccurate[11]–[13]. In scientific research, where the chain of evidence is vital, this error is fatal. The consequence is not just an occasional mistake, but a fundamental breach of trust, relegating these powerful tools to the status of thinking aids rather than reliable sources of knowledge. We have the linguistic engine, but it is detached from the core of the evidence and ethics. The root of this problem lies in grounding. LLM like GPT-3.5-turbo operates based on its parametric knowledge, a static, frozen snapshot of the world as perceived through its training data [14], [15]. When asked a question that goes outside the scope of this snapshot, such as the conclusions of a preprint published last month, it must generalize, often leading it to fabricate a plausible but incorrect answer. The ideal situation would be a system that seamlessly combines the inferential and linguistic capabilities of an LLM with direct, dynamic access to a reliable corpus. The dominant approach to filling this gap is retrieval-augmented generation (RAG)[5], [16]–[21]. By first retrieving relevant documents from an external knowledge base and then conditioning the LLM's generation on that specific context, RAG improves adaptability, factuality, and responsiveness to up-to-date content, and promises to link fluency with factuality[22][23][24]. The conceptual model is elegant in its separation of matters: a non-parametric memory for facts and a parametric model for reasoning and language.

However, most original work on RAG, such as the innovative framework in [19], has focused on architectural innovation, joint training of retrievers[25][26][27] and generators[28][29], or scaling to massive generic corpora. Although these are crucial advances, their evaluation has been mostly imprisoned to benchmarks such as Natural Questions, which test general faithful memorization from sources such as Wikipedia. This creates a significant and largely unexplored translation gap. Scientific literature is not a series of isolated facts; it is a dense tapestry of arguments, methodologies, results, and critiques. An improved system for retrieving a fact about a historical event may fail miserably when asked to retrieve the methodological justification for a particular experimental design, especially if the knowledge base is prepared with a generic segmentation of the text based on length, which turns a coherent argument into incoherent fragments. As a result, many existing implementations of RAG, despite their sophistication, can still produce inaccurate results in scientific contexts. The indirect consequence is a slowdown in their potential adoption; researchers cannot rely on the results without verifying them themselves exhaustively, which runs counter to the promised efficiency. This directly highlights the knowledge gap that our study attempts to address. We do not clearly understand how practical choices at the implementation level in the construction of a scientific RAG system, particularly the structuring of the knowledge base, influence its actual performance in reducing hallucinations. Previous studies have provided us with powerful engines, but little guidance on how to map their course in the specialized domain of science.

*Corresponding author

This work is based on a pragmatic conceptual model: for scientific RAG, the most effective improvements do not inevitably come from additional pre-training of massive models, but rather from an intentional and domain-appropriate design of the retrieval pipeline that feeds the generator. We build on the established RAG paradigm in accordance with the best practices mentioned in [30], but our approach differs. Instead of modifying the base LLM or training a new retrieval system from scratch, we investigate a minimal, zero-shot configuration using robust, off-the-shelf components. Our study asks a direct and practical question: Can careful knowledge base construction and retrieval setup, applied to a standard LLM, yield an impressive reduction in hallucination without any fine-tuning model?

Therefore, the objectives of this research are threefold:

- Implementation: To develop KnowRAG, a zero-shot pipeline utilizing semantic search on a processed arXiv corpus and generation via GPT-3.5-turbo.
- Evaluation: To rigorously test whether this pipeline improves factual accuracy across four domain-specific test sets compared to a standalone baseline using the LLM-as-a-Judge approach.
- Diagnosis Analysis: To perform a granular failure analysis, determine if errors stem from knowledge base gaps, retrieval misses, or generation failures.

Our primary contribution is the transformation of the LLM-as-a-Judge into a diagnostic instrument. By isolating specific systemic bottlenecks, we provide an actionable framework that moves beyond aggregate accuracy scores toward a reproducible model for reliable, inexpensive scientific assistants.

The remainder of this study mirrors our investigation process: Section II reviews related work and analyzes existing literature, while Section III details the methodology used to provide a diagnostic analysis of comprehensive RAG systems. In Section IV, we outline the experimental setup, including the creation of specialized test sets and a performance comparison against a reference LLM. Section V presents our results, offering both a quantitative evaluation of accuracy and a qualitative diagnosis of failure modes. We discuss these findings in Section VI and address the study's recommendations and limitations in Section VII. Finally, the study concludes in Section VIII.

II. Literature Review

Over the past five years, Retrieval-Augmented Generation has evolved in a field increasingly focused on algorithmic innovation [25]–[29]. Advances such as dense passage retrieval, end-to-end training of joint retriever-generator models, and more refined methods for combining retrieved content have led to measurable improvements on open-domain question answering tasks [19]. Yet this advancement in our knowledge relies that the documents used to build these models are organized, consistent, and unbiased but this is rarely true. Such a perspective cannot be upheld any longer. In narrow domains such as scientific literature, where texts contain rhetorical and informational layers and information is concentrated in different sections, the way in which the knowledge base is built becomes a defining factor for system performance. Key findings that are buried under narrative digressions, methodologies that are cut off mid-sentence, and indexing strategies that force distorted structural content into flat, indistinct lumps all highlight fundamental limitations. These are not deficiencies that can be solved by even the best-optimized retriever. Therefore, recognizing data primacy shifts the perception of the knowledge base from that of a passive archive to an active factor that shapes the maximum achievable retrieval accuracy, and consequently the factual integrity of the generated responses.

One of the greatest challenges is evaluating RAG systems as composite systems. How can we objectively assess the relevance and accuracy of a response to its sources? How can we obtain a performance measure for each component of the system? Traditional metrics from natural language processing: BLEU, ROUGE, and lexical matching do not reflect the semantic nuances of the generated text or its contextual accuracy. Human annotations are expensive and difficult to maintain throughout multiple experiments [31], [32]. The increasing use of large language models as evaluators could solve a long-standing dilemma in evaluation methodology [33]. Recent studies show that, when equipped with clear evaluation criteria and carefully designed prompts, LLMs can assess the relevance of research and the fidelity of generation, producing results that are very similar to human judgments and reaching levels of agreement comparable to those between annotators [35]. However, the significance of this change goes beyond simply replicating human evaluations effectively. The true value of the LLM as a judge lies not only in its ability to assign accurate scores but also in its ability to provide meaningful characterizations. A low relevance score in itself provides little information; what is more informative is when the model explains the failure, via chain-of-thought reasoning or dimension-specific scoring, and delivery of why relevance failed, whether due to missing content, rhetorical misalignment, or extractive fragmentation, renders visible the otherwise opaque influence of data curation decisions. While frameworks such as RAGAS [31] evaluate fidelity and accuracy, they focus on a model-centric approach: comparing retrievers in a static pipeline. This study explores a different path, using LLMs not just for final evaluation, but as tools to analyze the inner workings of the entire RAG pipeline.

This study, therefore, occupies a specific and, in our view, necessary place in this literature. Whereas previous work has optimized components in isolation or compared architectures under constant corpus conditions, we reverse the experimental logic: the algorithms are kept unchanged while the data preparation strategies are systematically modified. The goal is not to crown a single optimal configuration, but to trace, with empirical specificity, how each preservation choice propagates through retrieval and generation to shape measurable results in terms of contextual fidelity and factual accuracy. To do this, we use LLM-as-a-judge not as a complementary metric, but as a central diagnostic tool, capable of distinguishing, for example, a model that misread a passage from a model that never received enough passages to respond. This distinction, which conventional metrics cannot make, is precisely where the 'data-first' hypothesis becomes verifiable. Our implementation, KnowRAG, applies this logic in a reproducible, zero-shot pipeline designed for scientific corpora. We emphasize that this is not a proposal for a new architecture, but an argument for a

new unit of analysis. The fundamental contributions of existing RAG and their successors have equipped the field with its tools. What remains to be written, and what we aim, in part, to provide, is a systematic account of how these tools succeed or fail not in spite of the data, but because of it.

III. Methodology

This study employed a controlled experimental methodology to implement and evaluate KnowRAG, a zero-shot pipeline designed for scientific domains. Fig. 1 summarizes the KnowRAG methodology. Our approach is structured across three sequential phases to systematically measure improvements in correctness and to perform diagnostic failure analysis.

A. Knowledge Base Construction and Benchmarking

The study built a domain-specific corpus by scraping ten recent RAG-related publications from arXiv.

- Preprocessing: PDFs were converted to raw text and partitioned into 1,000-token chunks with a 30-token overlap using a recursive splitter.
- Vectorization: Each chunk was embedded using text-embedding-ada-002 and stored in a FAISS vector database for cosine similarity searching.
- Benchmarking (Q&A Test Sets): GPT-4 was utilized to generate 240 gold-standard questions (four sets of 60) directly from the corpus to ensure a verifiable "ground truth" for each query.

B. Experimental Pipeline

A comparative "Zero-Shot" procedure was executed to isolate the impact of the retrieval mechanism on a generative core (GPT-3.5-turbo). The choice of this model serves as a methodological control; by using a well-documented, stable baseline rather than a frontier model, we minimize "parametric contamination" (where the model answers from its own training memory) and more clearly isolate the performance delta attributable strictly to the RAG architecture.

- Baseline (Condition A- No retrieval): The LLM answered questions using only its internal pre-trained knowledge via a simple factual prompt.
- KnowRAG (Condition B- Retrieval-Augmented): The system embedded the query, retrieved the top-3 relevant chunks from the FAISS index, and injected them into a structured prompt. The LLM was strictly initiated to answer based only on the provided context and cite specific source IDs.
- Controls: To ensure reproducibility, parameters were fixed at a low temperature (0.1) and top_p=1.0, with no fine-tuning involved.

C. Evaluation Metric and Diagnostic Framework

To move beyond simple accuracy, we used GPT-4 as an automated judge to provide binary correctness scores and identify specific systemic bottlenecks. To mitigate evaluation circularity, we employed a hierarchical model strategy: utilizing the superior reasoning capabilities of GPT-4 to assess the outputs of the smaller GPT-3.5-Turbo generator. Following the LLM-as-a-Judge approach, GPT-4 analyzed the original question, the generated response, and the ground-truth context to determine if an answer was "correct" or "incorrect". By moving from subjective grading to objective verification, this method ensures the consistency and scale required for a deep diagnostic study. Importantly, this approach aligns closely with human expert judgment as demonstrated in [33], [34].

To understand the results, the pipeline tracked two essential metrics:

- Knowledge Base Coverage: Did the corpus contain the answer?
- Retriever Hit Rate (Top-3): Was the correct information successfully ranked within the top three documents?

These metrics are critical because they define the system's performance limits; a correct answer can only be generated if the information is both present in the corpus and successfully retrieved by the pipeline.

To understand the root cause of every error, we categorized incorrect answers into three distinct failure modes:

- Knowledge Gap: The answer was missing from the source documents entirely.
- Retrieval Failure: The information was in the corpus, but the retrieval failed to find it.
- Generation Failure: The correct context was successfully retrieved, but the LLM failed to present it faithfully.

This stepwise, instrumented methodology moves beyond reporting a simple performance delta. It provides a diagnostic framework that explains, with reasonable precision, why a zero-shot RAG system succeeds or fails in the scientific domain.

IV. Experiments

We conducted a series of controlled experiments to evaluate KnowRAG against the standalone GPT-3.5-turbo baseline. The experiments were designed to answer three core research questions. First, does retrieval augmentation significantly improve factual correctness in scientific question answering? Second, how consistent is this improvement across different types of scientific queries? And third, where do failures originate in a zero-shot RAG pipeline, and what do these failures reveal about the relative importance of knowledge base quality, retrieval accuracy, and generation fidelity?

A. Experimental Setup

All experiments were executed in a standardized computational environment. Document processing and embedding generation were performed using Python 3.10 with the LangChain, PyMuPDF, and FAISS libraries. All LLM inferences, both for generation and evaluation, were conducted via the OpenAI API. The temperature parameter was fixed at 0.1 for all generation tasks to ensure deterministic outputs, while the LLM-as-a-Judge evaluations using GPT-4 were conducted at temperature 0.0 for maximum consistency. The complete knowledge base and the generated test sets are available in the Appendix section to ensure full reproducibility.

B. Dataset and Benchmark Characteristics

Our evaluation leveraged four distinct QATestSets, each containing sixty question-answer-context triples derived from our curated arXiv corpus. The four test sets in the KnowRAG diagnostic framework were designed to vary in topic coverage to test the system's robustness across different levels of scientific complexity and domain specificity:

- QATestSet1: 4 topics, 60 questions
- QATestSet2: 7 topics, 60 questions
- QATestSet3: 6 topics, 60 questions
- QATestSet4: 4 topics, 60 questions

In each test set, the questions covered six classifications representative of scientific questions: Complex, conversational, distracting element, double, simple, and situational. For each question, GPT-4 produced a reference answer and cited the specific source identifier, thus providing verifiable truth. The final benchmark comprised 240 question-answer-context triplets, covering four distinct categories of reasoning.

To facilitate reproducibility, the four Q&A Test Sets files, including ground truth annotations, are publicly available at link.

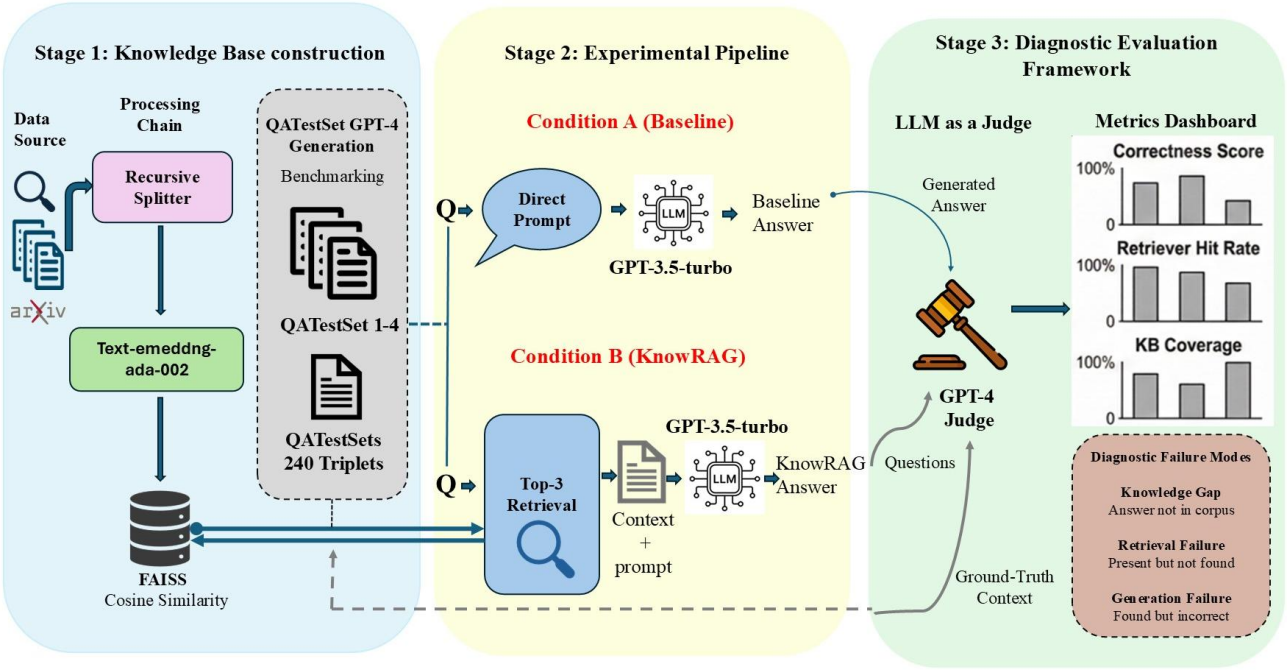

Fig. 1. The KnowRAG approach.

V. Evaluation and Results

In this section, we analyze the quantitative and qualitative performance of KnowRAG. For a detailed analysis of the evaluation Test Sets and extended diagnostic logs, please refer to Appendix A and Appendix B.

A. Comparative Performance: Baseline vs. KnowRAG

We first evaluated the factual correctness of both conditions across all four test sets. The results are presented in Table I. KnowRAG increased the average factual correctness of the baseline by more than double, from 27.5 per cent to 59.5 per cent. This improvement was consistent across all four test sets with gains ranging from 21 to 40 percentage points. The largest improvement of forty points was observed on QATestSet3, while the smallest gain of twenty-one points was observed on QATestSet2. This variation was our first indication that not all test sets, and by extension not all knowledge bases, are created equal.

TABLE I. Correctness Scores (%) by Method and Test Set Correctness is Measured via LLM-as-a-Judge (GPT-4)

Method	QATestSet1	QATestSet2	QATestSet3	QATestSet4	Average
Baseline (No Retrieval) GPT-3.5-turbo	28	29	23	**30**	27.5
KnowRAG (Retrieval-Augmented)	60	50	63	**65**	59.5
Improvement	+32	+21	+40	+35	+32

B. Diagnostic Analysis: Knowledge Base Coverage and Retriever Hit Rate

To understand why KnowRAG succeeded and where it failed, we conducted a diagnostic analysis using the metrics described in the Methodology section. Table II shows the results by Test Set.

The Knowledge Base Coverage metric reveals a catastrophic failure of our chunking strategy for certain test sets. QATestSet1

achieved excellent coverage of nearly 86%, with most of the answers to its questions preserved complete in our processed chunks. However, QATestSet2 suffered coverage of only 18%. This means that, despite the answers being present somewhere in the original PDFs, more than four out of five questions in this set had no retrievable answer after our processing.

The correlation with performance is immediate and clear. QATestSet2, which had the lowest coverage, produced the lowest correctness score (50%) and the smallest improvement over the baseline score (+21 points). In contrast, QATestSet1 and QATestSet4, with higher coverage, achieved substantially better performance. The causal relationship is clear. When the knowledge base is fragmented and incomplete, performance suffers.

With an average Retriever Hit Rate of 61%, the news report is somewhat more encouraging. When the answer was present in the knowledge base, our semantic search successfully retrieved the correct source chunk in the top three results for almost two-thirds of the questions. Performance varied, with particularly strong retrieval on QATestSet4 (70%) and weaker retrieval on QATestSet2 (45%). This suggests that methodological descriptions may be more challenging to match using cosine similarity on dense embeddings. Questions such as "What is the main contribution of paper X?" have a clear semantic center of gravity, whereas questions such as "How does the training procedure in method Y differ from standard approaches?" are more diffuse and may be harder to align with a single procedural passage.

TABLE II. INTERMEDIATE RETRIEVAL METRICS BY TEST SET (%)

Metric	QATestSet1	QATestSet2	QATestSet3	QATestSet4	AVG
Knowledge Base Coverage	85.71	18.18	44.44	66.67	53.75
Retriever Hit Rate (Top-3)	65	45	65	70	61.25

C. Failure Mode Analysis

By correlating the correctness scores with these intermediate metrics, we were able to categorize each incorrect KnowRAG response as one of three failure modes. Table III shows the resulting distribution.

The failure mode analysis yielded two profound insights that fundamentally shaped our understanding of where the real work lies in building reliable scientific RAG systems.

Firstly, the Knowledge Gap Failure is the dominant failure mode, accounting for nearly half of all errors on average and a vast 82% of errors on QATestSet2. This is not a failure of the retriever or the generator. It is a failure of the knowledge base itself. When the answer is not present in a retrievable chunk, the system has no path to correctness, regardless of the sophistication of its components. The implication is clear: the single greatest bottleneck in our zero-shot RAG pipeline is not the neural architecture of the retriever, nor the instruction-following capability of the LLM, but the mundane, understated work of deciding how to split a PDF into chunks.

Second, Generation Failure is extremely rare, accounting for only 4% of errors on average. When the correct context was successfully retrieved and provided to GPT-3.5-turbo with our structured prompt, the model almost constantly produced a faithful, accurate answer. This is a validation of the zero-shot approach. It suggests that a general-purpose instruction-tuned LLM, when given high-quality, relevant context, requires no fine-tuning to perform scientific question answering with high faithfulness. The model can read; it can follow instructions; it can synthesize. What it cannot do is conjure information that was never provided. The bottleneck is upstream.

TABLE III. FAILURE MODE DISTRIBUTION BY TEST SET (%)

Failure Mode	QATestSet1	QATestSet2	QATestSet3	QATestSet4	AVG
Knowledge Gap Failure	14.29	81.82	55.56	33.33	46.25
Retrieval Failure	20.71	36.18	19.44	16.67	23.25
Generation Failure	5.00	0.00	6.00	5.00	4.00

Retrieval Failure accounted for approximately 23% of errors, indicating meaningful room for improvement in embedding strategies, search techniques, or the choice of k. But this is, in the context of our findings, a secondary concern. Improving retrieval from 61% to 80% would be valuable, certainly, but it would still leave the system crippled if Knowledge Base Coverage remains at 53%. The order of operations matters, and our results suggest that coverage is the initial constraint.

D. Qualitative Analysis

To complement our quantitative findings, we examined specific examples from each failure category. These cases, presented in Table IV, illustrate the concrete mechanisms behind our statistical aggregates.

As illustrated in Table IV and Fig. 2 to Fig. 9, the qualitative examples confirm our quantitative results. A consistent pattern emerges: bibliographic queries regarding titles and authors fail across all test sets, despite high knowledge coverage. We attribute this to a structural bottleneck, as metadata is typically confined to specific document regions (e.g., first pages or citation blocks), making it systematically difficult to retrieve.

Furthermore, the lower Knowledge Base Coverage in Q&A Test Set 2 is directly linked to the document format. Unlike other sets, the source articles for Test Set 2 rely heavily on tabular data; because our pipeline struggled to ingest these tables, the relevant information was never indexed. This created a structural knowledge gap where the data existed in the original paper but remained absent from the system's searchable corpus.

TABLE IV. QUALITATIVE EXAMPLES OF FAILURE MODES

Question	Retrieved Context	Agent Answer	Ground Truth	Failure Mode	Key Insight
What is the title of the work by Xiaoqing Li, Jiajun Zhao, ...?	Document 105: Annual Meeting of the Association...	The title of the work by Xiaoqing Li, Jiajun Zhao... is "R2AG: Incorporating Retrieval Information into Retrieval-Augmented Generation."	The title of the work by Xiaoqing Li, Jiajun Zhao... is "Non-parametric unsupervised domain adaptation for neural machine translation."	Knowledge Gap Failure	Required paper not in corpus or metadata chunk missing.
According to Table I, which method demonstrated the highest results for the NQ dataset?	Document 22: 0.2808 0.1276 0.3882 0.2253 L...	According to Table I, the method that demonstrated the highest results for the NQ dataset is R^{2}AG (R-Squared AG).	The method R2AG+RAFT showed the highest results for the NQ dataset.	Generation Failure	Correct table retrieved but misinterpreted.
What are the three main ones? (After discussing retrieval-augmented text generation components)	Document 56: A Survey on Retrieval-Augmented T...	The three main datasets mentioned in the provided context are NQ, TQA, and HotpotQA.	The three key components of the retrieval-augmented generation paradigm are Retrieval, Augmentation, and Generation.	Retrieval Failure	Conversational thread lost; retrieved datasets instead of components.
Could you tell me the title of the academic paper by Edward J. Hu and his colleagues?	Document 119: Harrison Chase. 2022. LangChain....	I don't know.	The title of the paper by Edward J. Hu and his colleagues is "LoRA: Low-Rank Adaptation of Large Language Models."	Knowledge Gap Failure	Paper not present in corpus (low coverage).

Beyond this, three recurring failure modes stand out:

- First, knowledge gap failures often occurred not because information was absent from the corpus, but because it was fragmented by our chunking strategy. By failing to respect the original PDF structure, the strategy separated key ideas from their authors or sources.
- Second, retrieval failures occurred when embedding similarity caused semantic misalignment. For example, a question regarding training differences between REALM and DPR incorrectly retrieved a chunk on the BERT architecture. This mismatch stemmed from terminological overlap; shared terms like "encoder" and "pre-training" produced a high similarity score even though the correct passage was conceptually different. This highlights a fundamental limitation: in scientific contexts, functional relevance does not always correlate with semantic similarity.
- Third, generation failures, while rare at only 4%, stemmed from instruction-following errors. In these cases, the LLM defaulted to its internal parametric knowledge or misread the provided context, as seen in the second question of Table IV. While occasional, these instances prove that even with perfect context, accurate generation is not guaranteed. However, their low frequency confirms that modern instruction-tuned LLMs demonstrate high faithfulness when provided with relevant context.

VI. DISCUSSION

Our findings challenge the prevailing model-centric research paradigm and encourage a fundamental reorientation toward data-centric engineering in scientific RAG systems. By utilizing LLM-as-a-judge, not merely as a performance metric, but as a diagnostic instrument, we successfully isolated structural failures in retrieval coverage from cognitive failures in generation. This moves RAG evaluation toward an actionable diagnosis rather than a simple aggregate scoring.

A. Knowledge Base Coverage is the Key Obstacle

The dominant failure mode reveals a fundamental flaw in standard document processing. Standard tools, like the RecursiveCharacterTextSplitter, operate on "blind" logic, slicing text by character counts rather than rhetorical meaning. In a scientific context, this is catastrophic. A research paper is a web of connected arguments; when chunking separates a Method from its Constraint, it "break down" the knowledge. No model can reason through an answer that was physically deleted during ingestion.

B. Beyond Semantic Matching

We observe that semantic similarity is often a poor proxy for functional relevance [35]–[37]. Scientific queries often seek "differences" or "limitations". These concepts are "diffuse", meaning they are spread across paragraphs and lack the specific keywords found in the query. Dense retrievers are easily misled by "lexical noise", pulling chunks that look similar on the surface but are logically unrelated. This highlights a limitation: vectors measure how much text looks alike, not how ideas relate logically.

C. The Reliability of the LLM "Brain"

The remarkably low rate of Generation failure suggests that concerns over "hallucinations" may be misplaced in RAG systems. When provided with high-quality context, the LLM acts as a reliable synthesis engine. This implies that hallucinations are often a symptom of data absence, not a failure of logic. If the model lacks the correct data, it defaults to its own memory. Therefore, resources spent on fine-tuning models might be more effectively redirected toward expert data engineering.

D. Systematic Bias in Metadata

The consistent failure of bibliographic queries (Title/Author) reveals a positional bias. Because metadata is concentrated in specific regions like citation blocks, naive chunking splits this information into "weak" chunks that the retriever ignores. This

is not a random error; it is a systematic failure to treat document structure as a core part of the retrieval process.

These findings argue for a rebalancing of research priorities. The field has offered immense energy to improving retrievers, scaling models, and designing sophisticated fusion mechanisms. For scientific RAG, the path to reliability lies in upstream interventions. We must move away from "naive" text splitting and toward structure-aware ingestion; systems that understand a paper's layout are just as important as its texts.

VII. Limitations and Future Works

Several limitations should be considered when interpreting the findings of this study:

- Methodological circularity: Using an LLM (GPT-4) to judge a pipeline powered by an LLM (GPT-3.5) introduces potential circularity. While this study provides a diagnostic framework, these assessments have not yet been calibrated against human expert annotation.
- Generalizability: Our findings are derived from a single pipeline architecture and ten targeted articles. It remains unclear whether the "coverage-over-generation" failure mode persists across different chunking strategies, larger corpora, or specialized scientific knowledge graphs.
- To transition toward expert data engineering, future research must move beyond basic text ingestion. We suggest three concrete recommendations:
- Adopt structure-aware chunking strategies that preserve the rhetorical organization of academic writing.
- Implement hybrid retrieval approaches combining dense and sparse methods to capture functional relevance beyond semantic similarity.
- Adopt a multimodal approach that specifically addresses tabular data and equations, which were identified in Q&A Test Set 2 as a primary cause of coverage loss.

VIII. Conclusion

This study presents KnowRAG, a zero-shot retrieval-augmented generation (RAG) pipeline that achieves notably better factual accuracy compared to a Large Language Model alone. Our analysis reveals that the main limitation in performance is not retrieval or generator quality, but rather the extent of Knowledge Base coverage, with missing knowledge responsible for most errors. These results shift the focus away from scaling models, suggesting that reliable scientific AI relies more on advanced data engineering than larger models. The most impactful improvements in scientific RAG systems come from ambitious enhancements such as structure-aware document chunking, hybrid retrieval methods and more expressive knowledge representations. In essence, the effectiveness of an RAG system hinges on the quality of its knowledge base. By focusing on how we transform raw publications into retrievable knowledge, we can bridge the gap between existence in a PDF and presence in a model's response.

References

[1] D. Carmel, Y. Chang, H. Deng, and J. Y. Nie, "Future Directions of Query Understanding," Inf. Retr. Ser., vol. 46, pp. 205–224, 2020, doi: 10.1007/978-3-030-58334-7_9.

[2] C. Zhai, "Large Language Models and Future of Information Retrieval: Opportunities and Challenges," pp. 481–490, doi: 10.1145/3626772.3657848.

[3] A. Moutaoukkil and A. El Mezouary, "From Queries to Understanding: Designing the Next-Generation Search Engine," 2025 5th Int. Conf. Innov. Res. Appl. Sci. Eng. Technol. IRASET 2025, pp. 1–3, 2025, doi: 10.1109/IRASET64571.2025.11008326.

[4] N. Kandpal, H. Deng, A. Roberts, E. Wallace, and C. Raffel, "Large Language Models Struggle to Learn Long-Tail Knowledge," Proc. Mach. Learn. Res., vol. 202, pp. 15696–15707, 2023.

[5] S. Wu, Y. Cui, N. Guan, and C. J. Xue, "Retrieval-Augmented Generation for Natural Language Processing : A Survey."

[6] A. Vaswani, "Attention Is All You Need," no. Nips, 2017.

[7] N. R. Mannuru, "Large Language Models (LLMs) as a Tool to Facilitate Information Seeking Behavior," 2024.

[8] Y. Zhu et al., "Large Language Models for Information Retrieval: A Survey," pp. 1–35, 2023, [Online]. Available: http://arxiv.org/abs/2308.07107.

[9] Q. Ai et al., "Information Retrieval meets Large Language Models: A strategic report from Chinese IR community," AI Open, vol. 4, pp. 80–90, 2023, doi: 10.1016/j.aiopen.2023.08.001.

[10] A. Moutaoukkil, A. Idarrou, and I. Belahyane, "Information retrieval approaches: A comparative study," Int. J. Electr. Comput. Eng. Syst., vol. 13, no. 10, pp. 961–970, 2022, doi: 10.32985/ijeces.13.10.11.

[11] Z. Xu, "Hallucination is Inevitable : An Innate Limitation of Large Language Models."

[12] Y. Zhang et al., "Siren's Song in the AI Ocean: A Survey on Hallucination in Large Language Models," 2023, [Online]. Available: http://arxiv.org/abs/2309.01219.

[13] Z. Li, "The Dark Side of ChatGPT: Legal and Ethical Challenges from Stochastic Parrots and Hallucination," pp. 2–4, 2023, [Online]. Available: http://arxiv.org/abs/2304.14347.

[14] T. B. Brown et al., "Language Models are Few-Shot Learners," pp. 1154–1156, 2020, [Online]. Available: http://arxiv.org/abs/2005.14165.

[15] A. Radford, J. Wu, R. Child, D. Luan, D. Amodei, and I. Sutskever, "Language Models are Unsupervised Multitask Learners," 2018.

[16] J. Li, Y. Yuan, and Z. Zhang, "Enhancing LLM Factual Accuracy with RAG to Counter Hallucinations: A Case Study on Domain-Specific Queries in Private Knowledge-Bases," 2024, [Online]. Available: http://arxiv.org/abs/2403.10446.

[17] A. A. Khan, M. T. Hasan, K. K. Kemell, J. Rasku, and P. Abrahamsson, "Developing Retrieval Augmented Generation (RAG) based LLM Systems from PDFs: An Experience Report," pp. 1–36, 2024, [Online]. Available: http://arxiv.org/abs/2410.15944.

[18] A. Balaguer et al., "RAG vs Fine-tuning: Pipelines, Tradeoffs, and a Case Study on Agriculture," 2024, [Online]. Available: http://arxiv.org/abs/2401.08406.

[19] P. Lewis et al., "Retrieval-augmented generation for knowledge-intensive NLP tasks," Adv. Neural Inf. Process. Syst., vol. 2020-Decem, no. NeurIPS, 2020.

[20] P. Zhao, H. Zhang, Q. Yu, Z. Wang, and Y. Geng, "Retrieval-Augmented Generation for AI-Generated Content : A Survey," pp. 1–22.

[21] H. Li, Y. Su, D. Cai, Y. Wang, and L. Liu, "A Survey on Retrieval-Augmented Text Generation," 2022, [Online]. Available: http://arxiv.org/abs/2202.01110.

[22] N. Thakur, N. Reimers, J. Daxenberger, and I. Gurevych, "Augmented SBERT: Data Augmentation Method for Improving Bi-Encoders for Pairwise Sentence Scoring Tasks," vol. 2, pp. 296–310, 2021, doi: 10.18653/v1/2021.naacl-main.28.

[23] Y. Ke et al., "Development and Testing of Retrieval Augmented Generation in Large Language Models -- A Case Study Report," 2024, [Online]. Available: http://arxiv.org/abs/2402.01733.

[24] H. Emekci, "Maximizing RAG efficiency : A comparative analysis of RAG methods," pp. 1–25, 2024, doi: 10.1017/nlp.2024.53.

[25] K. Guu, K. Lee, Z. Tung, P. Pasupat, and M. W. Chang, "REALM: Retrieval-Augmented language model pre-training," 37th Int. Conf. Mach. Learn. ICML 2020, vol. PartF16814, pp. 3887–3896, 2020.

[26] K. Lee, M. W. Chang, and K. Toutanova, "Latent retrieval for weakly supervised open domain question answering," ACL 2019 - 57th Annu. Meet. Assoc. Comput. Linguist. Proc. Conf., pp. 6086–6096, 2020, doi: 10.18653/v1/p19-1612.

[27] W. Jiang, S. Zhang, B. Han, J. Wang, B. Wang, and T. Kraska, "PipeRAG: Fast Retrieval-Augmented Generation via Algorithm-System Co-design," 2024, [Online]. Available: http://arxiv.org/abs/2403.05676.

[28] H. Le et al., "FlauBERT: Unsupervised language model pre-training for French," Lr. 2020 - 12th Int. Conf. Lang. Resour. Eval. Conf. Proc., pp. 2479–2490, 2020.

[29] L. Martin et al., "CamemBERT: a Tasty French Language Model," pp. 7203–7219, 2020, doi: 10.18653/v1/2020.acl-main.645.

[30] Y. Wu, Z. Xu, T. Shi, Z. Wang, and S. Li, "Searching for Best Practices in Retrieval-Augmented Generation."

[31] S. Es, J. James, L. Espinosa-Anke, and S. Schockaert, "RAGAS: Automated Evaluation of Retrieval Augmented Generation," EACL 2024 - 18th Conf. Eur. Chapter Assoc. Comput. Linguist. Proc. Syst. Demonstr., pp. 150–158, 2024, doi: 10.18653/v1/2024.eacl-demo.16.

[32] G. B and A. Purwar, Evaluating the Efficacy of Open-Source LLMs in Enterprise-Specific RAG Systems: A Comparative Study of Performance and Scalability, vol. 1, no. 1. Association for Computing Machinery, 2024.

[33] L. Zheng et al., "Judging LLM-as-a-Judge with MT-Bench and Chatbot Arena," no. NeurIPS, pp. 1–29, 2023, [Online]. Available: http://arxiv.org/abs/2306.05685.

[34] G. Hardy Chen, S. Chen, Z. Liu, F. Jiang, and B. Wang, "Humans or LLMs as the Judge? A Study on Judgement Bias - analiza biasów ludzkich i modeli (są inaczej zbiasowane)," pp. 8301–8327, 2024, [Online]. Available: https://github.com/.

[35] H. Zamani and W. B. Cro, "Relevance-based Word Embedding," pp. 505–514, 2017.

[36] J. Guo, Y. Fan, Q. Ai, and W. B. Croft, "A deep relevance matching model for Ad-hoc retrieval," Int. Conf. Inf. Knowl. Manag. Proc., vol. 24-28-Octo, pp. 55–64, 2016, doi: 10.1145/2983323.2983769.

[37] K. Hui, A. Yates, K. Berberich, and G. de Melo, "PACRR: A position-aware neural IR model for relevance matching," EMNLP 2017 - Conf. Empir. Methods Nat. Lang. Process. Proc., no. July, pp. 1049–1058, 2017, doi: 10.18653/v1/d17-1110.

APPENDIX A: CORRECTNESS BY TOPICS IN FOUR QATESTSET

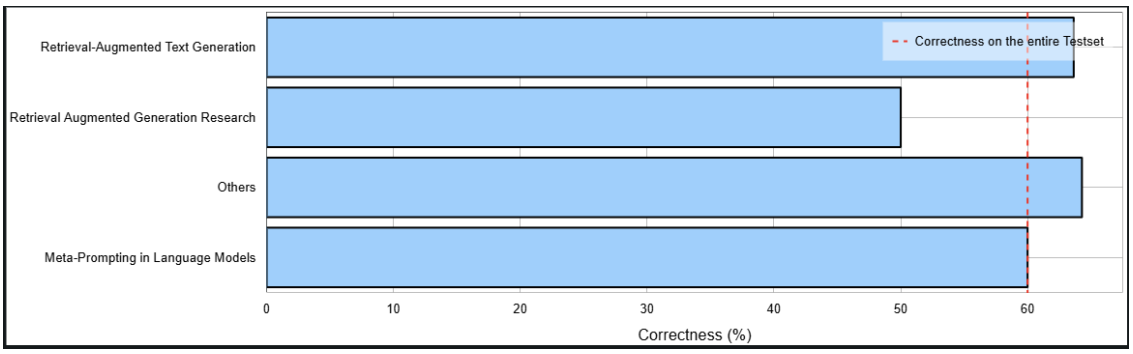

Fig. 2. Correctness by topics in the QATestSet1.

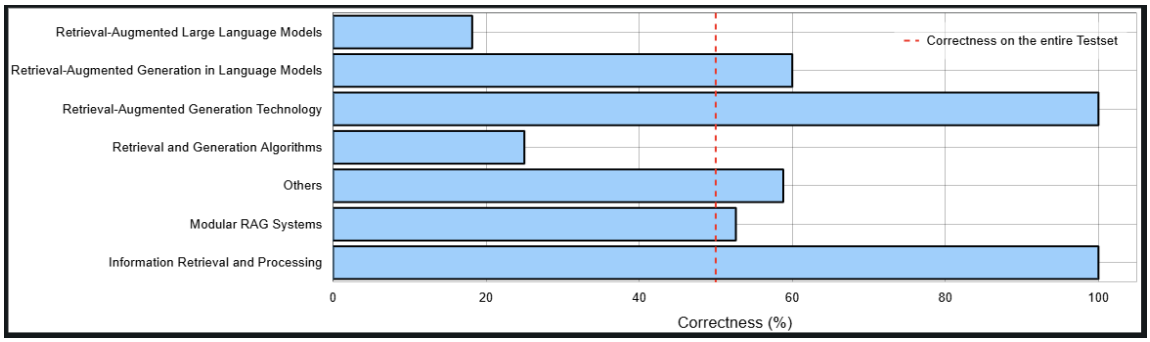

Fig. 3. Correctness by topics in the QATestSet2.

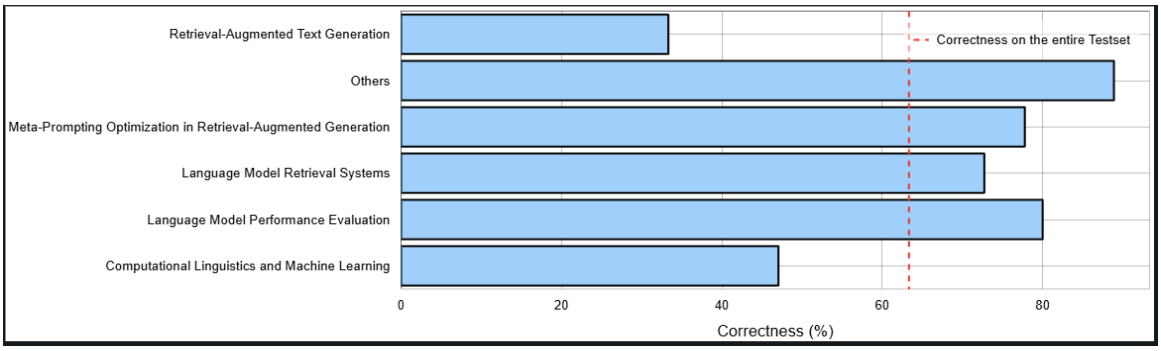

Fig. 4. Correctness by topics in the QATestSet3.

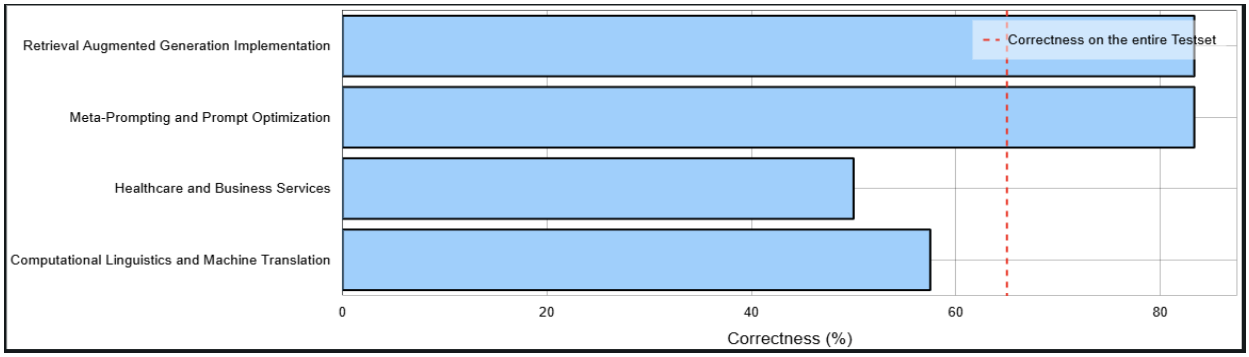

Fig. 5. Correctness by topics in the QATestSet4.

APPENDIX B: AN OVERVIEW OF THE KNOWLEDGE BASE OF FOUR QATESTSET

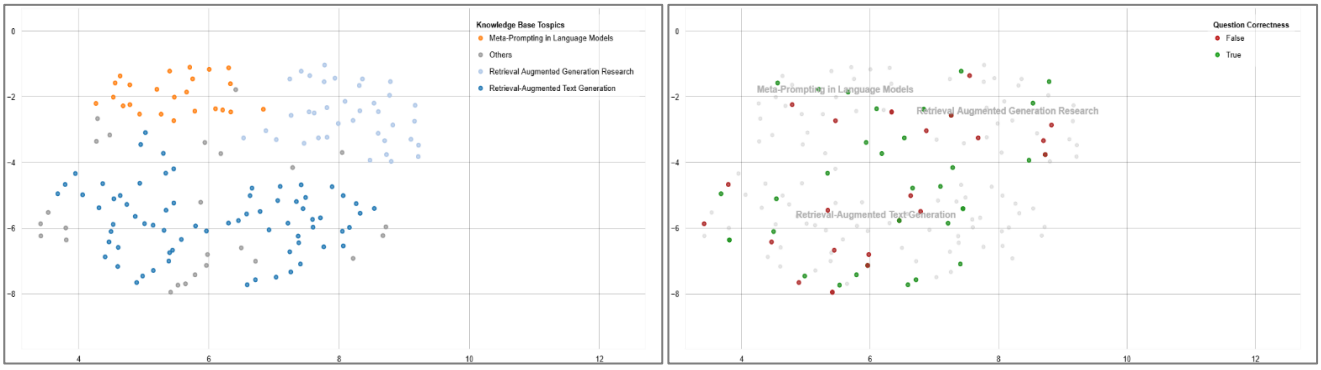

Fig. 6. An overview of the knowledge base of QATestSet1: Topics exploration and failures.

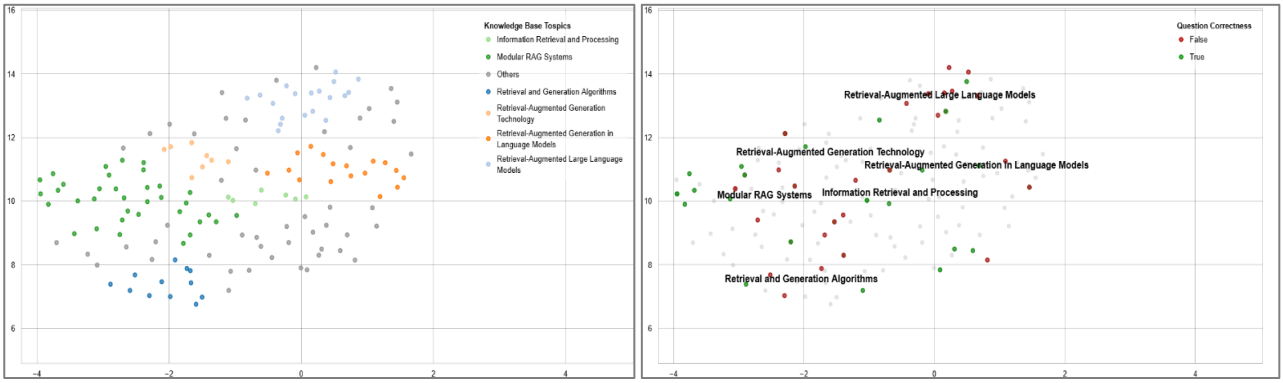

Fig. 7. An overview of the knowledge base of QATestSet2: Topics exploration and failures.

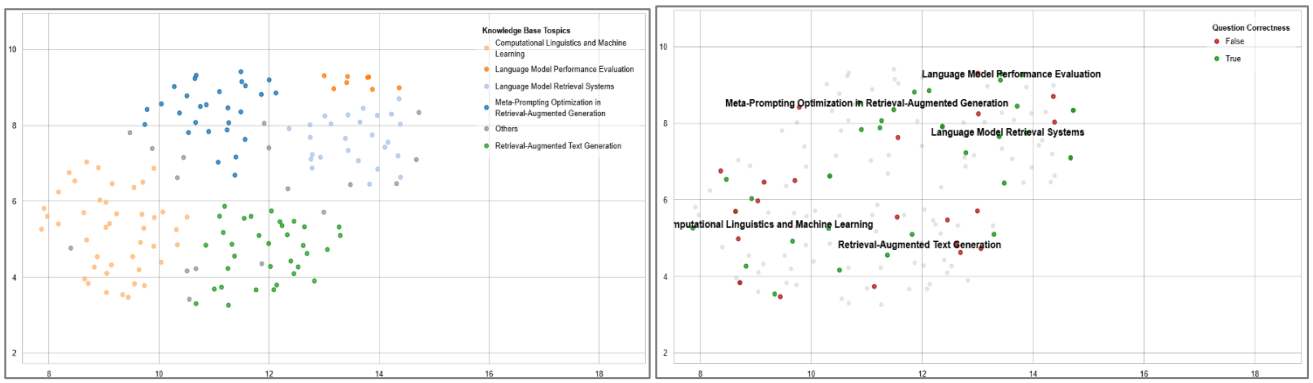

Fig. 8. An overview of the knowledge base of QATestSet3: Topics exploration and failures.

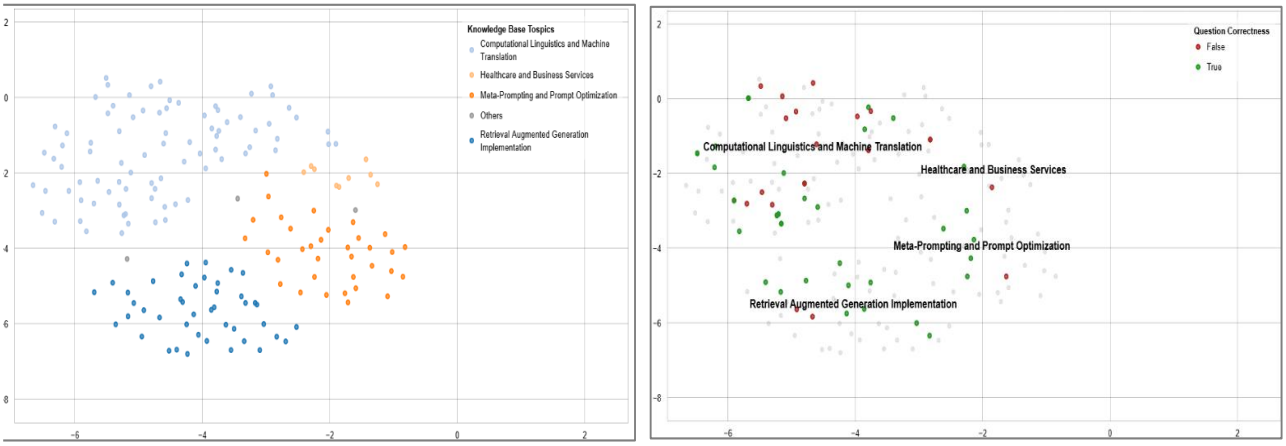

Fig. 9. An overview of the knowledge base of QATestSet4: Topics exploration and failures.