Towards AI-Native Memory: Tuning LLMs by RAG Principles

Anonymous ACL submission

Abstract

Long-context large language models (LLMs) and retrieval-augmented generation (RAG) are the two mainstream solutions to incorporate additional information into the generation process. In this paper, we first renovate and create three datasets, systematically compare these two types of solutions, and empirically show that (1) LLM-native solutions, although more expensive, shall be easier to capture the big picture and better answer *global* queries; and (2) when the queries concern *local* information, RAG solutions shall be more competitive especially when the keywords can be explicitly matched. Therefore, we further propose a novel method RAG-Tuned-LLM which fine-tunes a relative small (e.g., 7B) LLM using the data generated following the RAG principles, so it can combine the advantages of both solutions. This method leverages AI-native memory, enabling the model to more effectively integrate and retrieve information. Extensive experiments on three datasets demonstrate that RAG-Tuned-LLM can beat long-context LLMs and RAG methods on both local and global queries.

1 Introduction

Long-context large language models (LLMs) is an LLM-native solution to enable longer inputs to LLMs. For example, GPT-4-turbo and GPT-4o models can process 128K tokens (Achiam et al., 2023), and Gemini 1.5 claimed a context window of 1M or 10M tokens (Reid et al., 2024). There are also attempts towards "unlimited" context lengths from perspectives like length extrapolation (Peng et al., 2023; Xiao et al., 2023; Han et al., 2023; Zhang et al., 2024) and position bias (Liu et al., 2024; Peysakhovich and Lerer, 2023; An et al., 2024).

Retrieval-augmented generation (RAG), is another mainstream approach to incorporate additional information into the generation process. Specifically, RAG solutions (Lewis et al., 2020; Kočiskỳ et al., 2018; Pang et al., 2022; Trivedi et al., 2022; Edge et al., 2024) rely on a retriever to identify a smaller number of query-relevant contexts from a large corpus, and then feed them into a LLM to answer the query. Compared with long-context LLMs, its serving and inference cost is more affordable, and therefore, it is perhaps even more popular than long-context LLMs in the real-world applications.

In this paper, we first systematically compare these two types of methods on three datasets. We start with two public datasets, namely *news* articles (Tang and Yang, 2024) and *podcast* transcripts (Scott, 2024), follow the general ideas mentioned in Edge et al. (2024) to generate the queries and references, and further categorize the queries into two types, *local* and *global*. Specifically, local queries are retrieval-focused and detail-oriented, targeting specific information and concrete answers from small chunks of data. Global queries, on the other hand, require a broader, interpretive approach to synthesize insights by considering the dataset as a whole. We further introduce a new proprietary dataset containing *journaling* articles and user-provided local and global queries and their expected answers from a journaling app.[1]

Intuitively, (1) LLM-native solutions, although more expensive, shall be easier to capture the big picture and better answer *global* queries; and (2) when the queries concern *local* information, RAG solutions shall be more competitive especially when the keywords can be explicitly matched. Based on these three datasets, we run competitions between a vanilla RAG (Lewis et al., 2020) and Gemini 1.5 (Reid et al., 2024), the win rate results as shown in Table 2, confirmed our intuitions. It is worth mentioning that RAG surpasses long-context LLMs in all evaluation metrics when handling local

[1]The app name is masked for the blind review purpose.

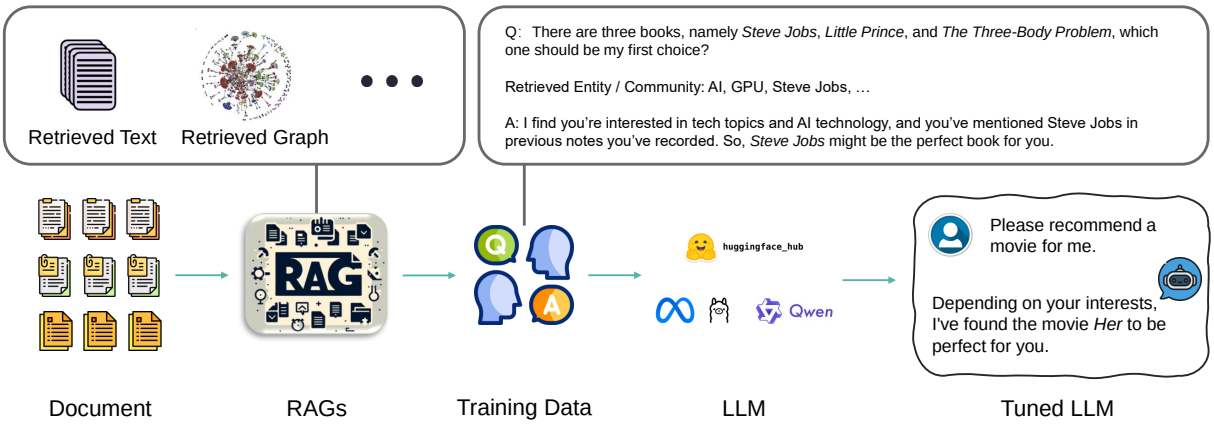

Figure 1: Overview of our method pipeline. **Stage 1**: Rag techniques serve as a foundation for synthesizing training data. **Stage 2**: The synthesized data is then employed to fine-tune a large language model (LLM) using Lora. **Stage 3**: Inference is conducted solely with AI-native memory. The fine-tuned model integrates the advantages of both the LLM-native solution and RAG.

queries, yet under-performs in addressing global ones.

Following our findings, we further propose a novel method RAG-Tuned-LLM which fine-tunes a relatively small (e.g., 7B) LLM using the data generated following the RAG principles, so it can combine the advantages of RAG and LLM-native solutions. It leverages AI-native memory, enabling the model to not only retrieve and utilize external information effectively but also to internalize knowledge in a way that allows it to maintain contextual coherence and handle different types of queries more naturally and efficiently. Specifically, as illustrated in Figure 1, we follow the Graph RAG (Edge et al., 2024) principles to extract entities, their interrelations, and the communities formed by these entities and relationships from plain text documents. We then generate data from both local and global perspectives: (1) local data synthesis concentrates on generating content-specific QA pairs, and (2) global data synthesis focuses on producing query-answering pairs that integrate insights across entities and relationships. With the data synthesized, fine-tuning the LLM becomes an intuitive task, and we employed the widely adopted LoRA technique (Hu et al., 2021).

Extensive experiments on three datasets demonstrate that RAG-Tuned-LLM can beat long-context LLMs and RAG methods on both local and global queries. The relatively small LLM obtained by RAG-Tuned-LLM consistently outperforms other methods, e.g., Graph RAG, long-context LLM, and vanilla RAG, on all metrics, particularly on the podcast and journaling datasets. RAG-Tuned-LLM excels in providing insightful and user-friendly responses, especially in tasks requiring deep comprehension.

Our contributions are summarized as follows.

- We create three datasets with local and global queries with their references, and then systematically compare LLM-native and (vanilla) RAG solutions, showing their unique advantages. It is worth mentioning that on one dataset, both the queries and references are manually created by human users.
- We follow the comparison results and propose a novel method RAG-Tuned-LLM to combine the advantages of RAG and LLM-native solutions.
- Extensive experiments on three datasets demonstrate that RAG-Tuned-LLM can indeed outperform long-context LLMs and (advanced) RAG methods on both local and global queries.

Reproducibility. We plan to release the *local* and *global* queries and their references on the two public datasets as well as our codes (including all the prompts) in a GitHub repo upon acceptance.

2 LLM-native vs. RAG Comparison

In this section, we systematically compare LLM-native and RAG solutions. An LLM-native solution to enable long inputs is to leverage **long-context LLMs**. For a fair comparison, we choose `Gemini-1.5-pro-001` model, because it is arguably the LLM that can handle the longest context window, i.e., 2 million tokens, so in our experiments, no truncation is necessary. We implemented a simple RAG solution, **VanillaRAG**, using standard embedding and reranking techniques from the

Table 1: Dataset statistics. **Training split:** This includes the number of instances, total tokens, and the average length of the input (Q) and output (A), respectively. **Evaluation split:** This includes the number of global and local queries, total tokens, and the average length of these queries.

Dataset	Training				Evaluation			
	# Instances	# Tokens	Avg Q Len	Avg A Len	Global	Local	# Tokens	Avg Q Len
Podcast	54,627	15,695,632	23.29	264.04	125	125	5,573	22.30
News	155,896	46,258,361	23.54	273.19	125	125	5,506	22.02
Journaling	18,355	10,995,778	36.46	562.60	45	15	2,374	39.57

Table 2: Wining rate of Gemini-1.5 against VanillaRAG on both local and global queries across three datasets.

Dataset	Metric	Local	Global	Overall
Podcast	Usefulness	81.60	86.40	84.00
	Richness	87.20	90.40	88.80
	Insightfulness	90.40	90.40	90.40
	User-Friendliness	85.60	88.80	87.20
News	Usefulness	46.40	56.60	51.20
	Richness	48.80	56.80	52.80
	Insightfulness	49.60	58.40	54.00
	User-Friendliness	46.40	58.40	52.40
Journaling	Usefulness	53.33	93.33	83.33
	Richness	46.67	88.80	80.00
	Insightfulness	53.33	91.11	81.67
	User-Friendliness	53.33	93.33	83.33

Langchain framework [2]. Specifically, it employs the `text-embedding-ada-002` model for initial chunk retrieval, selecting the top-10 most relevant chunks. These chunks are then refined using Cohere's `rerank-english-v3.0` model, which filters the set down to 3 chunks. We use `gpt-4o-mini` as the language model for answer generation due to its efficiency and performance. While `gpt-4o` offers the highest capability in text comprehension and generation, evaluations indicate that `gpt-4o-mini` delivers comparable results with significantly lower computational cost, making it a more practical choice for our tasks. By incorporating both embedding-based recall and reranking, VanillaRAG serves as a strong baseline for standard RAG approaches.

We create three datasets for our comparisons and subsequent extensive experiments. Table 1 presents key statistics for these datasets. The construction process is detailed in the following sections.

2.1 Dataset Renovation: News and Podcast

We start with two public datasets, namely **news** articles (Tang and Yang, 2024) and **podcast** transcripts (Scott, 2024), follow the general ideas mentioned in Edge et al. (2024) to generate the queries and references, and further categorize the queries into two types, *local* and *global*. Local queries focus on extracting precise, detailed information from specific, isolated segments of the data, prioritizing accuracy and concrete responses. In contrast, global queries demand a more holistic and interpretive approach, synthesizing insights by analyzing broader patterns and overarching themes throughout the entire dataset.

Query Generation. We follow Graph RAG (Edge et al., 2024) and generate both local and global queries for the news and podcast datasets, and handmade 60 queries for the journaling dataset. Overall, this method produces a total of 250 queries for two public datasets, ensuring a comprehensive and balanced evaluation set. Details of the queries can be found in Table 1.

Each query is carefully aligned with the user's persona and task description, ensuring that it is relevant to the specific challenge the user is addressing. This persona-task alignment guarantees that the queries are both meaningful and actionable, providing users with a well-rounded exploration of the dataset.

2.2 Dataset Creation: Journaling

We further introduce a new proprietary dataset containing **journaling** articles and user-provided local and global queries and their expected answers from a journaling app. The app name is masked for the blind review purpose. Specifically, it contains 60 carefully crafted queries. These queries were designed to assess various reasoning capabilities in more complex and nuanced data scenarios, emphasizing the reasoning capabilities instead of simple retrieval.

To align with our evaluation framework, we further divided these four reasoning categories into local and global subsets. The local subset includes Memory-related queries, which are primarily focused on retrieving specific pieces of information

[2] LangChain: https://www.langchain.com/

from a narrow portion of the dataset. The global subset, in contrast, includes queries that require broader reasoning across the dataset, such as those aimed at understanding patterns, predicting future trends, and making recommendations. These queries demand interpretive analysis that synthesizes information from various parts of the dataset, requiring models to generate more comprehensive and context-aware responses. This distinction between local and global queries enables us to systematically assess a model's ability to handle both detailed retrieval and more holistic, global reasoning tasks.

This dataset provides a robust tool for evaluating models' capacity to engage with complex reasoning tasks, offering insights into how well they can both retrieve specific details and perform higher-order reasoning. By integrating this benchmark into our experiments, we can evaluate the models' performance in diverse real-world reasoning scenarios beyond simple fact retrieval.

2.3 Evaluation Metrics

We design our evaluation criteria to emphasize the user-centered perspective, ensuring that the generated answers are not only accurate but also practically useful and accessible for real-world tasks. This shift from system-centric evaluation to user-oriented assessment is reflected in the following four metrics:

- **Usefulness / Correctness**: Assesses accuracy, relevance, and value of the answer in addressing the query.
- **Richness / Diversity**: Measures the breadth, depth, and variety of perspectives in the answer.
- **Insightfulness / Deep Understanding**: Evaluates the depth of comprehension and originality of insights.
- **User-Friendliness**: Focuses on clarity, readability, and ease of understanding.

A more detailed and explicit explanation is deterred to the appendix B.

We employ an LLM-based evaluator, specifically `GPT-4o-mini`[3], to evaluate the performance of various methods against the established metrics. This evaluator is provided with the query, target metric, and a pair of answers. Additionally, we incorporate the context information as a reference during the evaluation. This context, collected manually, serves as a baseline for comparing the two answers, ensuring that the evaluator had a more robust framework to judge the relevance and quality of the responses. We repeat this evaluation process multiple times to account for stochastic variation.

We employ a winning rate approach for evaluation, focusing specifically on comparisons between our method and each individual baseline method. Direct comparisons between the baseline methods themselves were not conducted.

2.4 Results

The results in Table 2 support our initial assumptions: (1) LLM-native models, while costlier, perform better on *global* queries requiring broader context understanding, and (2) RAG models excel on *local* queries, particularly when explicit keyword matching is involved. Across the three datasets, Gemini 1.5 (Reid et al., 2024) outperforms vanilla RAG (Lewis et al., 2020) in global query metrics, especially in the Journaling dataset where it achieves over 90% win rates for usefulness and user-friendliness. Conversely, RAG consistently surpasses LLMs on local queries, confirming the complementary strengths of both approaches.

3 Our RAG-Tuned-LLM

Our comparison between LLM-native and RAG solutions suggests the potential benefits of combining RAG with LLMs fine-tuned through supervised learning (SFT). Specifically, RAG's strength in generating detailed, high-quality local and global QA pairs can complement LLM's broader understanding, creating a more capable system. By training on data synthesized from both granular (local) and comprehensive (global) perspectives, the LLM can leverage the advantages of both methods, improving its ability to deliver precise and context-aware responses.

This section details the methodology for our data synthesis process, focusing on creating QA pairs from entities, relationships, and text units. Each component integrates structured and unstructured information to generate high-quality training data, ensuring the robustness and relevance of the final model.

3.1 Graph RAG: A Recent Advanced RAG

To rigorously evaluate our proposed method, we compared it against several baseline approaches. These baselines were selected to represent diverse strategies for retrieval-augmented generation (RAG) and long-context query answering systems:

[3]We selected this model due to GPT-4o-mini's competitive performance and low query costs.

Table 3: Graph statistics for the Podcast, News, and Journaling datasets.

Dataset	Entities	Relationships	Communities
Podcast	5,182	8,631	837
News	17,877	26,208	3,534
Journaling	2,930	3,751	547

- **Graph RAG**: We adopted the state-of-the-art RAG, Graph RAG, for comparison, using four specific conditions (C0 to C3) to ensure a comprehensive evaluation. Graph RAG's multi-stage RAG mechanism allows for rich context representation and retrieval. By using these distinct configurations, we aimed to test how well each condition performed in capturing and utilizing relevant information from the dataset for answering complex querys.
- **Long-context LLM**: To account for the impact of long-context processing capabilities, we included a long-context LLM in our baselines. Specifically, we used the gemini-1.5-pro-001 model, which supports a 2 million token context window. This was chosen to ensure that no relevant contextual information would be lost due to limitations in context length, providing a fair comparison against other methods, including our own, that integrate long-range dependencies from the dataset.
- **Vanilla RAG**: As another baseline, we implemented a simplified Vanilla RAG, using standard embedding and reranking techniques from the Langchain framework. This method employed the text-embedding-ada-002 model for initial chunk retrieval, selecting the top 10 most relevant chunks. These chunks were then refined using Cohere's rerank-english-v3.0 model, which filtered the set down to 3 chunks. The final stage involved using gpt-4o-mini as the language model for generating the answers. By incorporating both embedding-based recall and reranking, Vanilla RAG serves as a strong baseline for standard RAG approaches.

These baseline methods allowed for a thorough and multi-faceted comparison, enabling us to demonstrate the advantages of our approach across various retrieval, summarization, and reasoning tasks.

3.2 Global Data Synthesis

Our method begins by focusing on entities, which serve as the foundation of global data synthesis. For each entity, we generate descriptive training data using templates that explore key attributes in depth. These templates produce natural language querys that encourage a broad exploration of the entity's role within a larger context. To ensure detailed and coherent answers, we apply a chain-of-thought (CoT) (Wei et al., 2022) reasoning framework that integrates the entity's description with relevant global information. This structured approach involves:

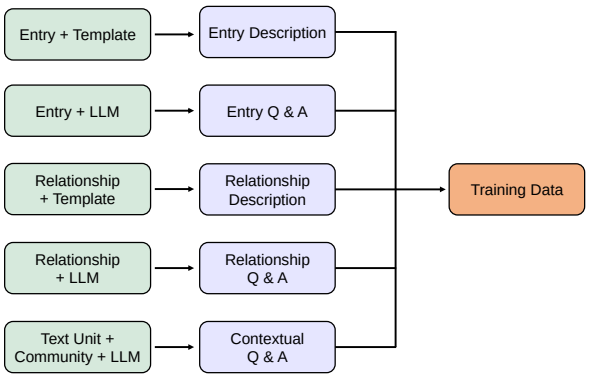

Figure 2: Overview of the data synthesis process used in RAG-Tuned-LLM. The process begins with the generation of entity descriptions and question-answer (Q&A) pairs through the combination of templates and large language models (LLMs), applied to both individual entries and relationships between entities. Additionally, text units enriched by community input are used to enhance the Q&A generation around entities & relationships. These components collectively form a comprehensive set of training data.

1. **Restating the context**: Ensures coherence and smooth flow in the response.
2. **Integrating entity description**: Merges entity attributes with the global context to generate meaningful responses.
3. **Constructing a detailed answer**: Provides a thorough explanation, with a minimum of 300-500 words based on available context.

Markdown formatting, like subheadings and bullet points, enhances clarity and structure. This approach ensures that entity-based QA pairs capture both detailed and global information, critical for training models with comprehensive understanding.

Next, we focus on relationships between entities, which are crucial for understanding global context through their interactions. We generate querys using relationship-specific templates that probe the nature and dynamics between entities within their global context. The CoT reasoning framework is also applied here to derive thoughtful answers by:

1. **Restating the relationship context**: Maintains coherence by anchoring answers in global context.
2. **Incorporating relationship descriptions**: Integrates relationship dynamics with contextual information.
3. **Providing comprehensive responses**: Ensures clarity and relevance, with 300-500 word answers depending on complexity.

By highlighting how entities and their relationships fit within the global context, this method ensures rich, globally-informed QA pairs that enhance model performance in broader scenarios.

3.3 Local Data Synthesis

The third aspect of our data synthesis focuses on generating querys from text chunks that contain multiple entities and relationships, but with a focus on local details. These text chunks serve as the contextual information, helping to refine querys that probe specific, localized aspects of the entities or relationships. The workflow involves:

1. **Assessing local information**: The text chunk is first evaluated to identify the relevant entities or relationships. The focus is on the specific details that these elements present in the given context.
2. **Generating detailed, context-specific querys**: querys are then crafted based on how these entities or relationships operate within the localized scope, utilizing the text chunk as the immediate context.

These localized querys target specific interactions or characteristics within the text, offering detailed insights into the smaller components of the data. This method emphasizes how entities or relationships function within a confined context, making it particularly effective for training models to handle nuanced, local-level reasoning.

3.4 LM tuning

By combining entity-based, relationship-based, and localized context-based query generation, this framework ensures that the training data captures both the granular details and broader patterns. Therefore, we can fine-tune a LLM to render the knowledge extracted through RAG as AI-native. Typically, there are two primary strategies for fine-tuning an LLM: full fine-tuning (Lv et al., 2023) and parameter-efficient fine-tuning (PEFT) (Ding et al., 2023). Full fine-tuning adjusts all the parameters within the LLM, generally offering a higher performance ceiling but at the cost of significant computational resources and more extensive training data. In contrast, PEFT focuses on fine-tuning only a small subset of the LLM's parameters (e.g., BitFit (Zaken et al., 2021) only fine-tunes bias terms) or incorporating a lightweight module (Houlsby et al., 2019; Hu et al., 2021), leaving the base LLM intact. Over recent years, PEFT has demonstrated impressive results comparable to full fine-tuning while requiring significantly fewer GPU resources. For this reason, we predominantly adopt the PEFT approach, specifically LoRA (Hu et al., 2021), grounded in the auto-regressive principle in this paper.

4 Experiments

4.1 Setup

We choose `Qwen-2-7B-instruct` (Bai et al., 2023a) for its good performance in both Chinese and English, and LoRA as the method to parameterize the knowledge. In our training, the LoRA rank is set to 64, and we adopt a cosine learning rate scheduler with a max learning rate of 0.0001. The total fine-tuning epoch is set to 3. We set the decoding temperature as 0 for more stable results.

4.2 Results

Table 4 presents the winning rates of our RAG-tuned model, equipped with AI-native memory, compared against three baseline methods—Graph RAG (C0 to C3), long-context LLM, and Vanilla RAG—across three distinct datasets: podcast, news, and our custom journaling dataset. Each dataset contains either 250 cases (podcast, news) or 60 cases (journaling), and the percentages represent the proportion of cases where our model's responses were judged superior by the LLM evaluator based on four key metrics: *Usefulness*, *Richness*, *Insightfulness*, and *User-Friendliness*. A score exceeding 50% indicates that our model outperformed the baseline in that particular comparison.

Podcast dataset.. For the podcast dataset, our model demonstrates strong performance across all metrics in comparison to the Graph RAG conditions, with notable victories in *User-Friendliness*, achieving over 70% winning rates across all Graph RAG configurations. It consistently excelled in *Insightfulness* as well, particularly in the C0 condition with a 59.20% win rate. This indicates our model's capability in delivering insightful and nuanced answers to queries derived from the conver-

Table 4: The overall win rates, averaged across local and global queries, of our fine-tuned model—equipped with AI-native memory—against Graph RAG, long-context LLM, and Vanilla RAG across three datasets (Podcast and News with 250 cases each, and Journaling with 60 cases). After aligning with human evaluation, we found that for 250 cases, the human assessments matched the LLM's evaluation in 215 cases, achieving an alignment rate of 86%. Results of local and local subsets can be found at Table 6 and Table 5.

Dataset	Metric	Graph RAG C0	Graph RAG C1	Graph RAG C2	Graph RAG C3	Long-Context LLM	Vanilla RAG
Podcast	Usefulness	55.60	54.40	52.40	52.40	66.80	96.40
	Richness	52.40	49.60	47.20	46.80	60.40	96.80
	Insightfulness	59.20	53.60	52.00	50.80	61.20	99.20
	User-Friendliness	81.20	76.40	72.80	70.40	78.00	89.60
News	Usefulness	52.40	53.20	50.40	51.20	96.00	96.80
	Richness	50.80	49.20	46.00	46.40	95.20	99.20
	Insightfulness	56.40	55.60	51.60	52.00	96.80	99.60
	User-Friendliness	78.80	73.60	70.80	68.80	91.20	85.20
Journaling	Usefulness	56.67	53.33	53.33	51.67	63.33	88.33
	Richness	53.33	56.67	55.00	53.33	71.67	96.67
	Insightfulness	68.33	66.67	61.67	66.67	80.00	96.67
	User-Friendliness	56.67	55.00	56.67	53.33	71.67	86.67

sational nature of podcast transcripts, where deep comprehension and detailed responses are crucial.

In comparison with the long-context LLM, our model displayed competitive performance, particularly in *Usefulness* (66.80%) and *User-Friendliness* (78.00%), proving its ability to deliver useful and accessible content despite not having a large context window. Moreover, RAG-Tuned-LLM significantly outperformed Vanilla RAG among all perspectives, which shows that the ability to retrieve and rerank a large number of relevant chunks can't lead to precise and detailed answers.

News dataset.. In the news dataset, our model exhibits a performance trend similar to that of the podcast dataset, demonstrating strengths in both *User-Friendliness* and *Insightfulness*. However, the long-context LLM performs notably poorly across all aspects, as reflected in its high loss rate. Based on the statistics in Table 1 and the experimental results in Table 4, it is evident that the larger the total token count in the dataset, the worse the performance of the long-context model. Given that the total token count in the news dataset is significantly higher than in the other two datasets, we attribute the catastrophic drop in performance to the excessive length of the context.

Journaling dataset.. In the journaling dataset, which focuses on complex reasoning tasks, our model excels in both *Insightfulness* and *Richness*. It achieves an impressive win rate of 88.33% in *Insightfulness* against Vanilla RAG and performs notably well in *Richness* across the Graph RAG conditions. This highlights the model's ability to offer not only insightful responses but also a richer, more nuanced perspective when reasoning through complex, real-world tasks that demand both depth and diversity in content.

4.3 Cross-Dataset Observations

Across all datasets, our model consistently shows strength in *Usefulness* and *Insightfulness*, metrics that are most critical from the user's perspective. These metrics highlight our model's ability to provide correct, relevant answers while offering deep, thoughtful insights. In contrast, *Richness* remains a challenge, but the fact that our model performs at a competitive level against Graph RAG in this regard is already a significant achievement. Our goal is not to surpass but to closely align with Graph RAG in terms of *Richness*, which we believe is a realistic and attainable objective in the near future.

5 Related Works

5.1 From RAG to Graph RAG

Pre-trained language models, such as Qwen (Bai et al., 2023b) and Llama (Touvron et al., 2023a), demonstrated remarkable query-answering capabilities. However, they fall short when confronted with problems requiring knowledge beyond their training data, particularly querys related to users' personal information. In this context, retrieval-augmented generation (RAG) (Lewis et al., 2020) provides a solution by sourcing the necessary information from an external knowledge base. While RAG has proven both practical and effective, certain challenges persist. For instance, conventional RAG retrieves only fine-grained details relevant to the query, with broader comprehension

largely reliant on the LLM itself. Consequently, abstract querys, such as inquiries regarding the user's MBTI, often yield unsatisfactory answers. To address these shortcomings, Graph RAG (Edge et al., 2024) has been introduced, constructing a knowledge graph based on the LLM. Upon receiving a query, Graph RAG can provide hierarchical responses, ranging from detailed local insights to more global abstractions, leveraging the knowledge graph for enhanced understanding.

5.2 External Knowledge vs. Internal Memory

To mitigate the emergence of hallucinations, contemporary LLMs typically draw upon additional knowledge from external sources, such as RAG. In the quest for enhanced efficiency and effectiveness, RAG has undergone a continuous evolution since its inception, advancing from Vanilla RAG (Lewis et al., 2020) to advanced RAG (Ilin, 2023) and Modular RAG (Yu et al., 2022; Shao et al., 2023), as summarized by (Zhu et al., 2023). Most recently, Microsoft introduced the Graph RAG project (Edge et al., 2024), further broadening the scope of the RAG paradigm. In contrast to RAG's dependence on external knowledge, another avenue of research concentrates on enhancing the intrinsic memory and capabilities of LLMs themselves (Shang et al., 2024; Salemi and Zamani, 2024; Tan et al., 2024), typically achieved through fine-tuning. We contend that such AI-native memory empowers the model to attain superior abstraction capabilities and a more comprehensive understanding of the context on a global scale. Alternatively, providing long-context LLMs (Team et al., 2023; Achiam et al., 2023) with the complete document may also facilitate the abstraction of information pertinent to user queries and yield a broader understanding. However, as articulated by Hsieh et al. (2024); Shang et al. (2024), the context may surpass the constraints of the LLM's context window, which is typically much narrower than reported, thereby leading to the potential loss of vital information amidst a vast expanse of text. This paper endeavors to demonstrate that within the realm of RAG-Tuned-LLM, those endowed with AI-native memory about users surpass the efficacy of relying on external knowledge.

5.3 Fine-Tuning LLMs

To incrementally expand the knowledge of a pre-trained LLM or to align it with human preferences, fine-tuning stands as one of the most prevalent approaches, encompassing methods such as supervised fine-tuning (SFT), reinforcement learning from human feedback (RLHF) (Ouyang et al., 2022), and direct preference optimization (DPO) (Rafailov et al., 2024). Despite their effectiveness, these techniques are notably annotation-heavy and computationally intensive, rendering the fine-tuning of an LLM using these methods prohibitively costly. To circumvent the extensive computational demands of full fine-tuning, which can reach into tens of billions, numerous parameter-efficient fine-tuning (PEFT) methods have been explored, including adapter (Houlsby et al., 2019) and Lora (Hu et al., 2021). In this paper, we primarily employ Graph RAG and LLM to generate training datasets, subsequently utilizing Lora to fine-tune a RAG-Tuned-LLM.

6 Conclusion and Future Work

In this paper, we validate RAG's fine-grained retrieval abilities and the global abstraction strengths of LLM-native solutions, such as long-context LLMs. However, RAG lacks holistic understanding, and long-context models tend to lose key information over extended contexts. We integrate these strengths by fine-tuning an LLM within a RAG framework for data generation. This work is the first to explore LLM and RAG integration within a unified framework, bridging open-domain and domain-specific QA tasks. Our RAG-tuned LLM, equipped with AI-native memory, outperforms both standard RAG methods and long-context LLMs across diverse datasets, demonstrating superior performance in handling hierarchical queries.

Future Work. Building on this study, several future directions are worth exploring to further validate and enhance our proposed method. First, we plan to extend RAG-Tuned-LLM to more diverse datasets and domains, enabling us to evaluate its generalizability across different tasks, including complex challenges like multi-hop reasoning and multi-modal QA. This will provide a clearer understanding of RAG-Tuned-LLM's effectiveness in both open-domain and domain-specific contexts. Additionally, we will experiment with various foundational models (e.g., the Llama series (Touvron et al., 2023a,b)), testing RAG-Tuned-LLM's adaptability to different architectures and model scales. This will highlight the trade-offs between model size, computational efficiency, and performance when combining RAG and LLM-native methods.

Limitations

While our proposed method RAG-Tuned-LLM demonstrates notable advantages over long-context LLMs and RAG in handling both global and local questions, it also presents several limitations that warrant further investigation.

First, scalability remains a challenge. Although our method effectively combines RAG and LLM-native capabilities, fine-tuning even relatively small models (e.g., 7B parameters) can be computationally intensive, particularly when applied to large-scale datasets or complex domains. This poses a constraint on its applicability in resource-constrained environments where computational power or memory is limited. Further optimization of training and inference efficiency would be necessary to make RAG-Tuned-LLM more accessible for broader use cases. To mitigate bias introduced by a single model family, future studies should incorporate a more diverse range of models.

Second, the generalizability of RAG-Tuned-LLM across diverse tasks and domains has not yet been fully tested. Our experiments focus primarily on open-domain and domain-specific QA tasks within a controlled environment. While the results are promising, applying RAG-Tuned-LLM to more complex or specialized tasks, such as multi-modal or multi-hop reasoning, is needed to fully understand its versatility. Moreover, certain niche or domain-specific datasets may require tailored adaptations of the model to achieve optimal performance.

Due to the privacy protection protocol, the journaling dataset constructed from personal users remains inaccessible to the public[4]. However, this dataset significantly broadens the scope of the study, rendering the conclusions of this paper more general.

Another limitation relates to the *AI-native memory* concept. While our method introduces a memory component by fine-tuning LLMs on data generated with RAG principles, the current exploration of memory retention and recall mechanisms remains preliminary. The memory module's long-term effectiveness in highly dynamic contexts or in cases where the context evolves over multiple interactions has not been fully evaluated. Further development of more sophisticated memory-aware strategies, as well as experiments in iterative, evolving environments, is necessary to better understand the potential of AI-native memory.

Additionally, the performance trade-offs between model size, computational cost, and overall effectiveness remain an area of concern. While we show that smaller models fine-tuned with RAG principles can outperform larger long-context LLMs, the relationship between model size, efficiency, and performance in more varied scenarios has yet to be fully understood. The extent to which these findings can be generalized to different foundational models (e.g., Llama) with different architectures is an open question.

Finally, despite the positive results in controlled settings, our method's real-world applicability requires more extensive testing. In dynamic, real-time environments, such as interactive personal assistant applications, models often need to adapt quickly to changing contexts and evolving user needs. While initial experiments with real-world data have shown promise, evaluating RAG-Tuned-LLM in more diverse and interactive scenarios will be critical to understanding its practical limits and potential for continual learning.

In summary, while RAG-Tuned-LLM pushes the boundaries of current QA models by integrating RAG and LLM-native solutions, several limitations related to scalability, generalizability, memory mechanisms, performance trade-offs, and real-world application remain to be addressed in future work.

[4]It is important to note that we have obtained the users' approval for the purposes of this paper.

References

Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, et al. 2023. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*.

Shengnan An, Zexiong Ma, Zeqi Lin, Nanning Zheng, and Jian-Guang Lou. 2024. Make your llm fully utilize the context. *arXiv preprint arXiv:2404.16811*.

Jinze Bai, Shuai Bai, Yunfei Chu, Zeyu Cui, Kai Dang, Xiaodong Deng, Yang Fan, Wenbin Ge, Yu Han, Fei Huang, Binyuan Hui, Luo Ji, Mei Li, Junyang Lin, Runji Lin, Dayiheng Liu, Gao Liu, Chengqiang Lu, Keming Lu, Jianxin Ma, Rui Men, Xingzhang Ren, Xuancheng Ren, Chuanqi Tan, Sinan Tan, Jianhong Tu, Peng Wang, Shijie Wang, Wei Wang, Shengguang Wu, Benfeng Xu, Jin Xu, An Yang, Hao Yang, Jian Yang, Shusheng Yang, Yang Yao, Bowen Yu, Hongyi Yuan, Zheng Yuan, Jianwei Zhang, Xingxuan Zhang, Yichang Zhang, Zhenru Zhang, Chang Zhou, Jingren Zhou, Xiaohuan Zhou, and Tianhang

Zhu. 2023a. Qwen technical report. *arXiv preprint arXiv:2309.16609*.

Jinze Bai, Shuai Bai, Yunfei Chu, Zeyu Cui, Kai Dang, Xiaodong Deng, Yang Fan, Wenbin Ge, Yu Han, Fei Huang, et al. 2023b. Qwen technical report. *arXiv preprint arXiv:2309.16609*.

Ning Ding, Yujia Qin, Guang Yang, Fuchao Wei, Zonghan Yang, Yusheng Su, Shengding Hu, Yulin Chen, Chi-Min Chan, Weize Chen, et al. 2023. Parameter-efficient fine-tuning of large-scale pre-trained language models. *Nature Machine Intelligence*, 5(3):220–235.

Darren Edge, Ha Trinh, Newman Cheng, Joshua Bradley, Alex Chao, Apurva Mody, Steven Truitt, and Jonathan Larson. 2024. From local to global: A graph rag approach to query-focused summarization. *arXiv preprint arXiv:2404.16130*.

Chi Han, Qifan Wang, Wenhan Xiong, Yu Chen, Heng Ji, and Sinong Wang. 2023. Lm-infinite: Simple on-the-fly length generalization for large language models. *arXiv preprint arXiv:2308.16137*.

Neil Houlsby, Andrei Giurgiu, Stanislaw Jastrzebski, Bruna Morrone, Quentin De Laroussilhe, Andrea Gesmundo, Mona Attariyan, and Sylvain Gelly. 2019. Parameter-efficient transfer learning for nlp. In *International conference on machine learning*, pages 2790–2799. PMLR.

Cheng-Ping Hsieh, Simeng Sun, Samuel Kriman, Shantanu Acharya, Dima Rekesh, Fei Jia, and Boris Ginsburg. 2024. Ruler: What's the real context size of your long-context language models? *arXiv preprint arXiv:2404.06654*.

Edward J Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. 2021. Lora: Low-rank adaptation of large language models. *arXiv preprint arXiv:2106.09685*.

Ivan Ilin. 2023. Advanced rag techniques: an illustrated overview.

Tomáš Kočiskỳ, Jonathan Schwarz, Phil Blunsom, Chris Dyer, Karl Moritz Hermann, Gábor Melis, and Edward Grefenstette. 2018. The narrativeqa reading comprehension challenge. *Transactions of the Association for Computational Linguistics*, 6:317–328.

Patrick Lewis, Ethan Perez, Aleksandra Piktus, Fabio Petroni, Vladimir Karpukhin, Naman Goyal, Heinrich Küttler, Mike Lewis, Wen-tau Yih, Tim Rocktäschel, et al. 2020. Retrieval-augmented generation for knowledge-intensive nlp tasks. *Advances in Neural Information Processing Systems*, 33:9459–9474.

Nelson F Liu, Kevin Lin, John Hewitt, Ashwin Paranjape, Michele Bevilacqua, Fabio Petroni, and Percy Liang. 2024. Lost in the middle: How language models use long contexts. *Transactions of the Association for Computational Linguistics*, 12:157–173.

Kai Lv, Yuqing Yang, Tengxiao Liu, Qinghui Gao, Qipeng Guo, and Xipeng Qiu. 2023. Full parameter fine-tuning for large language models with limited resources. *arXiv preprint arXiv:2306.09782*.

Long Ouyang, Jeffrey Wu, Xu Jiang, Diogo Almeida, Carroll Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, et al. 2022. Training language models to follow instructions with human feedback. *Advances in neural information processing systems*, 35:27730–27744.

Richard Yuanzhe Pang, Alicia Parrish, Nitish Joshi, Nikita Nangia, Jason Phang, Angelica Chen, Vishakh Padmakumar, Johnny Ma, Jana Thompson, He He, et al. 2022. Quality: Question answering with long input texts, yes! In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 5336–5358.

Bowen Peng, Jeffrey Quesnelle, Honglu Fan, and Enrico Shippole. 2023. Yarn: Efficient context window extension of large language models. In *The Twelfth International Conference on Learning Representations*.

Alexander Peysakhovich and Adam Lerer. 2023. Attention sorting combats recency bias in long context language models. *arXiv preprint arXiv:2310.01427*.

Rafael Rafailov, Archit Sharma, Eric Mitchell, Christopher D Manning, Stefano Ermon, and Chelsea Finn. 2024. Direct preference optimization: Your language model is secretly a reward model. *Advances in Neural Information Processing Systems*, 36.

Machel Reid, Nikolay Savinov, Denis Teplyashin, Dmitry Lepikhin, Timothy Lillicrap, Jean-baptiste Alayrac, Radu Soricut, Angeliki Lazaridou, Orhan Firat, Julian Schrittwieser, et al. 2024. Gemini 1.5: Unlocking multimodal understanding across millions of tokens of context. *arXiv preprint arXiv:2403.05530*.

Alireza Salemi and Hamed Zamani. 2024. Comparing retrieval-augmentation and parameter-efficient fine-tuning for privacy-preserving personalization of large language models. *arXiv preprint arXiv:2409.09510*.

Kevin Scott. 2024. [link].

Jingbo Shang, Zai Zheng, Xiang Ying, Felix Tao, and Mindverse Team. 2024. Ai-native memory: A pathway from llms towards agi. *arXiv preprint arXiv:2406.18312*.

Zhihong Shao, Yeyun Gong, Yelong Shen, Minlie Huang, Nan Duan, and Weizhu Chen. 2023. Enhancing retrieval-augmented large language models with iterative retrieval-generation synergy. *arXiv preprint arXiv:2305.15294*.

Zhaoxuan Tan, Qingkai Zeng, Yijun Tian, Zheyuan Liu, Bing Yin, and Meng Jiang. 2024. Democratizing large language models via personalized parameter-efficient fine-tuning. *arXiv preprint arXiv:2402.04401*.

Yixuan Tang and Yi Yang. 2024. Multihop-rag: Benchmarking retrieval-augmented generation for multi-hop queries. *arXiv preprint arXiv:2401.15391*.

Gemini Team, Rohan Anil, Sebastian Borgeaud, Yonghui Wu, Jean-Baptiste Alayrac, Jiahui Yu, Radu Soricut, Johan Schalkwyk, Andrew M Dai, Anja Hauth, et al. 2023. Gemini: a family of highly capable multimodal models. *arXiv preprint arXiv:2312.11805*.

Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, et al. 2023a. Llama: Open and efficient foundation language models. *arXiv preprint arXiv:2302.13971*.

Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, et al. 2023b. Llama 2: Open foundation and fine-tuned chat models. *arXiv preprint arXiv:2307.09288*.

Harsh Trivedi, Niranjan Balasubramanian, Tushar Khot, and Ashish Sabharwal. 2022. Musique: Multi-hop questions via single-hop question composition. *Transactions of the Association for Computational Linguistics*, 10:539–554.

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Fei Xia, Ed Chi, Quoc V Le, Denny Zhou, et al. 2022. Chain-of-thought prompting elicits reasoning in large language models. *Advances in neural information processing systems*, 35:24824–24837.

Guangxuan Xiao, Yuandong Tian, Beidi Chen, Song Han, and Mike Lewis. 2023. Efficient streaming language models with attention sinks. In *The Twelfth International Conference on Learning Representations*.

Wenhao Yu, Dan Iter, Shuohang Wang, Yichong Xu, Mingxuan Ju, Soumya Sanyal, Chenguang Zhu, Michael Zeng, and Meng Jiang. 2022. Generate rather than retrieve: Large language models are strong context generators. *arXiv preprint arXiv:2209.10063*.

Elad Ben Zaken, Shauli Ravfogel, and Yoav Goldberg. 2021. Bitfit: Simple parameter-efficient fine-tuning for transformer-based masked language-models. *arXiv preprint arXiv:2106.10199*.

Peitian Zhang, Zheng Liu, Shitao Xiao, Ninglu Shao, Qiwei Ye, and Zhicheng Dou. 2024. Soaring from 4k to 400k: Extending llm's context with activation beacon. *arXiv preprint arXiv:2401.03462*.

Yutao Zhu, Huaying Yuan, Shuting Wang, Jiongnan Liu, Wenhan Liu, Chenlong Deng, Haonan Chen, Zhicheng Dou, and Ji-Rong Wen. 2023. Large language models for information retrieval: A survey. *arXiv preprint arXiv:2308.07107*.

A Global and Local Queries Definition

A significant innovation in our query generation method is the distinction between **global** and **local** queries, similar to GraphRAG's approach but implemented with greater focus on user-driven tasks.

- **Global Queries**: These queries are designed to prompt high-level, interpretive answers that require the user to consider the dataset in its entirety. They address broad trends, themes, and insights that emerge from the data, guiding the user towards macro-level analysis. The global queries necessitate the synthesis of multiple chunks of the dataset, ensuring that the user engages with the dataset as a whole rather than focusing on specific details.
- **Local Queries**: These queries are retrieval-focused and aim to guide the user towards specific pieces of information within the dataset. Each local query is designed to be answerable by referencing a particular section or chunk of the dataset, encouraging a detail-oriented analysis. Local queries require precision in information retrieval and are geared towards users who need clear, concrete answers to more focused queries.

By separating the queries into these two categories, we ensure that the evaluation of RAG systems covers both granular detail retrieval and broader, sense-making tasks, thus providing a more complete assessment of the system's ability to engage with the dataset at multiple levels.

B Metric Explaination

- **Usefulness / Correctness**: This metric assesses the accuracy and reliability of the answer in relation to the posed query. It evaluates whether the answer directly addresses the query and provides useful, relevant information. Answers that demonstrate clear correctness and provide valuable content score higher on this metric.
- **Richness / Diversity**: Here, we measure the variety and depth of the content provided in the answer. An answer that covers multiple perspectives or offers detailed explanations from different angles is considered more diverse and rich. This metric values comprehensiveness and the ability to present a nuanced view of the dataset or topic.
- **Insightfulness / Deep Understanding**: This metric gauges the level of understanding demonstrated by the answer. Insightful responses showcase deep comprehension of the subject matter and may offer thoughtful or original insights that go beyond surface-level retrieval. Answers that synthesize data meaningfully to provide novel or insightful interpretations are rated more highly.
- **User-Friendliness**: This criterion focuses on the clarity, readability, and structure of the response. An answer that is well-organized, concise, and easy for a user to understand will score higher. This metric ensures that even complex answers remain accessible and comprehensible to the target audience, balancing the need for depth with usability.

C Results of Local and Global Subsets

Here we present the experimental results of our model evaluation. Table 5 shows the performance metrics for local queries across three datasets, while Table 6 demonstrates the corresponding results for global queries.

The comparison between global and local query performance reveals a notable pattern in our evaluation results. While both query types demonstrate robust performance, global queries exhibit marginally superior outcomes across most metrics. Specifically, as shown in Table 6, the global query performance achieves higher scores, particularly in the Usefulness and Richness metrics, compared to the local query results presented in Table 5. Nevertheless, the local query performance maintains competitive effectiveness, with only a modest gap in performance metrics, suggesting that our model retains strong capabilities in handling both granular and comprehensive information retrieval tasks.

Table 5: Detailed performance comparison for local queries across different approaches.

Dataset	Metric	Graph RAG C0	Graph RAG C1	Graph RAG C2	Graph RAG C3	Long-Context LLM	Vanilla RAG
Podcast	Usefulness	56.80	53.60	52.00	52.80	65.60	95.20
	Richness	52.80	49.60	47.20	48.00	59.20	96.00
	Insightfulness	59.20	54.40	50.40	51.20	60.00	99.20
	User-Friendliness	80.00	76.00	72.00	71.20	77.60	88.80
News	Usefulness	52.00	52.80	49.60	50.40	95.20	95.20
	Richness	50.40	49.60	45.60	46.40	94.40	99.20
	Insightfulness	56.00	55.20	51.20	51.20	96.00	99.20
	User-Friendliness	78.40	73.60	70.40	68.80	90.40	84.80
LPM	Usefulness	53.33	46.67	46.67	46.67	60.00	73.33
	Richness	46.67	53.33	46.67	46.67	66.67	86.67
	Insightfulness	66.67	60.00	53.33	60.00	73.33	86.67
	User-Friendliness	53.33	46.67	53.33	46.67	66.67	80.00

Table 6: Detailed performance comparison for global queries across different approaches. The results demonstrate how each model performs on broader, more comprehensive queries, with notable differences in performance patterns compared to local queries.

Dataset	Metric	Graph RAG C0	Graph RAG C1	Graph RAG C2	Graph RAG C3	Long-Context LLM	Vanilla RAG
Podcast	Usefulness	54.40	55.20	52.80	52.00	68.00	97.60
	Richness	52.00	49.60	47.20	45.60	61.60	97.60
	Insightfulness	59.20	52.80	53.60	50.40	62.40	99.20
	User-Friendliness	82.40	76.80	73.60	69.60	78.40	90.40
News	Usefulness	52.80	53.60	51.20	52.00	96.80	98.40
	Richness	51.20	48.80	46.40	46.60	96.00	99.20
	Insightfulness	56.80	56.00	52.00	52.00	97.60	100.00
	User-Friendliness	79.20	73.60	71.20	68.80	92.00	85.60
LPM	Usefulness	57.78	55.56	55.56	53.33	64.44	93.33
	Richness	55.56	57.78	57.78	55.56	73.33	100.00
	Insightfulness	68.89	68.89	64.44	68.89	82.22	100.00
	User-Friendliness	57.78	57.78	57.78	55.56	73.33	88.89

Figure 3: Case 1 of examples from the News dataset.

Examples from the News dataset: Tuned LLM defeats GraphRAG

Case 1:
Q: What is the expected release date of Britney Spears' memoir?

Garph RAG:
Britney Spears' highly anticipated memoir, titled *The Woman in Me*, is set to be released on **October 24, 2023**. This memoir is expected to delve into significant moments in her life, including her experiences with fame, motherhood, and the impact of her conservatorship, which has been a focal point of public interest and discussion.

The memoir is described as a candid account of her life, providing insights into her personal struggles and triumphs, and is anticipated to resonate deeply with her fans and the general public alike.

Tuned LLM:
Britney Spears' memoir, titled *The Woman in Me*, is anticipated to be released on **October 24, 2023**. This highly anticipated book is set to delve into her life experiences and challenges, particularly focusing on her conservatorship, a defining aspect of her public persona for many years. The memoir promises to provide readers with an intimate look at her journey through fame, personal struggles, and the complexities of her relationships, including her marriage to Sam Asghari.

Following a tumultuous period marked by the conservatorship, which was finally terminated in November 2021, Britney's memoir comes at a pivotal moment in her life. The end of the conservatorship allowed her to regain control over her life and career, offering her a sense of newfound freedom. In interviews, she has expressed sadness about the loss of her childhood home, describing it as a place where she felt most like herself. This emotional context adds depth to her narrative, as she explores themes of identity, resilience, and the impact of fame on her personal relationships.

The memoir is not merely a recounting of events; it is also a reflection on her personal growth and the lessons she has learned along the way. Britney has been open about her struggles with mental health and the pressures of public life, making her story relatable to many who have faced similar challenges. The memoir is also expected to offer insights into her relationship with Sam Asghari, whom she married in June 2022. Their marriage ended in July 2023, adding another layer of complexity to her story.

Overall, *The Woman in Me* is poised to be a significant literary work, capturing the essence of Britney Spears' life and offering readers a chance to understand her perspective on fame, motherhood, and her quest for autonomy. As the release date approaches, there is growing anticipation about how Britney will share her story and what new revelations may come to light.

Figure 4: Case 2 of examples from the News dataset.

Examples from the News dataset: Tuned LLM defeats GraphRAG

Case 2:
Q: How many counts of fraud and money laundering is Sam Bankman-Fried (SBF) facing in his trial?

Graph RAG:
Sam Bankman-Fried (SBF) is facing a total of seven counts related to his trial, which include charges of fraud and conspiracy. Specifically, these charges encompass wire fraud and money laundering, among other allegations tied to the operations of FTX and the misappropriation of customer funds.

The trial has garnered significant attention due to the serious nature of these allegations, which are considered some of the largest financial frauds in recent history.

Tuned LLM:
In the ongoing trial of Sam Bankman-Fried, commonly referred to as SBF, he is facing serious legal charges that include multiple counts of fraud and money laundering. The context surrounding this trial is quite significant, as it has garnered considerable attention due to the collapse of FTX, a cryptocurrency exchange that was once valued at an impressive $32 billion before its downfall. This dramatic shift in fortune has led to a high-profile legal battle that has captured the interest of both the public and legal experts alike.

The trial itself is set against a backdrop of allegations that SBF engaged in fraudulent activities while managing FTX and its affiliated trading firm, Alameda Research. Specifically, the prosecution claims that SBF orchestrated a scheme that misappropriated customer funds, leading to substantial financial losses for investors and customers. The prosecution's case is expected to unfold over several weeks, with the government presenting evidence that will detail how SBF allegedly misled investors and customers about the safety and management of their assets.

As part of the trial proceedings, SBF has entered a plea of not guilty, which indicates his intent to contest the charges against him. The defense's strategy will likely revolve around challenging the prosecution's narrative and presenting arguments that could potentially exonerate him. However, the prosecution's case is bolstered by the testimony of key witnesses, including former associates and industry experts who can provide insight into the operations of FTX and the broader implications of the alleged fraud.

In summary, Sam Bankman-Fried is facing multiple counts of fraud and money laundering in his trial, which is a critical juncture for both him and the cryptocurrency industry as a whole. The outcome of this trial could have far-reaching consequences, not only for SBF but also for regulatory frameworks governing cryptocurrency exchanges moving forward. As the trial progresses, it will be essential to monitor how the evidence presented impacts the jury's perception and ultimately influences the verdict.