





Evaluating GraphRAG’s Role in Improving Contextual Understanding of News in Newsrooms

Balazs Mosolygo * , Bahareh Fatemi* , Fazle Rabbi , and Andreas L. Opdahl 

University of Bergen, Bergen, Norway

{Balazs.Mosolygo,Bahareh.Fatemi,Fazle.Rabbi,Andreas.Opdahl}@uib.no

Abstract. In a newsroom, journalists are frequently tasked with reporting on complex events, such as conflicts, where understanding the broader context and nuanced details is crucial for accurate and insightful reporting. The challenge lies in processing and synthesizing vast amounts of information from various sources to build a comprehensive picture of the event. This requires not only retrieving specific facts but also understanding the interconnections between different pieces of information. The advent of Large Language Models (LLMs) has brought advancements in text processing, offering the capability to quickly retrieve and generate content from extensive datasets. However, the use of LLMs in newsrooms comes with challenges, particularly concerning their static knowledge and hallucination, where models produce responses that are plausible but incorrect. To address these challenges, Retrieval-Augmented Generation (RAG) has been developed, While RAG is effective for straightforward queries where the relevant information is contained within specific documents, it has limitations when dealing with complex queries that involve synthesizing information from multiple sources or understanding intricate relationships between entities. GraphRAG offers to overcome such limitations by leveraging knowledge graphs, which offer to combine information from multiple sources in a structured manner. In this work, we design a set of experiments to compare GraphRAG’s capabilities to that of existing general LLM and RAG based approaches when it comes to understanding and accurately representing complex issues.

Keywords: Retrieval augmented generation (RAG) · AI-assisted journalism · Large language models · Knowledge graphs · GraphRAG.

1 Introduction

In today’s fast-paced news environment, the accurate and timely reporting is crucial. Contextualizing complex events as stories evolve is often time-consuming

* Equal contribution

for journalists, particularly in sensitive areas like conflict reporting. Accessing the necessary historical, social, and political context can be challenging, leading to fragmented reporting. This problem is exacerbated by the overwhelming volume of information from various sources, which requires journalists to sift through extensive articles and reports—a time-consuming process that detracts from their ability to focus on analysis and verification [8]. While large language models (LLMs) hold promise for streamlining some tasks, their limitations in accuracy and reliability hinder their acceptance in sensitive fields like journalism, where ensuring factual correctness is paramount. LLMs are trained on large amounts of data, and the knowledge gained from this training is usually referred to as enormous knowledge [17], which is embedded in the form of model weights or parameters. However, this extensive general knowledge often lacks the precision needed for domains that demand factual accuracy. Additionally, LLMs can generate hallucinations that pose particular challenges in trust-critical contexts such as newsrooms [10].

Various approaches have been proposed to address this issue using knowledge graphs (KGs). In order to embed factual knowledge into language models, knowledge graphs have been incorporated into the pre-training process, either by integrating KGs into the training objectives [27,23] or by including them directly in the model’s input [22,15]. A challenge with this method is that inductive learning models need retraining to update with new knowledge, leading some studies to suggest delaying the integration of external knowledge until the inference stage. For instance, the Retrieval-Augmented Generation (RAG) framework [14] leverages both parametric and non-parametric memory types: general knowledge is maintained in parametric memory, while factual knowledge is managed through non-parametric memory. This hybrid approach enables the model to dynamically access and integrate up-to-date factual knowledge without the need for constant retraining. Although RAG has been evaluated across various knowledge-intensive NLP tasks and has demonstrated strong performance in different analyses, it has limitations [9]. Specifically, RAG may struggle with complex tasks that involve synthesizing information from multiple sources or comprehending broader concepts within large datasets, as it relies heavily on retrieving discrete pieces of information from individual documents, which can limit its ability to form coherent, high-level insights that span across various contexts.

In a further development of this approach, GraphRAG [6] integrates knowledge graphs with LLMs in a retrieval-augmented generation framework, but with a more sophisticated retrieval strategy that uses community detection to partition the graph into distinct mutually exclusive communities that the LLM can summarize. To perform this partitioning, the authors use the Leiden algorithm [24], which identifies hierarchical community structures within large-scale graphs. GraphRAG’s was shown to be effective in query-focused summarization tasks which, rather than being limited to an explicit retrieval task, requires accurately answering more general queries such as, *What are the main themes in the....* This approach may be particularly well-suited for the needs of newsrooms.

While traditional RAG is effective for answering specific, localized questions, where the answer can be found within a few documents, it falls short in addressing the kind of broader, more global questions that journalists often face. This limitation becomes evident when providing comprehensive background information on complex conflict events, where synthesizing information from diverse sources and understanding overarching themes is essential. GraphRAG’s ability to handle such global queries should make it a more fitting solution in these contexts.

In this work, we contribute to the field of AI-assisted journalism by evaluating the effectiveness of GraphRAG in comparison to traditional retrieval-augmented generation approaches. We specifically focus on its ability to understand and accurately represent complex queries within the news domain. The rest of the paper is organized as follows: Section 2 provides an overview of the background and related works, Section 3 details our evaluation framework, Section 4 presents our findings and, finally, in Section 5 we conclude the study.

2 Background and related works

2.1 Language models

Language models (LMs) have revolutionized natural language processing, particularly with the advent of transformer-based architectures that leverage self-attention mechanisms for improved context understanding. These models can be categorized into two primary types: generative and discriminative. Generative models, such as GPT (Generative Pre-trained Transformer) [3], focus on producing coherent text by modeling the joint probability $p(x, y) = p(y | x) \times p(x)$, making them ideal for tasks like text generation and dialogue systems. In contrast, discriminative models, exemplified by BERT (Bidirectional Encoder Representations from Transformers) [5], aim to estimate only the conditional probability $p(y | x)$, which makes them highly effective for tasks like text classification and sentiment analysis. Additionally, encoder-decoder architectures, like those utilized in T5 [19], combine the strengths of both generative and discriminative approaches.

2.2 Knowledge graphs

Knowledge graphs (KGs) are structured representations of data that model relationships between entities and provide a framework for capturing and organizing knowledge in a way that is easily accessible for computational tasks. They serve as powerful tools in natural language processing by enhancing the understanding of context and relationships inherent in data. There are several types of knowledge graphs, including general knowledge graphs, which contain broad and widely applicable information, for instance Wikidata [25] or YAGO [21], and domain-specific knowledge graphs that focus on specialized fields, such as

CMeKG (chinese medical Knowledge graph)¹. The integration of LLMs with KGs has gained traction to enhance LLMs’ performance by grounding their responses in structured knowledge. This synergistic relationship can help mitigate issues such as hallucination, where models produce incorrect or nonsensical information, while also improving their reasoning abilities and enhancing explainability [17]. In the context of RAG systems, KGs provide a rich source of knowledge that can enhance both retrieval and generation components. In the following section we will see how KGs are used in RAG systems.

2.3 Retrieval augmented generation

Retrieval-Augmented Generation (RAG) systems combine the generative capabilities of large language models with information retrieval techniques. RAG systems enhance standard language models by incorporating relevant documents or data retrieved from external knowledge bases into the model’s context before generating answers. The key components of RAG systems are the **retriever** and the **generator**. The retriever fetches relevant information from a database or knowledge graph utilizing methods such as sparse retrieval, dense retrieval, graph-based retrieval, etc.

- **Sparse retrieval** methods can be used in RAG systems due to their simplicity and efficiency. These methods, such as TF-IDF or BM25 [20], represent documents and queries as sparse vectors, typically based on word frequencies. Unlike dense retrieval, which encodes the semantics of entire documents and queries into dense embeddings, sparse retrieval relies on exact or near-exact token matching between the query and documents.
- **Dense retrieval** represents the documents and queries using dense embedding vectors and build Approximate Nearest Neighbor indexes to speed up the retrieval [28]. In RAG Models, the goal is to retrieve the top-K documents and marginalize over them to compute the probability of generating the entire output sequence y as follows:

$$p(y|x) = \sum_{z \in \text{top-k}(p(\cdot|x))} p_{\eta}(z|x)p_{\theta}(y|x, z) \quad (1)$$

Where x, y and z are respectively the query, the output and the retrieved text documents and where η, θ , are learnable parameters. [12] demonstrated that dense retrieval can outperform the traditional sparse retrieval component in open-domain question answering. [14] introduce RAG models that leverage dense retrieval techniques to find relevant documents. Their retrieval component is based on dense retrieval with BERT, where the retriever uses a bi-encoder architecture to encode both documents and queries as dense vectors.

- **Graph-based retrieval** leverages the structural properties of graphs to enhance information retrieval processes. Unlike traditional retrieval methods that primarily focus on textual similarity, graph-based retrieval extracts

¹ <https://cmekg.pcl.ac.cn/>

information from a structured graph database that contains relational knowledge. This approach enables the retrieval of specific graph elements—such as nodes, triples, paths, or subgraphs—that are relevant to the user’s query, which allows for more structure and granularity within the knowledge augmentation. This capability is crucial for addressing complex queries that require not just localized information but also a broader relational context perspective on the data landscape. Similar to Equation 1, in this setting we have:

$$p(y|x, \mathcal{G}) = \sum_{G \subseteq \mathcal{G}} p_{\eta}(G|x, \mathcal{G}) p_{\theta}(y|x, G) \quad (2)$$

where \mathcal{G} is the underlying text-attributed graph and G s are its various subgraphs containing relevant information. Different approaches have been used to retrieve relevant information in graph-based RAG systems, broadly categorized into three main types: non-parametric, LM-based, and GNN-based retrievers [18].

- Non-parametric retrievers rely on heuristic rules or traditional graph algorithms for efficient retrieval. For example, [4] builds the shortest path relating query entities in the graph and retrieve text chunks mapped to the shortest path entities and their neighbour edges, while [7] applies community detection to identify clusters of related nodes within the graph for retrieval which are aggregated to create a comprehensive response to user queries.
- LM-based retrievers leverage the NLP capabilities of language models (LMs) to interpret queries and retrieve relevant information. KG-GPT [13] uses LLMs to map sub-sentences to relevant relations within a knowledge graph. [26] employs RoBERTa and embedding-based similarity to compute the relevance of relations in a graph, guiding the expansion of paths to form a subgraph for answering queries.
- GNN-based retrievers employ Graph Neural Networks (GNNs) to encode graph structures and score nodes or subgraphs based on their similarity to the query. These models are particularly effective in capturing the intricate structure of graphs. For instance, GNN-RAG [16] leverages GNNs to assess the relevance of entities in the graph, classifying them as either relevant or not, and retrieves the most pertinent nodes based on a predefined threshold.

The generator component in RAG systems can vary depending on the downstream task it is designed for. In the NLP domain, the most common types of generators are generative models with transformer-based architectures such as BART [14] and GPT [7]. Other methods include LSTM models and diffusion models. GANs are also employed in specific scenarios like image generation, text-to-image tasks, and audio synthesis [28].

The retriever and generator can both be trained or fine-tuned to enhance their performance on specific tasks [2,14]. However, it is also possible to employ training-free approaches, where pre-trained models are utilized without further

adaptation [11,7]. In these cases, the effectiveness of retrieval and generation relies heavily on prompt engineering and the inherent capabilities of models like GPT-4 [1], allowing for flexible integration into various applications while minimizing the need for extensive training.

3 Methodology

As already mentioned, the goal of this paper is to evaluate the performance of various RAG approaches in the context of news analysis. This domain presents a unique challenge for information retrieval, as it not only requires the accurate recall of information but also the management of conflicting viewpoints, changes in narrative, and the continuous influx of evolving data related to ongoing or complex events. For this paper, we have constructed two experiments, both aiming to evaluate the models’ ability to provide accurate information to specific questions. The two experiments can be split according to the scope of the information required to answer questions they involve. The first, *local question* set aims to evaluate the different approaches in the context of recalling specific information. This is done by evaluating responses given to multiple choice questions, which are designed to require the recall of specific information. The answers to such questions do not require simultaneous understanding of large portions of data as answers can be found within a single sentence, given that such a sentence is unambiguous and universally accepted as true. The second, *global question* set aims to query for information that would require both understanding broader contexts and serving multiple pieces of key information at once, to be answered properly. Answering such questions accurately is key to practical usefulness in a newsroom, where gaining quick and up-to-date overviews of a complex situations is invaluable.

An overview of our evaluation approach is available in Figure 1. Here the uppermost orange path represents the method evaluation when it comes to high-level overview type questions (i.e., global), while the bottommost blue path shows the evaluation of low-level fact focused questions (i.e., local). The entire set of questions and answers is available on Zenodo², and the source code used during evaluation is also available on GitHub³.

3.1 Data Collection

Avoiding misinformation and bias is important when evaluating a RAG system, because the system will reflect errors in the underlying dataset, resulting in generated responses that contain inaccuracies that are not attributable to the inherent limitations of the system. Errors in the dataset can thus reduce the performance of an otherwise well-adjusted RAG system when it comes to generating responses that reflect reality. The source of this issue lies in the system’s ability

² <https://zenodo.org/records/13840244>

³ https://github.com/MBalazs8796/rag_newsroom_eval

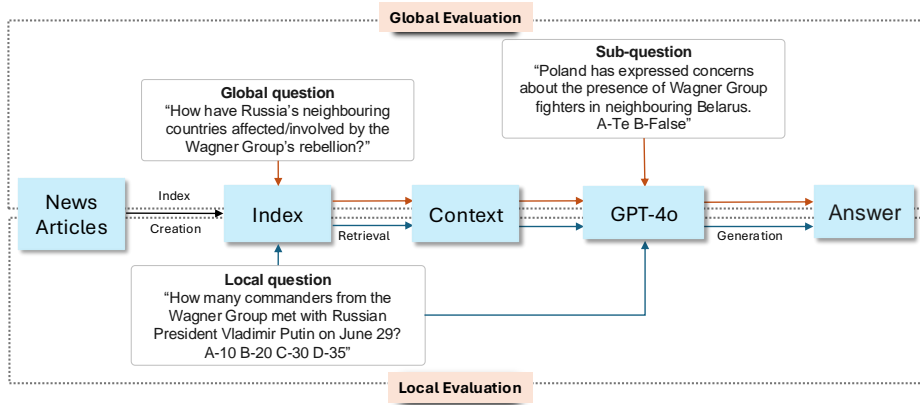


Fig. 1: Workflow diagram of our evaluation framework

Conflict name	Entities used	Keywords used
Sudanese civil war	Sudan (Q1049)	civil war
Wagner Group rebellion	Wagner Group (Q36597284)	rebellion
Russo-Ukrainian war	Russia (Q159), Ukraine (Q212)	war
Israel-Hamas war	Israel (Q801), Hamas (Q38799)	conflict
Yemeni civil war	Yemen (Q805)	war
Nigerien coup	Niger (Q1032)	coup

Table 1: A list of topics used for the generation of the dataset with the entities and keywords used for selecting relevant articles in Aylien

to faithfully reproduce incorrect or biased information, leading to the recall of erroneous or misleading facts. Additionally, less adaptable systems may receive higher scores during evaluation as their inability to adapt to the information provided by the dataset may lead them to ignore false or biased facts, raising their chances of responding correctly to questions related to such topics.

We therefore constructed a test dataset from popular, generally well-trusting news sources to reduce the likelihood of politically biased, or insufficiently fact checked information being present. However, fully eliminating bias is not possible, as even well-trusted and mainstream news sources can only report information from their own perspective. This being the case, we argue that selecting trusted news sources remains the best course of action because (1) it lowers the risk of misinformation and bias and (2) the expected gold-standard answers to our evaluation questions are likely to be informed by the same underlying sources. Thus bias is not eliminated, but is assumed to be generally consistent between an informed questioner and the information source. The list of the used news sources is as follows: *Aljazeera Magazine*, *Associated Press*, *BBC*, *CNN*, *DW*, *France 24*, *Reuters*, *The Guardian*, *The New York Times*, *The Washington Post*.

Question type	# of questions	Example Question text
Local Question (LQ)	662	1. How many years was Alexandra Skochilenko sentenced to prison for her antiwar protest? a) 3 years b) 5 years c) 7 years d) 9 years 2. When did the Yemen conflict, involving the Houthi group and a Saudi-backed government, begin? a) 2012 b) 2014 c) 2016 d) 2018
Global Question (GQ)	33	1. What major events lead up to Russia’s invasion of Ukraine in 2022? 2. What were the major events of the 2023 coup in Niger?
Sub-question	~ 5 per GQ	1.1 Which major historical event lead to the ousting of President Viktor Yanukovich? a) Arab Spring b) Kyrgyz Revolution c) Fishball Revolution d) Revolution of Dignity 2.1 What happened to Mohamed Bazoum during the coup in Niger? a) He was killed by a sniper b) He fled the country c) He was detained d) He was not involved

Table 2: Examples for three question types. Sub-question 1.1 belongs to global question 1, and question 2.1 belongs to global question 2, respectively.

From the above-mentioned sources, we have extracted 1000 recent news articles using the Aylien API ⁴. The data pull was completed using Aylien’s graphical user interface on the 6th of September 2024 with the starting date being the 2nd of August 2022 and the end date being the 6th of September 2024. To select relevant articles a combination of Aylien’s entity and keyword search functions have been used. The combination of keywords and entities can be found in Table 1. In cases where multiple entities are shown, the presence of at least one of them is required. We searched for keywords in both titles and article bodies.

This method of searching does not guarantee that each article is relevant to the given topic. The inclusion of unrelated articles leads only to a slight imbalance in the dataset, as some entity-keyword combinations are more efficient at returning only relevant articles than others. While this does lead to a minor inconsistency in terms of the breadth of information available in each case, it makes the experiments consistent with a real setting, in which thousands of largely irrelevant articles would be available and where slight differences in terms of coverage are to be expected.

As shown in Table 1 we have focused on recent and ongoing conflicts. This selection was made because ongoing conflicts showcase many of the challenges specific to the news domain: (1) Different perspectives and political motives may lead to an entirely different portrayal of recent events, such as civil wars. (2) Additionally, as new reports come in and conflicts develop, places, people, and events may change context or interpretation. (3) Finally, more accurate facts are continuously discovered while they are being reported, leading to a temporal inconsistency, where the number of displaced people for example may change over time. Managing such a complex environment could be challenging for RAG systems as contradicting information is likely to be present.

3.2 Local Questions

The retrieval of information from large language models is rendered impractical in situations where the data in question is of a fine-grained factual nature, as the potential for hallucinations producing fictitious responses, even if small, pose the requirement of cross referencing with more reliable sources. Evaluating RAG techniques in this setting is valuable nonetheless, as the ability to accurately

⁴ <https://aylien.com/>

recall information, especially in such a complicated setting as international political conflicts, is a key factor when it comes to answering more complex questions.

For the purposes of this paper, a question is defined as local if it could be answered based on a single sentence, without requiring additional context. Such questions would, for example, include the names of people involved in conflicts or dates of certain events. In some cases, seemingly local questions may extend their scope, if for example the fact in question is contested and not generally accepted. We do not consider such questions local, as representing all sides of an argument, or the multiple possible sources of disagreement, cannot in general be answered by a single sentence.

Creating local questions is relatively simple, as long as a sufficiently large set of verifiable facts is available. As such, we have automated the process using OpenAI’s GPT-4o model which allowed us to generate 1600 questions that were very likely to be answerable based on the information provided in the articles. To achieve this we have provided GPT-4o with an article and instructions to create a set of multiple-choice questions based on it, with the specific requirement of the question being answerable based on the provided article. Following this process, the questions were manually filtered to correct or eliminate questions that were formatted incorrectly, that were not answerable without the given article’s explicit presence (for example questions like “*According to this article...*”), that had incorrect answers according to the latest information available, or that relied on disputed facts. Once the filtering was completed the 662 questions left were considered adequate. Two examples of local questions can be found in Table 2.

3.3 Global Questions

The creation of broad event descriptions, and the reconciliation of different perspectives on the same issue, presents an important use-case for large language models in the newsroom. Their capabilities to quickly process large amounts of data may aid journalists in gaining an initial understanding of a problem situation. We have therefore designed a set of experiments to evaluate different RAG approaches when it comes to providing answers to complex questions.

Questions become complex when a single source of information, even if well-trusted in terms of accuracy, is not sufficient to provide a complete answer. Answers to such questions are also expected to be more nuanced, as in such cases a larger context, with potentially multiple sources may need to be represented. We identify such questions as global in scope, and refer to them as global questions. Global questions can often be broken down into a set of local questions, so that their answers when combined would produce a sufficient answer. An example of such a question is: “*What is the history behind Norway’s monarchy?*” A satisfactory answer to this a question would require the aggregation of multiple facts and an understanding of their relation to one another.

The assessment of the different RAG systems’ capacity to generate accurate and extensive responses to global questions is a pivotal component of the evaluation of their practical applicability in newsrooms. The construction of a scoring system that can accurately and effectively assess the quality of complex

responses, such as summaries of events, descriptions of contentious issues, and other similar types of responses, represents a significant challenge. For scores to be useful, they must be as objective as possible. However, determining how well a certain issue has been summarized may be influenced by the evaluators’ personal biases and preferences. To address this challenge, we have developed a multi-layered evaluation system to mitigate the influence of personal bias in the assessment of complex responses.

This evaluation includes 33 manually crafted global questions, specifically designed to require multiple sources of information as context to be answered adequately. In contrast to the local questions discussed in section 3.2, they are not inquiries for isolated facts. Global questions were created based solely on the conflicts listed in Table 1, without considering the contents of the collected dataset, meaning that there is no guarantee that the answers to them are present in the dataset.

The evaluation process starts with a global question being posed. To measure the quality of the response, a series of predefined sub-questions are asked based on the provided global answer. Examples of global questions and a few corresponding sub-questions can be found in Table 2. These sub-questions are manually designed multiple-choice questions with ground-truth sub-answers, specific to the particular global question. They are to be answered exclusively based on the global answer and are meant to cover some of the more important details of the respective global question, serving as a test for the completeness of the answer. If the global answer contains information that is factually incorrect about the sub-questions, the sub-questions will be answered by the available evidence leading to incorrect answers.

Sub-question responses are generated by GPT-4o, and the resulting output is subsequently evaluated for correctness according to the manually defined ground-truth sub-answers. It is important to note that GPT-4o is instructed to respond with an invalid answer if it finds that the answer to a sub-question is not present. This is meant to eliminate the likelihood of GPT-4o using external information or randomly guessing. Upon completion of all sub-questions, the score associated with each global question is determined by the overall accuracy of the answers provided in response to the local questions.

4 Results

This section presents the results of our evaluation, using the test set specifically designed for the journalism domain. The test set reflects the real-world queries a journalist might pose, categorized into local and global contexts, to assess the performance of different systems in providing accurate contextual background information.

- **GraphRAG** [7]: We use a community-based retrieval mechanism in conjunction with GPT-4o⁵ as the generator. Due to cost constraints, we utilized the GPT-4o mini model for constructing the knowledge graph.

⁵ <https://openai.com/index/hello-gpt-4o/>

Method	Global Mean Accuracy	Global Std. Deviation	Local Accuracy
GraphRAG	0.52	± 0.29	0.73
GPT-4O	0.35	± 0.25	0.77
Dense Retrieval	0.33	± 0.27	0.80
Sparse Retrieval	0.23	± 0.20	0.81

Table 3: Mean global results with standard deviation and local accuracy

- **Sparse Retrieval:** This system employs BM25 [20] for retrieval and GPT-4o as the generator. BM25 is a traditional sparse retrieval method based on term matching, which serves as a baseline for comparison.
- **Dense Retrieval:** We use all-MiniLM-L6-v2⁶ for dense retrieval. The model encodes both queries and documents into embeddings for similarity search, with GPT-4o as the generator.
- **GPT-4o:** As a baseline, we also evaluate GPT-4o alone, without retrieval augmentation, to assess the model’s generative performance when solely relying on its internal knowledge.

4.1 Global evaluation results

The evaluation of different methods for providing contextual information reveals a clear performance hierarchy according to Figure 2. GraphRAG emerges as the top performer, demonstrating the highest median score and the greatest potential for exceptional results, as evidenced by several outliers, some of which indicate a perfect score (100%). Low outliers are present as well. However there are fewer of them compared to the other methods. Additionally there are questions in the dataset where none of the tested methods reach above a 0% accuracy, indicating that the question may be too challenging or poorly phrased—a scenario that could easily occur in a real-world newsroom setting. However, the large interquartile range indicates a variability in performance across different queries. This suggests that while GraphRAG excels in certain contexts, its effectiveness can fluctuate depending on the specific nature of the questions being asked. This variability may also be attributed to the uncontrolled differences in difficulty among the global questions. GPT-4o and RAG show similar median performances, ranking second and third respectively, with GPT-4o exhibiting slightly more consistent results. Both methods outperform traditional keyword search, which lags significantly behind with the lowest median score. Notably, all methods produce very low scores for certain questions, suggesting that some queries remain challenging across all approaches. Overall, the low and high scores appear to be relatively consistent across methods, reinforcing the notion that the questions posed have varied levels of difficulty.

⁶ <https://huggingface.co/sentence-transformers/all-MiniLM-L6-v2>

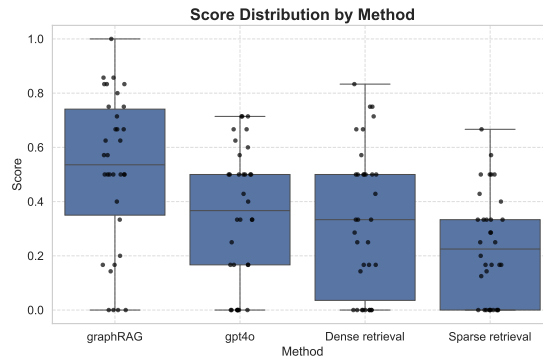


Fig. 2: Box plot representing of global results per method used. Each dot represents the respective method’s performance to a given global question

4.2 Local evaluation results

The results of the local evaluation are shown in Table 3. They show a different picture from the ones discussed during global evaluation. In this case, sparse and dense retrieval perform best, with sparse retrieval scoring highest with an accuracy of 81%, whereas GraphRAG underperforms our GPT-4o baseline by a significant margin. Traditional retrieval methods perform well in this setting because, as described in section 3.2, the local questions are constructed based on the dataset articles themselves. As the questions are likely to contain words or phrases that hint at the articles in which the correct answer is present, a term-matching based approach should perform well because it is likely to retrieve the correct article as input to the generator.

GraphRAG’s surprisingly poor performance may be explained by two factors. Firstly GraphRAG failed to create answers that adhered to the formatting requirements for the local questions it received. This resulted in a necessity of reevaluating its answers by asking GPT-4o what the answer to a given multiple choice answer is. This second step of evaluation has introduced a secondary source of error. Secondly GraphRAG’s underlying representation may be a disadvantage when answers are specific in nature and are known to be directly available in a relatively small dataset. Such specific information is difficult to represent in a graph meaning that the low accuracy when it comes to fine-grained information may be the result of a propagated error within the underlying graph itself.

5 Discussion and Conclusion

In this work, we evaluated the effectiveness of retrieval augmented generation systems using three different retrieval methods — sparse, dense, and graph-based — in the context of newsrooms to assist journalists with gaining an up-to-date

insight into complex ongoing events. Our goal was to assess which system best suits journalists’ needs, both in providing a broad understanding of events and in answering specific, localized questions. To achieve this, we designed an evaluation framework that included a diverse set of questions, specifically targeting local and global assessments, both manually crafted and automatically generated.

According to our experiments, GraphRAG [7] demonstrates superior performance in addressing broader, more general queries that require synthesizing information from multiple sources to create a wider scope of understanding. This advantage can be attributed to its community-based retrieval mechanism, which clusters related entities and extracts summaries that capture overarching themes. However, the local evaluation indicated that dense retrieval techniques also performed competitively. Dense retrieval methods not only yielded strong results but also offered the benefits of being more cost-efficient and faster compared to graph-based methods. This suggests that while GraphRAG is a good choice for generating a big-picture perspectives on complex events, dense retrieval may be more practical for newsroom applications where speed and efficiency in answering precise queries are essential. Overall, our findings underscore the importance of tailoring retrieval methods to the different types of information needs of journalists in dynamic reporting environments.

References

1. Achiam, J., Adler, S., Agarwal, S., Ahmad, L., Akkaya, I., Aleman, F.L., Almeida, D., Altenschmidt, J., Altman, S., Anadkat, S., et al.: Gpt-4 technical report. arXiv preprint arXiv:2303.08774 (2023)
2. Bordes, A., Usunier, N., Chopra, S., Weston, J.: Large-scale simple question answering with memory networks. arXiv preprint arXiv:1506.02075 (2015)
3. Brown, T.B.: Language models are few-shot learners. arXiv preprint arXiv:2005.14165 (2020)
4. Delile, J., Mukherjee, S., Van Pamel, A., Zhukov, L.: Graph-based retriever captures the long tail of biomedical knowledge. arXiv preprint arXiv:2402.12352 (2024)
5. Devlin, J.: Bert: Pre-training of deep bidirectional transformers for language understanding. arXiv preprint arXiv:1810.04805 (2018)
6. Edge, D., Trinh, H., Cheng, N., Bradley, J., Chao, A., Mody, A., Truitt, S., Larson, J.: From local to global: A graph rag approach to query-focused summarization. arXiv preprint arXiv:2404.16130 (2024)
7. Edge, D., Trinh, H., Cheng, N., Bradley, J., Chao, A., Mody, A., Truitt, S., Larson, J.: From local to global: A graph rag approach to query-focused summarization (April 2024), <https://www.microsoft.com/en-us/research/publication/from-local-to-global-a-graph-rag-approach-to-query-focused-summarization/>
8. Gallofré Ocaña, M., Nyre, L., Opdahl, A.L., Tessem, B., Trattner, C., Veres, C.: Towards a big data platform for news angles (2019)
9. Gao, Y., Xiong, Y., Gao, X., Jia, K., Pan, J., Bi, Y., Dai, Y., Sun, J., Wang, H.: Retrieval-augmented generation for large language models: A survey. arXiv preprint arXiv:2312.10997 (2023)
10. Huang, L., Yu, W., Ma, W., Zhong, W., Feng, Z., Wang, H., Chen, Q., Peng, W., Feng, X., Qin, B., et al.: A survey on hallucination in large language models: Principles, taxonomy, challenges, and open questions. arXiv preprint arXiv:2311.05232 (2023)

11. Jin, B., Xie, C., Zhang, J., Roy, K.K., Zhang, Y., Wang, S., Meng, Y., Han, J.: Graph chain-of-thought: Augmenting large language models by reasoning on graphs. arXiv preprint arXiv:2404.07103 (2024)
12. Karpukhin, V., Oğuz, B., Min, S., Lewis, P., Wu, L., Edunov, S., Chen, D., Yih, W.t.: Dense passage retrieval for open-domain question answering. arXiv preprint arXiv:2004.04906 (2020)
13. Kim, J., Kwon, Y., Jo, Y., Choi, E.: Kg-gpt: A general framework for reasoning on knowledge graphs using large language models. arXiv preprint arXiv:2310.11220 (2023)
14. Lewis, P., Perez, E., Piktus, A., Petroni, F., Karpukhin, V., Goyal, N., Küttler, H., Lewis, M., Yih, W.t., Rocktäschel, T., et al.: Retrieval-augmented generation for knowledge-intensive nlp tasks. *Advances in Neural Information Processing Systems* **33**, 9459–9474 (2020)
15. Liu, W., Zhou, P., Zhao, Z., Wang, Z., Ju, Q., Deng, H., Wang, P.: K-bert: Enabling language representation with knowledge graph. In: *Proceedings of the AAAI Conference on Artificial Intelligence*. vol. 34, pp. 2901–2908 (2020)
16. Mavromatis, C., Karypis, G.: Gnn-rag: Graph neural retrieval for large language model reasoning. arXiv preprint arXiv:2405.20139 (2024)
17. Pan, S., Luo, L., Wang, Y., Chen, C., Wang, J., Wu, X.: Unifying large language models and knowledge graphs: A roadmap. *IEEE Transactions on Knowledge and Data Engineering* (2024)
18. Peng, B., Zhu, Y., Liu, Y., Bo, X., Shi, H., Hong, C., Zhang, Y., Tang, S.: Graph retrieval-augmented generation: A survey. arXiv preprint arXiv:2408.08921 (2024)
19. Raffel, C., Shazeer, N., Roberts, A., Lee, K., Narang, S., Matena, M., Zhou, Y., Li, W., Liu, P.J.: Exploring the limits of transfer learning with a unified text-to-text transformer. *Journal of machine learning research* **21**(140), 1–67 (2020)
20. Robertson, S., Zaragoza, H.: The probabilistic relevance framework: Bm25 and beyond. *Found. Trends Inf. Retr.* **3**(4), 333–389 (apr 2009). <https://doi.org/10.1561/15000000019>, <https://doi.org/10.1561/15000000019>
21. Suchanek, F.M., Kasneci, G., Weikum, G.: Yago: a core of semantic knowledge. In: *Proceedings of the 16th international conference on World Wide Web*. pp. 697–706 (2007)
22. Sun, Y., Wang, S., Feng, S., Ding, S., Pang, C., Shang, J., Liu, J., Chen, X., Zhao, Y., Lu, Y., et al.: Ernie 3.0: Large-scale knowledge enhanced pre-training for language understanding and generation. arXiv preprint arXiv:2107.02137 (2021)
23. Tian, H., Gao, C., Xiao, X., Liu, H., He, B., Wu, H., Wang, H., Wu, F.: Skep: Sentiment knowledge enhanced pre-training for sentiment analysis. arXiv preprint arXiv:2005.05635 (2020)
24. Traag, V.A., Waltman, L., Van Eck, N.J.: From louvain to leiden: guaranteeing well-connected communities. *Scientific reports* **9**(1), 1–12 (2019)
25. Vrandečić, D., Krötzsch, M.: Wikidata: a free collaborative knowledgebase. *Communications of the ACM* **57**(10), 78–85 (2014)
26. Zhang, J., Zhang, X., Yu, J., Tang, J., Tang, J., Li, C., Chen, H.: Subgraph retrieval enhanced model for multi-hop knowledge base question answering. arXiv preprint arXiv:2202.13296 (2022)
27. Zhang, Z., Han, X., Liu, Z., Jiang, X., Sun, M., Liu, Q.: Ernie: Enhanced language representation with informative entities. arXiv preprint arXiv:1905.07129 (2019)
28. Zhao, P., Zhang, H., Yu, Q., Wang, Z., Geng, Y., Fu, F., Yang, L., Zhang, W., Cui, B.: Retrieval-augmented generation for AI-generated content: A survey. arXiv preprint arXiv:2402.19473 (2024)