İSTANBUL UNIVERSITY PRESS

Journal of Data Analytics and Artificial Intelligence Applications

Journal of Data Analytics and
Artificial Intelligence Applications,
D3AI 1, 1 (January 2025): 1–13

https://doi.org/10.26650/d3ai.002

Submitted: 21.10.2024
Revision Requested: 16.11.2024
Last Revision Received: 18.12.2024
Accepted: 18.12.2024
Published Online 23.01.2025

Research Article

Open Access

Modern AI Models for Text Analysis: A Comparison of Chatgpt and Rag

Aslan Nurzhanov [1] & Altynbek Sharipbay [2]

[1] L.N.Gumilyov Eurasian National University, Faculty of Information Technology, Department of Information Security, Astana, Kazakhstan

[2] L.N.Gumilyov Eurasian National University, Faculty of Information Technology, Department of Artificial Intelligence Technology, Astana, Kazakhstan

Abstract

This study presents a comparative analysis of two text-processing models: ChatGPT and Retrieval-Augmented Generation (RAG).

ChatGPT, built on the Generative Pre-trained Transformer (GPT) architecture, excels at generating coherent and contextually appropriate texts, making it widely applicable in fields such as education, healthcare, and business. However, it has a significant limitation—it relies solely on pre-trained data, lacking the ability to access real-time information, which can affect the relevance of its responses in dynamic contexts.

In contrast, RAG integrates text generation with external data retrieval, offering a substantial advantage in terms of real-time data relevance. This feature enhances both the accuracy and completeness of the generated responses, especially for tasks that require up-to-date information. The study evaluates both models based on several key performance indicators, including accuracy, completeness, processing time, and scalability.

The conclusion highlights the strengths and weaknesses of each model and suggests potential improvements for their future application across various domains. By offering a deeper understanding of the capabilities and limitations of these technologies, this research contributes to their optimal use and further development.

Keywords Artificial intelligence (AI) · machine learning · natural language processing (NLP) · ChatGPT · RAG

Citation: Aslan Nurzhanov, and Altynbek Sharipbay. 2025. Modern AI Models for Text Analysis: A Comparison of Chatgpt and Rag. *Journal of Data Analytics and Artificial Intelligence Applications* 1, 1 (January 2025), 1-13. https://doi.org/10.26650/d3ai.002

Corresponding author: Aslan Nurzhanov as777an@gmail.com

1. INTRODUCTION

In recent years, AI models have become essential in text processing, including automatic text generation and data analysis. Popular solutions include the generative transformer-based model ChatGPT [1] and hybrid models like RAG [2].

ChatGPT, developed by OpenAI, generates coherent text by training on extensive datasets, allowing it to handle diverse queries. However, it is limited by its reliance on an internal knowledge base, affecting accuracy and relevance [3, 4].

RAG enhances language models by combining retrieval and generation, achieving state-of-the-art performance in many NLP tasks [5, 6].

This study offers a comparative analysis of ChatGPT and RAG, highlighting their advantages, limitations, and potential applications in various tasks.

2. ANALYSIS OF MODERN TEXT PROCESSING METHODS

In this study, over 40 research papers were reviewed, focusing on the application of modern models, including ChatGPT and RAG, in data processing, particularly in text analysis. These works cover a broad range of topics related to the use of these technologies across various fields—from education and healthcare to science and engineering [7]. The articles examine both the advantages of these models, such as high flexibility and improved accuracy through the use of external data, as well as their limitations, including issues with bias, inaccuracies in responses, and computational costs. The comparative analysis of the studies highlighted the key features of applying ChatGPT and RAG and developed a benchmark methodology for their effective use in various text processing scenarios [8, 9].

2.1. ChatGPT

ChatGPT is a large language model developed by OpenAI based on the GPT architecture. It has been trained on vast amounts of textual data to generate coherent and meaningful responses to text-based queries. The model can engage in conversations, answer questions, generate text, and even solve tasks that require complex contextual understanding and logic [1].

2.1.1. Advantages and disadvantages of using RAG for text processing

A comparative review of studies on ChatGPT highlights its diverse applications and varying strengths and weaknesses. In education, research indicates that ChatGPT enhances learning by providing quick answers, assisting with essay writing, and explaining complex concepts, leading to high user satisfaction [6, 10].

In medicine, ChatGPT aids in drafting reports and making recommendations, thus improving clinical efficiency [7, 11].

However, limitations exist, particularly regarding accuracy and reliability in technical fields, as some studies note that the model can generate plausible but incorrect responses [8, 12]. In addition, concerns about political bias and objectivity have been raised [9].

In summary, while ChatGPT shows significant potential, caution is necessary in fields where accuracy and impartiality are critical.

2.1.2. 2.1.2. Benchmark methodology for using ChatGPT

The benchmark methodology for ChatGPT in text generation assesses its ability to produce coherent, accurate, and contextually relevant responses. Key performance metrics include accuracy, recall, precision, and F1-Score, along with processing time to evaluate efficiency.

Testing involves large-scale datasets, such as SQuAD, and various response formats (e.g., structured answers and free-form text) to gauge performance flexibility. Since ChatGPT relies solely on internal knowledge, it generates high-quality responses with minimal latency, supporting scalable response generation for numerous queries [13].

2.2. 2.2. RAG

RAG is a powerful approach for text processing that combines the strengths of generative models with the ability to retrieve information from external data sources [14]. This approach offers several key advantages that make it appealing for various tasks, although it also comes with certain limitations.

2.2.1. 2.2.1. Advantages and disadvantages of using RAG for text processing

RAG excels in retrieving relevant data from external databases, enhancing the accuracy and reliability of responses, particularly in critical fields such as medicine and law [15]. It effectively handles long texts and complex queries, enriching responses with the necessary context, which is beneficial in multi-layered tasks such as legal analysis [16].

Additionally, RAG's adaptability to various domains, leveraging specialised databases, provides versatility compared to traditional models trained on limited datasets [17].

However, RAG's high computational complexity poses a disadvantage, especially with large datasets, leading to slower response times in time-sensitive tasks [18]. Setting up an RAG also requires substantial effort to integrate external databases efficiently, with potential performance issues if not carefully configured [19]. Furthermore, the effectiveness of the RAG relies on the quality of the external data; outdated or incorrect information can undermine its accuracy [20].

In summary, while RAG is a powerful tool for text processing in complex tasks requiring high accuracy, its implementation necessitates considerable computational resources, careful configuration, and access to high-quality data.

2.2.2. 2.2.3. Benchmark methodology for using the RAG

The benchmark methodology for RAG in text processing utilises objective metrics to evaluate performance, including Precision (relevance of retrieved documents), Recall (ability to retrieve all relevant documents), and Accuracy (overall correctness of predictions).

Processing Time measures the speed of document retrieval and response generation, while Scalability assesses the model’s capability to handle increased queries or dataset sizes without performance loss. These metrics provide a quantifiable basis for optimising RAG models in real-world applications, ensuring effective performance in complex text processing tasks [21].

3. RESEARCH METHODOLOGY

The methodology of this research is focused on evaluating the performance of the ChatGPT and RAG models using key metrics. For a systematic analysis, the study is divided into several critical aspects.

3.1. Literature review of the evaluation methods

The evaluation methodologies for the ChatGPT and RAG models incorporate various metrics, including accuracy, recall, and F1-score for ChatGPT, as well as precision and recall for RAG. A detailed overview of the evaluation methods is presented in Table 1 [22-24].

Table 1. Overview of the evaluation methods for the ChatGPT and RAG models

Study title	Model evaluated	Key metrics	Summary of the findings
Evaluating ChatGPT as a Question-Answering System: A Comprehensive Analysis and Comparison with Existing Models [22]	ChatGPT	Accuracy, Recall, and F1-score	Compared ChatGPT with traditional QA systems, testing various interaction modes and evaluation methods.
Evaluation of Retrieval-Augmented Generation: A Survey [23]	RAG	Precision, Recall	Discusses metrics for assessing RAG's retrieval capabilities and generated text quality, including answer relevance and faithfulness.
CRAG—Comprehensive RAG Benchmark [24]	RAG	Context Precision and Answer Relevance	Outlines the key evaluation metrics for the RAG, providing insights into the evaluation methodologies.

3.2. Experimental conditions for testing the ChatGPT and RAG

To test ChatGPT, QA datasets were used, including popular test sets like SQuAD, which evaluate the model's ability to generate accurate answers to complex questions. Various datasets and query types were employed to assess the performance in different generation modes [22, 25, 26].

For the RAG, testing required integrating external databases, using complex queries that necessitated information retrieval. The conditions included working with large datasets and evaluating the relevance of the retrieved data [23, 27].

3.3. The abilities of the ChatGPT and RAG models in processing and classifying extremist texts

To evaluate the effectiveness of ChatGPT and RAG in processing and classifying extremist texts, a specialised dataset was developed. It consisted of examples of extremist content across eight categories: political, religious, racial, national (ethnic), economic, social, youth, and environmental extremism. Additionally, the dataset included materials related to extremism, such as articles, reports, and publications, addressing various aspects of extremist activities, their consequences, and methods of prevention.

Experimental setup for ChatGPT: Input texts were processed directly without external data retrieval. The model's responses were evaluated based on binary classification (extremist or non-extremist) and type classification (correct identification of the specific type of extremism).

Experimental setup for the RAG: The model retrieved contextual data from an external database. Similar evaluation metrics were used, with additional emphasis on the relevance of the retrieved documents.

Key Evaluation Metrics:

- True Positives by type of extremism (TPv): The number of texts correctly classified not only as extremist but also by the correct type of extremism.

- False Positives by type of extremism (FPv): The number of texts with extremist content correctly identified as extremist but misclassified regarding the specific type of extremism.
- False Negatives by type of extremism (FNv): The number of texts containing extremist content of a specific type that the model either failed to classify correctly or failed to recognise as extremist (missed classification).

In the implementation of the RAG model, the external knowledge base was stored in MD (Markdown) files, allowing for a simple and structured format that facilitated efficient processing. These MD files contained texts organised into thematic segments, simplifying the retrieval of relevant information.

Vector representations of the data, stored in the Chroma database, were generated based on the content of these MD files. This setup ensured efficient data management and reduced the system load during query execution. Texts were split into chunks of 300 characters, ensuring consistent representation in the vector database and improving the accuracy of context retrieval.

Additionally, the LangChain library was used to orchestrate the processes of information retrieval and response generation. LangChain enabled seamless integration between the knowledge base, MD files, and vector search operations. During response generation, the RAG model used ChatGPT, leveraging its generative capabilities to analyse retrieved information and produce coherent and contextually relevant outputs. This approach ensured high accuracy and relevance in the tasks related to text classification and processing.

4. EXPERIMENTAL RESULTS

Two models, ChatGPT and RAG, were used in the experiments. Each model was tested in conditions as close as possible to real-world text processing scenarios, including tasks involving short, long, and implicit texts. ChatGPT was tested on various question-answer tasks using datasets like SQuAD, generating answers based on pre-trained data without external search [22]. In contrast, the RAG included a data retrieval component, enabling the system to find information in real databases before generating a response. This ensured the integration of additional sources to improve the accuracy and relevance of the answers [14].

4.1. Performance comparison of ChatGPT and RAG by criteria

Accuracy: ChatGPT performs well in providing contextually correct answers but struggles with complex, factually precise questions due to a lack of external data support [22]. In contrast, RAG shows higher accuracy, particularly in tasks requiring factual retrieval from external sources [25].

Recall: ChatGPT often fails to deliver complete answers, especially with implicit texts [28]. RAG demonstrates higher recall by effectively retrieving and integrating information from multiple sources [25].

Processing Time: ChatGPT is faster for tasks without information retrieval, while RAG takes longer due to its need to search external sources [29].

Scalability: ChatGPT handles numerous queries efficiently [3], whereas RAG faces scalability challenges with large datasets [30].

Relevance: ChatGPT's responses can be general or contextually limited due to reliance on pre-trained data [3]. RAG provides more relevant and accurate responses, especially for complex queries requiring current information [31].

Processing Long Texts: ChatGPT manages long texts but may lose critical information due to its context window limitations [32]. RAG effectively processes long texts by breaking them into chunks and retrieving relevant data as needed.

In the study of the performance of the ChatGPT and RAG models, the results of the comparative analysis are presented in Table 2 [18, 33-37].

Table 2. Comparison of the ChatGPT and RAG performance

Criteria	ChatGPT	RAG	Comments
Accuracy	50.5% (PubMedQA), 15.06% (HotpotQA)	56.42% (PubMedQA), 12.07% (HotpotQA)	RAG shows a slight improvement in accuracy compared to ChatGPT, especially when accessing external data [33, 34]
Recall	1.09% (PubMedQA), 22.63% (HotpotQA)	3.05% (PubMedQA), 25.05% (HotpotQA)	RAG provides better extraction of relevant information, especially on complex question-and-answer tasks [33]
Processing time	~0.3–0.5 sec	~1.5–2 sec	ChatGPT is faster because it does not require access to external data sources [35]
Scalability	High	Average	ChatGPT scales better due to the lower computational cost of extracting information [36]
Relevance of the response	Low for complex queries	Higher thanks to the external data	RAG is more relevant for tasks that require searching for relevant information [37]
Processing long texts	Moderate	High	RAG handles long texts better by extracting information from external sources [18]

These results provide a clear comparison of the performance of ChatGPT and RAG across different criteria, highlighting the strengths and limitations of each model in various text processing tasks.

4.2. Advantages of the RAG in handling long texts

RAG demonstrated clear advantages in working with long texts and texts with implicit content. Thanks to its data retrieval component, the model could locate and use relevant information from external sources, significantly improving the quality of its generated responses. In tasks where the text requires detailed processing and contextual understanding, RAG outperforms ChatGPT, which relies solely on internal model data. This was confirmed in studies where long documents and complex texts requiring in-depth analysis were tested [16].

Thus, RAG proved to be highly effective in complex scenarios that require the integration of external information, while ChatGPT remains more suitable for quicker, less complex tasks involving shorter texts [38].

4.2.1. Evaluation results of the ChatGPT and RAG models' capabilities in processing and classifying extremist texts

As part of the conducted scientific study, 160 extremist texts were tested, evenly distributed across eight types of extremism with 20 texts for each category. The analysis results are presented in Table 3.

Table 3. Comparative table of extremist text classification results by the ChatGPT and RAG models

Type of extremism	ChatGPT			RAG		
	TPv	FPv	FNv	TPv	FPv	FNv
Political	20	1	0	15	1	5
Religious	17	1	3	10	4	10
Racial	16	2	4	13	3	7
National (ethnic)	20	2	0	15	3	5
Economic	15	1	5	12	2	8
Social	16	1	4	11	3	9
Youth	16	0	4	12	3	8
Environmental	15	0	5	11	4	9

The ChatGPT model demonstrated high classification accuracy across most categories of extremism, particularly in tasks where the texts contained explicit content. It achieved maximum TPv values for political and national (ethnic) extremism (20 out of 20), while maintaining low FPv and FNv values. However, in some cases, such as religious and economic extremism, the model made more errors, with FNv reaching up to 5.

The RAG model, leveraging its ability to extract additional data from external sources, showed stable performance in classifying complex and veiled texts. However, its TPv values were generally lower than those of ChatGPT across almost all categories, especially for religious and social extremism, where FNv reached 10 and 9, respectively.

The classification results highlight the differences in the effectiveness of ChatGPT and RAG depending on the type of extremism.

For political extremism, ChatGPT correctly classified all texts (TPv = 20), with no omissions (FNv = 0) and only one FPv error. In contrast, RAG delivered a lower performance, correctly classifying 15 texts (TPv = 15), while missing 5 texts (FNv = 5) and maintaining a similar number of FPv errors.

In religious extremism, ChatGPT performed better, correctly classifying 17 texts (TPv = 17) with one FPv error and three FNv omissions. The RAG model was less accurate, correctly classifying only 10 texts (TPv = 10), making more FPv errors (4), and missing 10 texts (FNv = 10).

For racial extremism, ChatGPT achieved slightly better results, correctly classifying 16 texts (TPv = 16) with 4 FNv omissions. RAG performed worse, correctly classifying 13 texts (TPv = 13) and missing 7 (FNv = 7). Both models had comparable FPv errors (2–3).

In the category of national (ethnic) extremism, ChatGPT again demonstrated maximum accuracy, correctly classifying all texts (TPv = 20) with no omissions (FNv = 0). RAG underperformed, missing 5 texts (FNv = 5) and correctly classifying 15 texts (TPv = 15).

For economic extremism, ChatGPT correctly classified 15 texts (TPv = 15) with 5 FNv omissions, while RAG showed lower accuracy, correctly classifying 12 texts (TPv = 12) and missing 8 (FNv = 8).

The analysis of social extremism also highlighted ChatGPT's superiority, correctly classifying 16 texts (TPv = 16) with 4 FNv omissions. RAG demonstrated lower accuracy, correctly classifying 11 texts (TPv = 11) and missing 9 (FNv = 9).

For youth extremism, ChatGPT showed higher accuracy, correctly classifying 16 texts (TPv = 16) with 4 FNv omissions, while RAG correctly classified 12 texts (TPv = 12) and missed 8 (FNv = 8).

Finally, in the category of environmental extremism, ChatGPT correctly classified 15 texts (TPv = 15) with 5 FNv omissions. In comparison, RAG correctly classified 11 texts (TPv = 11), missed 9 texts (FNv = 9), and had more FPv errors (4 compared to 0 for ChatGPT).

These results demonstrate that ChatGPT achieves higher accuracy in classifying most types of extremism, particularly in cases involving explicit text content. In contrast, the RAG encounters challenges when classifying veiled or complex texts, which is reflected in its higher FPv and FNv error rates.

5. DISCUSSION

The discussion section focuses on a detailed comparison between the strengths and weaknesses of the two models used in this research, ChatGPT and RAG. This analysis is essential for understanding how each model performs in different contexts, highlighting their advantages and limitations in real-world applications. By evaluating these models, we can identify areas where improvements could further enhance their performance. The following subsections provide a breakdown of the key features and shortcomings of both models, followed by potential strategies for enhancing their capabilities in the field of text processing.

5.1. Strengths and weaknesses of each model

ChatGPT's key strengths include fast text generation, scalability, and versatility, enabling real-time query processing without the need for external database access [36]. However, it has significant drawbacks in terms of accuracy and reliability, as it relies on static training data, making it unsuitable for critical fields like medicine or law where current information is essential. Additionally, ChatGPT may produce plausible-sounding but factually incorrect responses [39].

In contrast, RAG improves the accuracy and relevance of responses by integrating with external databases, allowing access to real-time information, which is crucial for complex tasks in medicine and law [14]. Nonetheless, RAG has its own disadvantages, including the high computational costs associated with real-time information retrieval, which can slow down the performance. The complexity of setting up an RAG for integration with various databases also complicates its use across tasks [40].

5.2. Potential improvements in the text processing models

Improving ChatGPT's performance could involve integrating an information retrieval component like RAG to access real-time data, enhancing its accuracy in fields such as medicine and finance. Fine-tuning on specialised datasets for tasks such as medical or legal consultations could also increase relevance and reliability.

For RAG, optimising the data retrieval process with more efficient ranking algorithms could reduce search times and computational costs. In addition, incorporating automatic fact-checking and data verification systems would enhance accuracy and minimise irrelevant information.

In summary, both models possess unique strengths, and targeted improvements could enhance their effectiveness across various tasks.

5.3. Evaluation of the ChatGPT and RAG models' capabilities in processing and classifying extremist texts

The results of the comparative analysis revealed that the ChatGPT and RAG models exhibit varying levels of effectiveness in classifying extremist texts depending on the category and complexity of the content.

ChatGPT demonstrated high classification accuracy, particularly in tasks involving explicit content. The model achieved maximum TPv values for political and national (ethnic) extremism, correctly classifying all texts in these categories. Low FPv and FNv values across most categories confirm the model's ability to handle texts effectively that do not require deep analysis or the extraction of an additional context. However, in categories with veiled content, such as religious and economic extremism, ChatGPT delivered less accurate results, with more omissions (FNv up to 5).

RAG, on the other hand, showcased stable performance when working with veiled and complex texts due to its ability to retrieve additional information from external data sources. However, this capability did not always result in higher accuracy. In categories like religious and social extremism, the model exhibited a significant number of omissions (FNv up to 10) and classification errors related to extremist types (FPv up to 4). This may be attributed to the model's reliance on the quality of external data and the challenges associated with integrating these data into the classification process.

6. CONCLUSION

The comparison of the ChatGPT and RAG models is crucial for advancing artificial intelligence in specialised domains, such as processing extremist texts. ChatGPT serves as a powerful tool for text generation based on pre-trained knowledge, making it effective for quick analysis and tasks that do not require access to external data. However, its limitations become clear in situations where handling current or domain-specific information is essential.

In contrast, RAG, by integrating mechanisms for retrieving data from external sources, shows significant potential for processing texts that require deep analysis and contextual understanding. This capability is particularly important in fields such as medicine, law, or the analysis of veiled extremist content.

The comparison of these models helps identify the key aspects of their applications and determine the directions for further improvements. For instance, integrating the strengths of both models could lead to the development of hybrid systems that combine the accuracy and speed of ChatGPT with the data retrieval capabilities of RAG, offering significant potential for use in critically important areas.

Future research should focus on optimising each model to enhance its effectiveness in specialised applications. For ChatGPT, it would be beneficial to integrate data retrieval mechanisms, enabling the model to access real-time information. This could involve developing hybrid architectures that combine the model's pre-trained knowledge with contextual search capabilities. Additionally, fine-tuning ChatGPT on domain-specific datasets, such as those in medicine, law, or extremist content analysis, would significantly improve its accuracy for specific tasks. Developing algorithms to better analyse veiled texts, leveraging methods of deep context analysis to uncover nuanced meanings, is also essential. To improve the handling of long texts, the model's context window should be expanded, and mechanisms for segmenting text into chunks with subsequent interpretation integration should be implemented.

For the RAG, improving the quality of data retrieval is a priority. This can be achieved by employing more accurate ranking algorithms and implementing automatic verification mechanisms to minimise the impact of outdated or irrelevant information. Reducing the computational complexity of the model is also essential, which can be accomplished by optimising its architecture to accelerate search processes without compromising accuracy and employing dynamic selection of relevant sources. Additionally, RAG should be tailored to specialised tasks through fine-tuning on relevant databases, such as medical or legal sources, while strengthening its contextual retrieval mechanisms to better handle veiled texts.

A combined optimisation of these models could involve developing hybrid systems that merge the speed and accuracy of ChatGPT with RAG's external data retrieval capabilities. Such systems could dynamically adapt to various tasks, automatically switching between pre-trained knowledge generation and external data retrieval based on contextual demands. To enhance effectiveness, it is recommended to test these systems on real-world datasets that reflect practical application conditions and to develop new evaluation metrics that account for requirements in accuracy, speed, and scalability. Ethical and legal considerations are also critical, including measures to prevent the generation of biassed content and to ensure data confidentiality, particularly in sensitive areas such as medicine or counter-extremism efforts. These steps will help tailor the models to specialised applications and ensure their successful deployment across diverse contexts.

The scalability of the AI models and the reduction of computational costs are critical areas for future improvement. This study identifies several strategies that can be implemented to enhance these aspects for both the ChatGPT and RAG models.

For ChatGPT, computational efficiency can be improved through techniques such as parameter reduction or weight quantisation, which reduce resource demands without significantly impacting performance. Similarly, dynamic sampling strategies can be employed in RAG to minimise the number of external data queries, ensuring that only the most relevant information is retrieved.

Caching mechanisms present another promising avenue for optimisation. In the case of RAG, frequently accessed data can be cached, thereby reducing the retrieval times and computational overhead. For ChatGPT, the use of precomputed contexts for common queries could accelerate processing while maintaining accuracy. Additionally, parallel processing on high-performance computing platforms offers potential scalability improvements for both models, allowing them to handle larger datasets and more complex tasks efficiently.

This study makes a substantial contribution to the existing literature by highlighting the strengths and limitations of ChatGPT and RAG in text classification tasks. Specifically, it demonstrates that ChatGPT excels in tasks involving explicit content, while RAG is better suited for handling veiled or contextually complex texts. These findings enrich current knowledge by providing a clearer understanding of the contexts in which each model performs optimally.

Moreover, the results underscore the practical applicability of these models in real-world scenarios, such as healthcare and legal domains, which require high levels of accuracy and reliability. By bridging the gap between academic research and practical deployment, this study provides a valuable foundation for developing hybrid systems that combine the strengths of both ChatGPT and RAG. Such systems could enhance the precision and adaptability of AI models, making them more suitable for diverse and critical applications.

Finally, the findings of this research lay the groundwork for further exploration into computationally efficient hybrid architectures. The evaluation metrics developed and applied in this study, such as TPv, FPv, and

FNv, could serve as benchmarks for analysing the performance of AI models in other specialised tasks. These insights reinforce the importance of integrating computational efficiency, scalability, and practical adaptability into the development of next-generation AI systems.

Peer Review Externally peer-reviewed.

Author Contributions Conception/Design of Study- A.N., A.S.; Data Acquisition – A.N., A.S.; Data Analysis/Interpretation- A.N., A.S.; Drafting Manuscript- A.N., A.S.; Critical Revision of Manuscript- A.N., A.S.; Final Approval and Accountability- A.N., A.S.; Supervision- A.N., A.S.

Conflict of Interest The authors have no conflict of interest to declare

Grant Support The authors declared that this study has received no financial support.

Author Details

Aslan Nurzhanov

[1] L.N.Gumilyov Eurasian National University, Faculty of Information Technology, Department of Information Security, Astana, Kazakhstan

0009-0001-4617-7798

Altynbek Sharipbay

[2] L.N.Gumilyov Eurasian National University, Faculty of Information Technology, Department of Artificial Intelligence Technology, Astana, Kazakhstan

0009-0000-5511-7466

References

[1] Yuntao Wang, Yanghe Pan, Miao Yan, Zhou Su1, and Tom H. Luan. 2023. A Survey on ChatGPT: AI-Generated Contents, Challenges, and Solutions. arXiv:2305.18339. Retrieved from https://arxiv.org/abs/2305.18339

[2] Siyun Zhao, Yuqing Yang, Zilong Wang, Zhiyuan He, Luna K. Qiu, Lili Qiu. 2024. Retrieval Augmented Generation (RAG) and Beyond: A Comprehensive Survey on How to Make your LLMs use External Data More Wisely. arXiv:2409.14924. Retrieved from https://arxiv.org/abs/2409.14924

[3] Walid Harir. 2024. Unlocking the Potential of ChatGPT: A Comprehensive Exploration of its Applications, Advantages, Limitations, and Future Directions in Natural Language Processing. arXiv:2304.02017. Retrieved from https://arxiv.org/abs/2304.02017

[4] Biyang Guo, Xin Zhang, Ziyuan Wang, Minqi Jiang, Jinran Nie, Yuxuan Ding, Jianwei Yue, Yupeng Wu. 2023. How Close is ChatGPT to Human Experts? Comparison Corpus, Evaluation, and Detection. arXiv:2301.07597. Retrieved from https://arxiv.org/abs/2301.07597

[5] Shayekh Bin Islam, Md Asib Rahman, K S M Tozammel Hossain, Enamul Hoque, Shafiq Joty, Md Rizwan Parvez. 2024. Open-RAG: Enhanced Retrieval-Augmented Reasoning with Open-Source Large Language Models. arXiv:2410.01782. Retrieved from https://arxiv.org/abs/2410.01782

[6] Chengcheng Yu, Jinzhe Yan, Na Cai. 2024. ChatGPT in higher education: factors influencing ChatGPT user satisfaction and continued use intention. 2024. Frontiers in Education, 9, Article 2 (May 2024), 11 pages. https://doi.org/10.3389/feduc.2024.1354929

[7] Tirth Dave, Sai Anirudh Athaluri, Satyam Singh. 2023. ChatGPT in medicine: an overview of its applications, advantages, limitations, future prospects, and ethical considerations. Frontiers in Artificial Intelligence, 6, Article 64 (May 2023), 5 pages. https://doi.org/10.3389/frai.2023.1169595

[8] Karen D. Wang, Eric Burkholder, Carl Wieman, Shima Salehi, Nick Haber. 2024. Examining the potential and pitfalls of ChatGPT in science and engineering problem-solving. Frontiers in Education, 8, Article 5 (January 2024), 11 pages. https://doi.org/10.3389/feduc.2023.1330486

[9] Sasuke Fujimoto, Kazuhiro Takemoto. 2023. Revisiting the political biases of ChatGPT. Frontiers in Artificial Intelligence, 6, Article 174 (October 2023), 6 pages. https://doi.org/10.3389/frai.2023.1232003

[10] Anissa M. Bettayeb, Manar Abu Talib, Al Zahraa Sobhe Altayasinah, Fatima Dakalbab. 2024. Exploring the impact of ChatGPT: conversational AI in education. Frontiers in Education, 9, Article 1 (July 2024), 16 pages. https://doi.org/10.3389/feduc.2024.1379796

[11] Maria Grazia Maggio, Gennaro Tartarisco, Davide Cardile, Mirjam Bonanno, Roberta Bruschetta, Loris Pignolo, Giovanni Pioggia, Rocco Salvatore Calabrò, Antonio Cerasa. 2024. Exploring ChatGPT's potential in the clinical stream of neurorehabilitation. Frontiers in Artificial Intelligence, 7, Article 1 (January 2024), 15 pages. https://doi.org/10.3389/frai.2024.1407905

[12] Rex Bringula. 2024. ChatGPT in a programming course: benefits and limitations. Frontiers in Education, 9, Article 73 (February 2024), 6 pages. https://doi.org/10.3389/feduc.2024.1248705

[13] Zeyan Liu, Zijun Yao, Fengjun Li, and Bo Luo. 2024. Check Me If You Can: Detecting ChatGPT-Generated Academic Writing using CheckGPT. arXiv:2306.05524. Retrieved from https://arxiv.org/abs/2306.05524

[14] Yunfan Gao, Yun Xiong, Xinyu Gao, Kangxiang Jia, Jinliu Pan, Yuxi Bi, Yi Dai, Jiawei Sun, Meng Wang, Haofen Wang. 2024. Retrieval-Augmented Generation for Large Language Models: A Survey. arXiv:2312.10997v5. Retrieved from https://arxiv.org/abs/2312.10997v5

[15] Guangzhi Xiong, Qiao Jin, Xiao Wang, Minjia Zhang, Zhiyong Lu, Aidong Zhang. 2024. Improving Retrieval-Augmented Generation in Medicine with Iterative Follow-up Questions. arXiv:2408.00727. Retrieved from https://arxiv.org/abs/2408.00727

[16] Ziyan Jiang, Xueguang Ma, Wenhu Chen. 2024. LongRAG: Enhancing Retrieval-Augmented Generation with Long-context LLMs. arXiv:2406.15319. Retrieved from https://arxiv.org/abs/2406.15319

[17] Xun Xian, Ganghua Wang, Xuan Bi, Jayanth Srinivasa, Ashish Kundu, Charles Fleming, Mingyi Hong, Jie Ding. 2024. On the Vulnerability of Applying Retrieval-Augmented Generation within Knowledge-Intensive Application Domains. arXiv:2409.17275. Retrieved from https://arxiv.org/abs/2409.17275

[18] Zhuowan Li, Cheng Li, Mingyang Zhang, Qiaozhu Mei, Michael Bendersky. 2024. Retrieval Augmented Generation or Long-Context LLMs? A Comprehensive Study and Hybrid Approach. arXiv:2407.16833. Retrieved from https://arxiv.org/abs/2407.16833

[19] Ye Yuan, Chengwu Liu, Jingyang Yuan, Gongbo Sun, Siqi Li, Ming Zhang. 2024. A Hybrid RAG System with Comprehensive Enhancement on Complex Reasoning. arXiv:2408.05141. Retrieved from https://arxiv.org/abs/2408.05141

[20] Fei Wang, Xingchen Wan, Ruoxi Sun, Jiefeng Chen, Sercan Ö. Arık. 2024. Astute RAG: Overcoming Imperfect Retrieval Augmentation and Knowledge Conflicts for Large Language Models. arXiv:2410.07176. Retrieved from https://arxiv.org/abs/2410.07176

[21] Siyun Zhao, Yuqing Yang, Zilong Wang, Zhiyuan He, Luna K. Qiu, Lili Qiu. 2024. Retrieval Augmented Generation (RAG) and Beyond: A Comprehensive Survey on How to Make your LLMs use External Data More Wisely. arXiv:2409.14924. Retrieved from https://arxiv.org/abs/2409.14924

[22] Hossein Bahak, Farzaneh Taheri, Zahra Zojaji, Arefeh Kazemi. 2023. Evaluating ChatGPT as a Question Answering System: A Comprehensive Analysis and Comparison with Existing Models. arXiv:2312.07592. Retrieved from https://arxiv.org/abs/2312.07592

[23] Hao Yu, Aoran Gan, Kai Zhang, Shiwei Tong, Qi Liu, Zhaofeng Liu. 2024. Evaluation of Retrieval-Augmented Generation: A Survey. arXiv:2405.07437. Retrieved from https://arxiv.org/abs/2405.07437

[24] Xiao Yang, Kai Sun, Hao Xin, Yushi Sun, Nikita Bhalla, Xiangsen Chen, Sajal Choudhary, Rongze Daniel Gui, Ziran Will Jiang, Ziyu Jiang, Lingkun Kong, Brian Moran, Jiaqi Wang, Yifan Ethan Xu, An Yan, Chenyu Yang, Eting Yuan, Hanwen Zha, Nan Tang, Lei Chen, Nicolas Scheffer, Yue Liu, Nirav Shah, Rakesh Wanga, Anuj Kumar, Wen-tau Yih, Xin Luna Dong. 2024. CRAG – Comprehensive RAG Benchmark. arXiv:2406.04744. Retrieved from https://arxiv.org/abs/2406.04744

[25] Kunal Sawarkar, Abhilasha Mangal, Shivam Raj Solanki. 2024. Blended RAG: Improving RAG (Retriever-Augmented Generation) Accuracy with Semantic Search and Hybrid Query-Based Retrievers. arXiv:2404.07220v1. Retrieved from https://arxiv.org/abs/2404.07220v1

[26] Yizheng Huang, Jimmy Huang. 2024. Exploring ChatGPT for Next-generation Information Retrieval: Opportunities and Challenges. arXiv:2402.11203. Retrieved from https://arxiv.org/abs/2402.11203

[27] Yihang Zheng, Bo Li, Zhenghao Lin, Yi Luo, Xuanhe Zhou, Chen Lin, Jinsong Su, Guoliang Li, Shifu Li. 2024. Revolutionizing Database Q&A with Large Language Models: Comprehensive Benchmark and Evaluation. arXiv:2409.04475. Retrieved from https://arxiv.org/abs/2409.04475

[28] Yuntao Wang, Yanghe Pan, Miao Yan, Zhou Su, Tom H. Luan. 2023. A Survey on ChatGPT: AI-Generated Contents, Challenges, and Solutions. arXiv:2305.18339. Retrieved from https://arxiv.org/abs/2305.18339

[29] Weiwei Sun, Lingyong Yan, Xinyu Ma, Shuaiqiang Wang, Pengjie Ren, Zhumin Chen, Dawei Yin, Zhaochun Ren. 2023. Is ChatGPT Good at Search? Investigating Large Language Models as Re-Ranking Agents. arXiv:2304.09542. Retrieved from https://arxiv.org/abs/2304.09542

[30] Yu Bai, Yukai Miao, Li Chen, Dan Li, Yanyu Ren, Hongtao Xie, Ce Yang, Xuhui Cai. 2024. Pistis-RAG: A Scalable Cascading Framework Towards Trustworthy Retrieval-Augmented Generation. arXiv:2407.00072. Retrieved from https://arxiv.org/abs/2407.00072

[31] Xiaohua Wang, Zhenghua Wang, Xuan Gao, Feiran Zhang, Yixin Wu, Zhibo Xu, Tianyuan Shi, Zhengyuan Wang, Shizheng Li, Qi Qian, Ruicheng Yin, Changze Lv, Xiaoqing Zheng, Xuanjing Huang. 2024. Searching for Best Practices in Retrieval-Augmented Generation. arXiv:2407.01219. Retrieved from https://arxiv.org/abs/2407.01219

[32] Weizhi Fei, Xueyan Niu, Pingyi Zhou, Lu Hou, Bo Bai, Lei Deng, Wei Han. 2023. Extending Context Window of Large Language Models via Semantic Compression. arXiv:2312.09571. Retrieved from https://arxiv.org/abs/2312.09571

[33] Yuetong Zhao, Hongyu Cao, Xianyu Zhao, Zhijian Ou. 2024. An Empirical Study of Retrieval Augmented Generation with Chain-of-Thought. arXiv:2407.15569. Retrieved from https://arxiv.org/abs/2407.15569

[34] Yizheng Huang, Jimmy Huang. 2024. A Survey on Retrieval-Augmented Text Generation for Large Language Models. arXiv: 2404.10981v1. Retrieved from https://ar5iv.labs.arxiv.org/html/2404.10981v1

[35] Harry Guinness. 2024. What is RAG (retrieval augmented generation)? (August 2024). Retrieved October 15, 2024 from https://zapier.com/blog/retrieval-augmented-generation/.

[36] Frank Joublin, Antonello Ceravola, Joerg Deigmoeller, Michael Gienger, Mathias Franzius, Julian Eggert. 2023. A Glimpse in ChatGPT Capabilities and its impact for AI research. arXiv:2305.06087. Retrieved from https://arxiv.org/abs/2305.06087

[37] Patrick Lewis, Ethan Perez, Aleksandra Piktus, Fabio Petroni, Vladimir Karpukhin, Naman Goyal, Heinrich Küttler, Mike Lewis, Wen-tau Yih, Tim Rocktäschel, Sebastian Riedel, Douwe Kiela. 2021. Retrieval-Augmented Generation for Knowledge-Intensive NLP Tasks. arXiv:2005.11401. Retrieved from https://arxiv.org/abs/2005.11401

[38] Weiwei Sun, Lingyong Yan, Xinyu Ma, Shuaiqiang Wang, Pengjie Ren, Zhumin Chen, Dawei Yin, Zhaochun Ren. 2023. Is ChatGPT Good at Search? Investigating Large Language Models as Re-Ranking Agents. arXiv:2304.09542. Retrieved from https://arxiv.org/abs/2304.09542

[39] Christoph Leiter, Ran Zhang, Yanran Chen, Jonas Belouadi, Daniil Larionov, Vivian Fresen, Steffen Eger. 2023. ChatGPT: A Meta-Analysis after 2.5 Months. arXiv:2302.13795. Retrieved from https://arxiv.org/abs/2302.13795

[40] Ruiyang Qin, Zheyu Yan, Dewen Zeng, Zhenge Jia, Dancheng Liu, Jianbo Liu, Zhi Zheng, Ningyuan Cao, Kai Ni, Jinjun Xiong, Yiyu Shi. 2024. Robust Implementation of Retrieval-Augmented Generation on Edge-based Computing-in-Memory Architectures. arXiv:2405.04700. Retrieved from https://arxiv.org/abs/2405.04700