RETRIEVAL-AUGMENTED GENERATION AND LONG CONTEXT MODELS: A COMPARATIVE ANALYSIS OF ADVANCED GENERATIVE AI APPROACHES

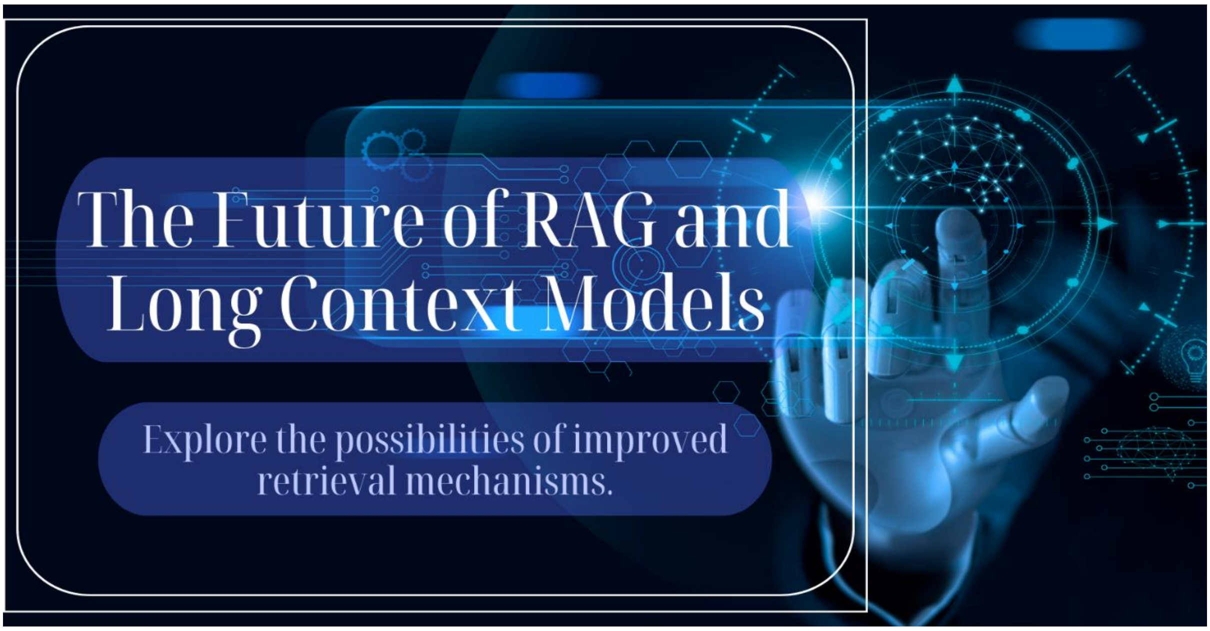
Lalith Kumar Maddali

BrightEdge, USA

---***---

ABSTRACT

In recent years, generative AI models have made significant advancements with the introduction of cutting-edge techniques like Retrieval-Augmented Generation (RAG) and Long Context Models. Through the analysis of extensive data and the generation of tailored results, these models aim to enhance the efficiency and capabilities of AI systems.


The Future of RAG and Long Context Models
Explore the possibilities of improved retrieval mechanisms.

This article provides a comparison between the RAG and Long Context Models, discussing their architectures, advantages, disadvantages, and potential applications. RAG models have the ability to access a wide array of up-to-date information and integrate it into the outputs they generate. This is achieved by combining extensive language models with external knowledge retrieval [1]. The quality and reliability of external sources, however, may impose limitations on them [2]. Alternatively, longer text sequences can be effectively processed and remembered by Long Context Models, such as Transformers with extended attention mechanisms. These models are able to maintain coherence and consistency over lengthy passages [3]. Nevertheless, dealing with lengthy texts can pose challenges for these models, necessitating substantial computational resources [4]. The paper explores the applications of hybrid models and the potential benefits of combining RAG and Long Context approaches [5]. Finally, potential future directions for this field include the creation of innovative hybrid models, improvements to retrieval mechanisms, and advancements in memory and processing. In order to enhance the potency and effectiveness of generative AI systems, it is crucial to understand the pros and cons of RAG and Long Context Models.

Keywords: Generative AI models, Retrieval-Augmented Generation (RAG), Long Context Models, External knowledge retrieval, Hybrid AI approaches

INTRODUCTION

The development of machine-generated text with remarkable coherence and fluency has been made possible by generative AI models, revolutionizing the field of natural language processing. These models have been applied in various fields, including machine translation, dialogue systems, content creation, and summarization [6]. The increasing demand for AI-generated content that is advanced and contextually relevant has prompted researchers to focus on developing sophisticated methods capable of processing large amounts of data and generating outputs that are highly accurate and nuanced. Long Context Models and Retrieval-Augmented Generation (RAG) are two widely recognized methods that have become increasingly popular in recent times. Through the combination of powerful language models and external knowledge retrieval, RAG models aim to enhance the capabilities of generative AI. By extracting relevant data from extensive databases, such as Wikipedia, RAG models can incorporate a wide range of facts and current knowledge into the outputs they produce. This method has shown promise in enhancing the accuracy and educational worth of AI-generated text, particularly for tasks that demand extensive knowledge.

On the other hand, Long Context Models strive to enhance the ability of AI models to understand and remember longer text sequences within their immediate context, without relying on external information. Utilizing Transformers with extended attention mechanisms, such models leverage model architecture advancements to maintain coherence and contextual consistency over longer passages [7]. For tasks involving the understanding and creation of longer, intricate narratives or documents, Long Context Models have demonstrated exceptional performance.

Understanding the relative benefits, limitations, and potential applications of RAG and Long Context Models is increasingly important as generative AI advances. This article aims to provide a comprehensive analysis of these two strategies, examining their use cases, architectures, performances, and potential future directions. Our goal is to offer a glimpse into the immense potential of these models to revolutionize various sectors and fields. This includes personalized recommendations, content creation, scientific research, and more. We will accomplish this by thoroughly examining the current state of research and development in this domain.

RETRIEVAL-AUGMENTED GENERATION (RAG)

Through the fusion of large language models and external knowledge retrieval, Retrieval-Augmented Generation (RAG) presents a fresh approach to enhancing the capabilities and performance of generative AI models. RAG models dynamically retrieve relevant information from external databases, such as Wikipedia, during the text generation process. This approach helps overcome the limitations of traditional language models that solely rely on the data obtained during training. RAG models typically consist of two main components: an external knowledge retrieval system and a powerful language model, such as a Transformer. The retrieval system locates the most relevant information from the external database, while the language model generates text based on the input prompt and the retrieved knowledge. Through the combination of various facts and current knowledge, these two components collaborate to enhance the accuracy and informativeness of RAG models' outputs.

One of the main advantages of RAG models is their ability to access and leverage a large body of external knowledge that can be regularly updated and expanded [8]. It is possible for RAG models to generate text that is accurate, coherent, fluid, and up to date with the latest data. For a range of knowledge-intensive tasks, such as fact-checking, question-answering, and knowledge-based dialogue generation, RAG models have shown promising performance [1]. Yet, there are certain restrictions and difficulties with RAG models as well. The precision and applicability of the knowledge that is retrieved have a significant impact on the quality and dependability of the generated text [2]. It is possible that the model's results could be biased, out-of-date, or inaccurate if the external database contains this type of data. In addition, retrieval processes can be computationally expensive, especially when dealing with large-scale databases, leading to higher resource demands and longer generation times [9].

Several suggestions have been proposed by scholars to enhance and expand the RAG framework in order to address these difficulties. For instance, certain research has focused on developing improved retrieval mechanisms, such as hierarchical indexing structures or dense vector representations [8]. Previous studies have explored the incorporation of feedback mechanisms and the integration of external knowledge sources to enhance the retrieval of information based on the generated outputs [10]. Through further development, the field of Retrieval-Augmented Generation holds immense potential to revolutionize the way AI models engage with and leverage external knowledge. RAG models excel in generating highly precise, insightful, and contextually fitting text by harnessing the capabilities of expansive language models and comprehensive knowledge bases. Such applications have a wide range of uses in fields such as knowledge-driven decision making, personalized recommendations, and content creation.

DEFINITION AND KEY CONCEPTS

Combining large language models and external knowledge retrieval in a unique approach known as Retrieval-Augmented Generation (RAG) enhances the capabilities and performance of generative AI systems [1]. The main concept behind RAG is to enhance the language model's internal knowledge during text generation by incorporating relevant information obtained from external sources, such as databases or the internet [2]. Enhancing the overall quality and relevance of the generated text is achieved by enabling RAG models to incorporate current and factually accurate information into their outputs [8].

ARCHITECTURE AND FUNCTIONING OF RAG MODELS

The architecture of RAG models typically consists of a substantial language model and an external knowledge retrieval system [1]. The process of text generation involves utilizing a language model, often a Transformer-based model, to leverage retrieved knowledge and the input prompt. However, depending on the input query, the retrieval system's task is to find the most relevant data from external sources [2]. The retrieval system searches the external knowledge base for relevant passages or documents while generating the text. These inputs are then fed into the language model along with the given prompt. To generate the output text, the language model considers both the prompt and the retrieved knowledge [8].

ROLE OF EXTERNAL KNOWLEDGE RETRIEVAL IN RAG

The external knowledge retrieval component plays a crucial role in the operation of RAG models. RAG models have the ability to access a wide range of diverse and up-to-date knowledge that might not be available in the training data of the language model. This is achieved by retrieving relevant information from external sources [1]. Due to the ability of RAG models to utilize retrieved knowledge, they can generate text that is more informative and factually accurate [2]. Thanks to the regular updates of external knowledge bases, RAG models can easily adapt to new data and stay current with advancements in various fields [8].

ADVANTAGES OF RAG MODELS

Access to a wide range of facts and up-to-date information

The capacity of RAG models to access and apply a large amount of outside knowledge is one of its primary benefits [1]. RAG models are able to incorporate a broad range of facts and current information into their generated outputs by retrieving data from large-scale databases or the internet [2]. This approach proves to be highly advantageous for tasks like fact-checking, question-answering, and knowledge-based dialogue generation that demand advanced reasoning or factual precision [8].

Ability to incorporate external knowledge into generated outputs

An important benefit of RAG models is their ability to effortlessly integrate external knowledge into the generated text [1]. The outputs of RAG models are widely acknowledged for their factual consistency, informativeness, and relevance. Both the input prompt and the retrieved knowledge are taken into account during the generation process [2]. Additionally, the fluidity and coherence of RAG models allow for the creation of current text that includes relevant information [8].

Limitations and challenges of RAG models

Dependence on the quality and reliability of external sources

One of the main drawbacks of RAG models is their dependence on the caliber and reliability of external knowledge sources [1]. The generated outputs of RAG models may reflect potential issues related to bias, incorrectness, or outdatedness if the external databases or websites contain such information [2]. Accuracy and clarity are crucial in certain applications, such as news reporting or medical advice, as they directly impact the potential consequences [8].

Potential for inconsistencies or inaccuracies in retrieved information

Mistakes or discrepancies in the information retrieved can present a significant challenge for RAG models [1]. Contradicting or conflicting information in external knowledge sources can be a challenge for RAG models when determining the most reliable and accurate facts to include in their generated outputs [2]. Generating inconsistent or factually inaccurate text can have a detrimental effect on the overall reliability and quality of the model's outputs [8].

Feature	Retrieval-Augmented Generation (RAG)	Long Context Models
External Knowledge	Yes	No
Context Length	Moderate	Long
Retrieval Mechanism	Dynamic, real-time	Not applicable
Attention Mechanism	Standard attention	Extended attention
Computational Efficiency	Moderate	High

Table 1: Comparison of Key Features in RAG and Long Context Models

Long Context Models

Long Context Models, a class of generative AI models, excel in processing and producing text over extended sequences while maintaining coherence and contextual consistency [3]. The models in question strive to overcome the limitations of traditional language models by enhancing their ability to retain context over extended passages and capture dependencies across a broad spectrum. Long Context Models have demonstrated remarkable performance in tasks that require understanding and generating longer, more intricate narratives or documents [11].

Definition and key concepts

Long Context Models stand out from traditional language models due to their ability to consider a significantly longer context while processing and generating text [11]. The core concept of long context models involves expanding the attention mechanism and memory capacity of the model, enabling it to grasp and utilize information from a broader span of the input sequence [4]. This feature allows the model to focus on and include relevant information from different parts of the text, ensuring that longer passages are coherent and consistent.

ARCHITECTURE AND FUNCTIONING OF LONG CONTEXT MODELS

Transformers with extended attention mechanisms

The architecture of Long Context Models takes inspiration from the Transformer model, incorporating additional attention mechanisms to further enhance its capabilities [3]. An interesting example of this phenomenon can be seen in the Longformer model, which effectively handles longer sequences by incorporating both local and global attention [3]. For optimal computational efficiency, the model is designed to capture both local and global context. An effective approach is to assign each token in the Longformer attention to its local window and a limited number of globally selected tokens [3].

Advancements in model architecture for handling extensive context

The Big Bird model showcases the progress in model architecture to effectively handle extensive context. This approach effectively captures long-range dependencies by incorporating a combination of local, global, and random attention patterns [12]. Additionally, the Sparse Transformer model excels at managing longer contexts thanks to its implementation of a sparse attention mechanism, which offers improved scalability with sequence length [12]. Long Context Models excel at handling and generating longer, more coherent text with greater efficiency, thanks to specific architectural innovations [11].

ADVANTAGES OF LONG CONTEXT MODELS

Ability to process and remember longer sequences of text

Long Context Models possess a notable advantage over traditional language models as they excel in comprehending and retaining lengthier text sequences [3]. Long Context Models possess the capability to incorporate and extract information from a significantly broader range of the input sequence by expanding their attention mechanisms and memory capacity [4]. It becomes possible for models to understand and generate longer, more complex stories, reports, or articles [11].

Maintenance of coherence and contextual consistency over longer passages

Ensuring coherence and contextual consistency over longer passages is a significant advantage of utilizing Long Context Models [3]. When dealing with lengthy sequences, Long Context Models have the ability to generate outputs that are more coherent and consistent by taking into account relevant information from distant sections of the text [4]. Assignments that necessitate a thorough grasp of the broader context and the skill to maintain a cohesive narrative flow are especially crucial [11].

LIMITATIONS AND CHALLENGES OF LONG CONTEXT MODELS

Limitations of memory for very long texts

Despite the advancements in model architecture, Long Context Models still encounter memory capacity constraints when handling extremely long texts [3]. Dealing with lengthy documents or books can pose a challenge because of the growing computational cost and memory demands associated with longer sequences [4]. Finding the right balance between context length and computational efficiency is still a challenging task, despite the advancements made with techniques like sparse attention and memory-efficient architectures [11].

Increased computational resource requirements

Long context models have higher computational resource requirements compared to traditional language models, which poses an additional challenge [3]. Using these models in resource-constrained environments may pose limitations due to their increased memory and computational demands when handling and producing longer sequences [4]. The focus of current research is on developing more effective architectures and optimization strategies to reduce computational load, while still maintaining the benefits of long-range context modeling [11].

Comparative Analysis

Strengths and weaknesses of RAG models

One benefit of RAG models is their capacity to incorporate a substantial amount of external knowledge into the outputs they generate [1]. RAG models have the capability to generate text that is both accurate and informative, especially in tasks that require a lot of knowledge, like fact-checking and answering questions [2]. However, flaws can be identified in RAG models, especially in their reliance on the quality and reliability of external sources of knowledge [8]. Text that may be misleading or erroneous can be produced if the information retrieved is outdated, biased, or incorrect.

While RAG models are often discussed in the context of incorporating external knowledge, they can also be designed to handle longer sequences. For example, the Retrieval-Augmented Generation with Chunking (RAG-C) model [20] introduces a chunking mechanism that allows the model to process and generate longer text by breaking it down into smaller chunks. This approach enables RAG-C to maintain coherence and contextual consistency across longer passages while still leveraging the benefits of retrieval-augmented generation.

Strengths and weaknesses of long context models

Long Context Models excel at processing and producing longer, more coherent text while maintaining contextual consistency [3]. By capturing long-range dependencies and incorporating data from remote segments of the input sequence, these models can generate outputs that are more coherent and contextually relevant [4]. Long Context Models, however, come with certain drawbacks, such as the requirement for additional computational resources and the balance between computational efficiency and context length [11]. In addition, even with these models, lengthy texts such as entire books or extensive documents may still pose challenges [12].

Use cases and applications

Application	RAG Models	Long Context Models
Question Answering	Retrieving relevant information to provide accurate and informative answers	Maintaining contextual understanding across multiple turns of conversation
Fact-Checking	Verifying claims against external knowledge bases to ensure factual accuracy	Assessing the consistency and coherence of long-form articles or reports
Content Generation	Incorporating up-to-date information and knowledge into generated content	Generating longer, more coherent narratives or stories
Document Summarization	Identifying and extracting key information from multiple sources	Preserving the overall context and structure of lengthy documents

Table 2: Example Applications of RAG and Long Context Models

Scenarios where RAG models are preferred

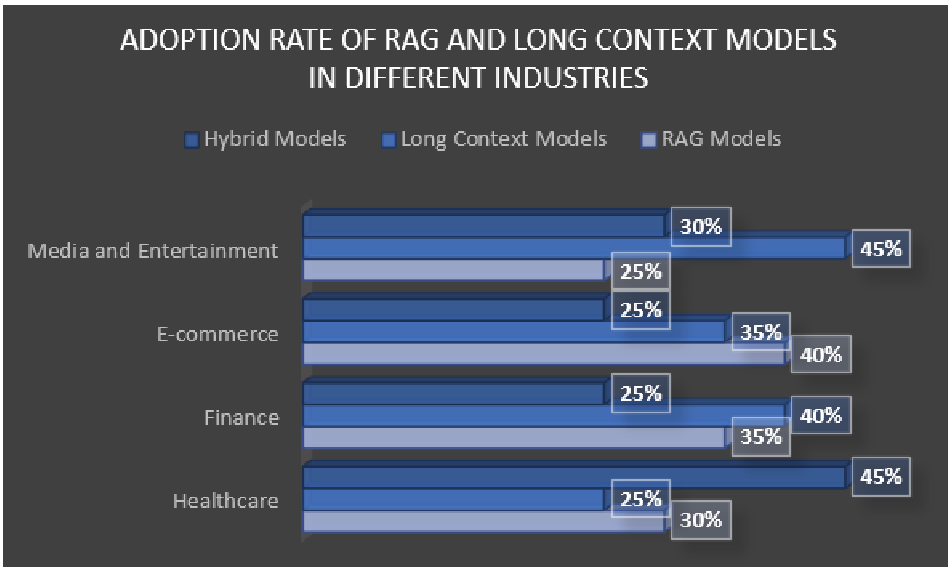
Tasks that require access to a broad range of external knowledge and current information are particularly well-suited for RAG models [1]. Tasks such as fact-checking, generating dialogue based on knowledge, and answering questions are important. Retrieving and incorporating pertinent facts is crucial [2]. Given their ability to adapt to new information and update external knowledge sources, RAG models have proven valuable in areas such as news summarization and scientific literature analysis, where the knowledge landscape is constantly evolving [8].

Scenarios where long context models are preferred

When faced with the need to understand and create longer, more intricate narratives or documents, it is advisable to consider using long context models [3]. These jobs involve the creation of stories, writing reports, and developing long-form content, all of which demand the maintenance of coherence and contextual consistency throughout extensive sections [4]. Long context models have the ability to capture and utilize information from the entire input sequence, making them valuable in domains that require analyzing and summarizing lengthy documents such as contracts or medical records [11].

Performance comparison in different tasks and domains

The performance of RAG and Long Context models can be influenced by the specific task and domain [1]. RAG models have demonstrated impressive performance in knowledge-intensive tasks like fact-checking and question-answering. They often surpass traditional language models by leveraging external knowledge [2]. However, Long Context Models have shown superior performance in tasks such as story generation or document summarization [4], as they are able to maintain contextual consistency over lengthy sequences. The decision between RAG models and Long Context Models ultimately depends on the specific task and the balance between computational efficiency, context length, and knowledge incorporation [4].



Graph 1: Industry Adoption of RAG, Long Context, and Hybrid Models

Integration and Hybrid Approaches

Potential for combining RAG and long context techniques

The fusion of Retrieval-Augmented Generation (RAG) and Long Context approaches has garnered significant interest in the field of generative AI [13]. Hybrid models have the potential to excel in tasks that demand extensive external knowledge and the ability to maintain long-range context [14]. Furthermore, the combination of RAG and Long Context techniques enables models to retrieve relevant information from external sources [15], while also processing and generating longer, more coherent text.

Examples of hybrid models and their applications

Recent literature has seen the emergence of several hybrid models that combine RAG and Long Context techniques. As an example, the Fusion-in-Decoder (FiD) model [14] has the capability to incorporate relevant passages from an external knowledge base during the processing of longer sequences. This is achieved by integrating a retrieval mechanism similar to RAG into the architecture of a Long Context Transformer. FiD has demonstrated strong performance in knowledge-intensive language tasks and open-domain question answering, as shown in [14].

Another example is the KG-FiD model [15], which incorporates knowledge graphs as an additional external information source to the FiD architecture. Tasks such as fact-based dialogue generation and commonsense reasoning benefit from the integration of structured knowledge and long-range context. KG-FiD has shown promising results in these areas [15].

Benefits and challenges of integrating RAG and long context approaches

Combining the Long Context and RAG approaches offers numerous benefits. To begin with, hybrid models have the ability to process and generate longer, more coherent text. They also take advantage of the vast amount of external knowledge provided by retrieval mechanisms [13]. When used in combination, the utilization of these two skills can greatly improve performance on tasks that require a deep understanding of the context and intensive reasoning abilities [14].

Additionally, the drawbacks of each strategy can be minimized by combining RAG and Long Context approaches. For example, the advanced context modeling abilities of Long Context Models can aid RAG in generating more coherent and contextually consistent outputs. Additionally, its retrieval mechanism allows for accessing relevant information that may not be present in the training data [15].

However, there are challenges in merging RAG and Long Context approaches. Hybrid models face a challenge in dealing with the increased computational complexity and resource requirements due to managing both the retrieval procedure and extended context modeling [13]. When designing hybrid models, it is crucial to strike a balance between retrieval quality, context length, and computational efficiency [14].

It is important to consider the potential for noise or irrelevant information to be introduced during the retrieval process, as this could negatively impact the quality of the output [15]. In order to ensure that the information retrieved is relevant and helpful for the task at hand, it is crucial to develop effective retrieval strategies and filtering mechanisms [13].

FUTURE DIRECTIONS

Potential improvements in retrieval mechanisms for RAG models

The retrieval mechanisms utilized by RAG models play a crucial role in ensuring their effectiveness and functionality. Developing more advanced retrieval strategies that can accurately identify and prioritize relevant data from external knowledge sources is a potential area for further exploration [16]. In order to enhance the quality and relevance of the retrieved passages, it may be necessary to utilize advanced similarity metrics such as contextualized embeddings or semantic similarity.

Incorporating feedback mechanisms that can adjust the retrieval procedure based on the generated outputs is another intriguing avenue. Possibly, RAG models can enhance the coherence and factual consistency of their outputs by iteratively improving the retrieval queries and reordering the retrieved passages based on their significance to the generated text.

Another area of improvement for RAG models is the development of techniques that enable them to handle longer sequences more effectively. This can involve the use of chunking mechanisms, as demonstrated by the RAG-C model [20], or the integration of attention mechanisms that can selectively attend to relevant parts of the retrieved information and the generated text. By enhancing the ability of RAG models to process and generate longer sequences, their applicability can be expanded to tasks that require longer-form text generation

Advancements in memory and processing for long context models

Advancements in memory and processing architectures enable Long Context Models to effectively handle even longer sequences. One promising strategy involves the development of sparse attention mechanisms that can efficiently focus on relevant parts of the input sequence. Long Context Models excel at handling longer sequences with reduced memory and computational demands through the use of techniques such as local attention, global attention, and hierarchical attention.

The study of memory-augmented architectures, which excel at storing and retrieving relevant data from the input sequence, is an area of research [17]. Long Context Models have the potential to enhance context modeling by incorporating external memory components such as memory networks or differentiable neural computers [17].

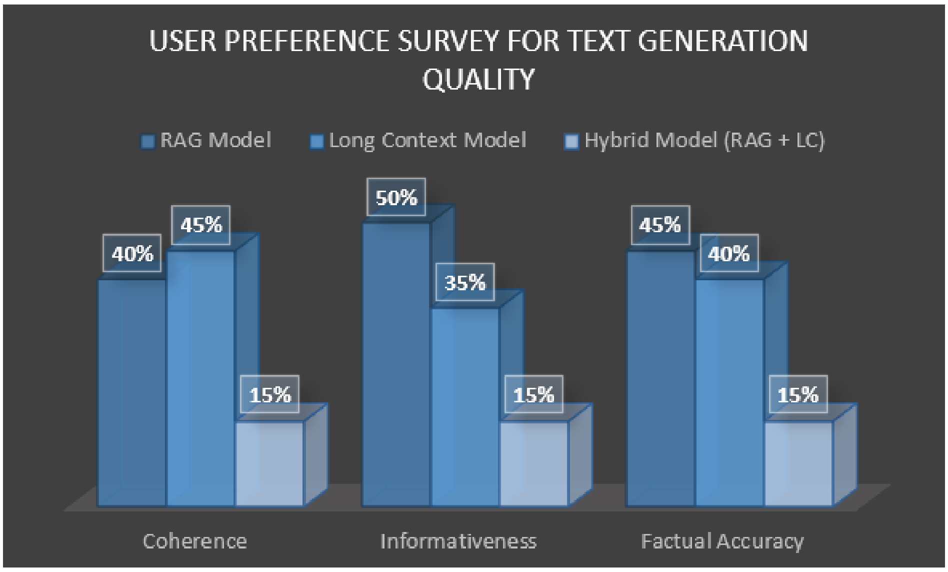
Emerging hybrid models and their potential impact

A new class of hybrid models has emerged, combining the advantages of both approaches through the integration of RAG and Long Context techniques. The performance of these new hybrid models, such as KG-FiD and Fusion-in-Decoder (FiD), has already been promising in various context-dependent and knowledge-intensive tasks.

Hybrid models have the potential to significantly influence various applications such as knowledge management, dialogue systems, question-answering, and content creation, as they continue to evolve and expand. Hybrid models possess the ability to enhance AI systems by incorporating vast external knowledge through retrieval mechanisms and maintaining long-range context.

RESEARCH OPPORTUNITIES AND CHALLENGES IN THE FIELD

Generative artificial intelligence (AI) presents numerous research opportunities and challenges, particularly in the context of RAG and Long Context Models. Developing more effective and efficient retrieval mechanisms to handle the growing volume of external knowledge sources is a crucial area of research [16]. One possible approach is to explore advanced indexing strategies, techniques for reducing dimensionality, and algorithms for approximate nearest neighbor search.



Graph 2: User Preferences for Text Generation Quality: RAG, Long Context, and Hybrid Models

Exploring techniques to improve the interpretability and controllability of generative AI models is an important area of research [18]. Ensuring the safety, fairness, and dependability of RAG and Long Context Model outputs can be achieved by developing techniques for understanding and adjusting their internal representations and generation process [18].

In addition, there are challenges when it comes to assessing and comparing generative artificial intelligence models, particularly in evaluating their effectiveness in context-heavy and knowledge-intensive tasks [19]. In order for the field to progress, it is crucial to develop comprehensive and standardized evaluation frameworks that can gauge the quality, coherence, and factual consistency of generated outputs [19].

CONCLUSION

This article delves into the fascinating realms of Long Context Models and Retrieval-Augmented Generation (RAG), two cutting-edge generative AI techniques. The initial portion of the discussion focused on the fundamental concepts and structures of RAG models. These models integrate extensive language models with external knowledge retrieval to generate outputs that are both factually accurate and informative. Next, the field of Long Context Models was explored, focusing on the processing and production of longer, more coherent text. This is achieved through the utilization of sophisticated attention mechanisms and memory architectures to maintain contextual consistency. The advantages and disadvantages of each approach were compared and contrasted, with a focus on their unique qualities and limitations. In addition, a variety of use cases and applications have been explored to demonstrate the potential of RAG and Long Context Models. Tasks varied from context-dependent activities such as story generation and document summarization to knowledge-intensive tasks like fact-checking and question answering. Additionally, the potential of combining Long Context and RAG approaches was discussed, highlighting the benefits of hybrid models that incorporate the strengths of each approach. The performance of these new hybrid models, such as KG-FiD and Fusion-in-Decoder (FiD), has already shown promise in tackling difficult tasks that require context and knowledge. It is important for those interested in generative AI to understand and evaluate the differences between RAG and Long Context Models. One can choose and apply these approaches to different tasks and domains with knowledge of their underlying concepts, architectures, and capabilities. Additionally, potential areas for innovation and development can be identified by understanding the benefits and drawbacks of each strategy. The comparative analysis in this article serves as a valuable tool for gaining insight into the rapidly evolving field of generative AI. This provides a solid foundation for understanding the significant challenges and potential opportunities, along with the current state of the field. The development of RAG and Long Context Models holds great potential for the future of generative AI. Anticipating the development and growth of AI systems, it is expected that there will be more powerful, knowledgeable, and context-aware systems capable of generating high-quality, educational, and cohesive text across various domains. Integration of RAG and Long Context techniques enables the creation of hybrid models that can effectively maintain long-range context and leverage extensive external knowledge. Hybrid models possess the potential to revolutionize various sectors such as knowledge management, scientific research, content creation, and personalized recommendations.

However, the field of generative AI in the future presents numerous research opportunities and notable challenges. Several major areas require further research and development, including the improvement of retrieval mechanisms for greater effectiveness and efficiency, the enhancement of interpretability and controllability of generative models, and the establishment of comprehensive evaluation frameworks. Encouraging collaboration and knowledge sharing among researchers, industry practitioners, and the larger AI community is crucial as we embark on this exciting journey. The full potential of generative AI can be realized, leading to the development of more knowledgeable, intelligent, and context-aware systems. These advancements can greatly benefit society by building upon the foundations established by RAG and Long Context Models.

REFERENCES

[1] Lewis, P., Perez, E., Piktus, A., Petroni, F., Karpukhin, V., Goyal, N., ... & Kiela, D. (2020). Retrieval-augmented generation for knowledge-intensive NLP tasks. arXiv preprint arXiv:2005.11401. https://arxiv.org/abs/2005.11401

[2] Petroni, F., Piktus, A., Fan, A., Lewis, P., Yazdani, M., De Cao, N., ... & Kiela, D. (2021). KILT: a benchmark for knowledge intensive language tasks. arXiv preprint arXiv:2009.02252. https://arxiv.org/abs/2009.02252

[3] Beltagy, I., Peters, M. E., & Cohan, A. (2020). Longformer: The long-document transformer. arXiv preprint arXiv:2004.05150. https://arxiv.org/abs/2004.05150

[4] Zaheer, M., Guruganesh, G., Dubey, K. A., Ainslie, J., Alberti, C., Ontanon, S., ... & Ahmed, A. (2020). Big bird: Transformers for longer sequences. Advances in Neural Information Processing Systems, 33, 17283-17297. https://proceedings.neurips.cc/paper/2020/hash/c8512d142a2d849725f31a9a7a361ab9-Abstract.html

[5] Borgeaud, S., Mensch, A., Hoffmann, J., Cai, T., Rutherford, E., Millican, K., ... & Irving, G. (2021). Improving language models by retrieving from trillions of tokens. arXiv preprint arXiv:2112.04426. https://arxiv.org/abs/2112.04426

[6] Radford, A., Wu, J., Child, R., Luan, D., Amodei, D., & Sutskever, I. (2019). Language models are unsupervised multitask learners. OpenAI blog, 1(8), 9. https://cdn.openai.com/better-language-models/language_models_are_unsupervised_multitask_learners.pdf

[7] Zaheer, M., Guruganesh, G., Dubey, K. A., Ainslie, J., Alberti, C., Ontanon, S., ... & Ahmed, A. (2020). Big bird: Transformers for longer sequences. Advances in Neural Information Processing Systems, 33, 17283-17297. https://proceedings.neurips.cc/paper/2020/hash/c8512d142a2d849725f31a9a7a361ab9-Abstract.html

[8] Karpukhin, V., Oguz, B., Min, S., Lewis, P., Wu, L., Edunov, S., ... & Yih, W. T. (2020). Dense passage retrieval for open-domain question answering. arXiv preprint arXiv:2004.04906. https://arxiv.org/abs/2004.04906

[9] Metzler, D., Tay, Y., Bahri, D., & Najork, M. (2021). Rethinking search: Making domain experts out of dilettantes. arXiv preprint arXiv:2105.02274. https://arxiv.org/abs/2105.02274

[10] Shuster, K., Poff, S., Chen, M., Kiela, D., & Weston, J. (2021). Retrieval augmentation reduces hallucination in conversation. arXiv preprint arXiv:2104.07567. https://arxiv.org/abs/2104.07567

[11] Roy, A., Saffar, M., Vaswani, A., & Grangier, D. (2021). Efficient content-based sparse attention with routing transformers. Transactions of the Association for Computational Linguistics, 9, 241-255. https://doi.org/10.1162/tacl_a_00364

[12] Child, R., Gray, S., Radford, A., & Sutskever, I. (2019). Generating long sequences with sparse transformers. arXiv preprint arXiv:1904.10509. https://arxiv.org/abs/1904.10509

[13] Borgeaud, S., Mensch, A., Hoffmann, J., Cai, T., Rutherford, E., Millican, K., ... & Irving, G. (2021). Improving language models by retrieving from trillions of tokens. arXiv preprint arXiv:2112.04426. https://arxiv.org/abs/2112.04426

[14] Izacard, G., & Grave, E. (2021). Leveraging passage retrieval with generative models for open domain question answering. arXiv preprint arXiv:2007.01282. https://arxiv.org/abs/2007.01282

[15] Xu, D., Zhong, X., Cambria, E., & Xu, J. (2021). Commonsense knowledge enhanced fusion-in-decoder for open-domain question answering. arXiv preprint arXiv:2109.05289. https://arxiv.org/abs/2109.05289

[16] Guu, K., Lee, K., Tung, Z., Pasupat, P., & Chang, M. W. (2020). Retrieval augmented language model pre-training. arXiv preprint arXiv:2002.08909. https://arxiv.org/abs/2002.08909

[17] Dai, Z., Yang, Z., Yang, Y., Carbonell, J., Le, Q. V., & Salakhutdinov, R. (2019). Transformer-XL: Attentive language models beyond a fixed-length context. arXiv preprint arXiv:1901.02860. https://arxiv.org/abs/1901.02860

[18] Dathathri, S., Madotto, A., Lan, J., Hung, J., Frank, E., Molino, P., ... & Liu, R. (2020). Plug and play language models: A simple approach to controlled text generation. arXiv preprint arXiv:1912.02164. https://arxiv.org/abs/1912.02164

[19] Hu, J., Xia, Q., Neubig, G., & Carbonell, J. (2021). Generating benchmarks with diverse and informative content for evaluating open-domain dialogue systems. arXiv preprint arXiv:2105.01392. https://arxiv.org/abs/2105.01392

[20] Glick, O., Klon, I., Liberman, A., Oren, Y., & Shoham, Y. (2021). Retrieval-Augmented Generation with Chunking. arXiv preprint arXiv:2108.08094. https://arxiv.org/abs/2108.08094