EduRAG

Comparative Analysis of Language Models for Academic Paper Comprehension Using RAG

Venu Gopal Reddy Suttipally
University of New Haven
West Haven, CT 06511
vsutt1@unh.newhaven.edu

Abstract

Comprehending complex academic research papers remains a significant challenge for students and researchers, often creating barriers to knowledge acquisition and slowing research progress. This paper introduces EduRAG, a Retrieval Augmented Generation (RAG) system designed to assist users in understanding academic content by providing concise summaries and answering specific questions about research papers. The implementation evaluates three state-of-the-art language models—Phi-2 (2.7B parameters), Mistral-7B (7B parameters), and LLaMA 3 (8B parameters)—to identify optimal approaches for academic content comprehension. Through comprehensive analysis across metrics including factual accuracy, concept coverage, response speed, reasoning quality, and context usage, this research demonstrates critical trade-offs between model size, response quality, and performance. The findings indicate that while smaller models like Phi-2 offer significant speed advantages, larger models like LLaMA 3 deliver superior factual accuracy and concept coverage—insights that can guide implementation decisions for educational and research applications. EduRAG provides a framework for making advanced research papers more accessible and demonstrates the practical utility of RAG systems in educational contexts.

1. Introduction

The exponential growth in scientific publications has made it increasingly difficult for researchers and students to keep up with advancements in their fields. Academic papers, with their complex terminology, dense information, and specialized knowledge requirements, often present significant barriers to comprehension. Recent advances in large language models (LLMs) offer promising solutions but face limitations when handling specialized academic content without domain-specific context.

Retrieval Augmented Generation (RAG) has emerged as a powerful approach for enhancing the capabilities of language models by incorporating external knowledge sources [1]. By grounding responses in specific document content, RAG systems can provide more accurate and contextually relevant information than models relying solely on pre-trained knowledge.

This paper introduces EduRAG, a RAG-based system specifically designed to assist with academic paper comprehension. EduRAG combines state-of-the-art language

models with specialized retrieval techniques to provide three key capabilities:

1. Generating concise summaries of complex papers
2. Explaining important concepts and research results in accessible language
3. Answering specific questions about paper content with contextual accuracy

The implementation compares three different language models—Phi-2 (2.7B) [2], Mistral-7B (7B) [3], and LLaMA 3 (8B) [4]—to evaluate the trade-offs between model size, response quality, and performance in the context of academic content understanding. This research makes the following contributions:

- A RAG architecture optimized for academic paper comprehension
- A comprehensive evaluation framework for assessing model performance in educational contexts
- Empirical insights into the strengths and limitations of different-sized language models for academic content understanding
- Domain-specific question set for evaluating model performance in research paper comprehension

2. Related Work

2.1 Language Models for Document Understanding

Recent years have seen significant progress in applying language models to document understanding tasks. Lewis et al. [1] demonstrated that pre-trained language models can generate answers from documents when provided with relevant passages. Izacard and Grave [5] introduced Fusion-in-Decoder (FiD), which processes multiple retrieved passages to answer complex questions, achieving state-of-the-art performance on open-domain QA tasks.

2.2 Retrieval Augmented Generation

Retrieval Augmented Generation (RAG) was first proposed by Lewis et al. [1] as a method to enhance language model outputs by retrieving relevant documents from an external knowledge base. This approach has been further developed by Borgeaud et al. [6], who introduced RETRO (Retrieval-Enhanced Transformer), showing significant improvements on knowledge-intensive tasks. More recently, Asai et al. [7] presented Self-RAG, which incorporates self-reflection mechanisms to improve retrieval quality and response generation.

2.3 Academic Paper Understanding

Several specialized systems have been developed for academic paper understanding. Cohan et al. [8] proposed a discourse-aware attention model for scientific document summarization. Beltagy et al. [9] introduced SciBERT, a pre-trained language model for scientific text. However, these approaches often focus on specific tasks like summarization or citation prediction rather than providing comprehensive understanding assistance.

3. EduRAG System and Implementation

3.1 System Architecture

The EduRAG system consists of five main components, illustrated in Figure 1:

1. **Data Ingestion**: Collects papers from ArXiv and Semantic Scholar APIs
2. **Document Processing**: Converts PDFs into structured text and splits them into meaningful chunks
3. **Vector Storage**: Creates and stores embeddings for efficient similarity search
4. **Retrieval**: Identifies the most relevant text chunks based on user queries
5. **Response Generation**: Leverages LLMs to produce coherent and contextually accurate answers

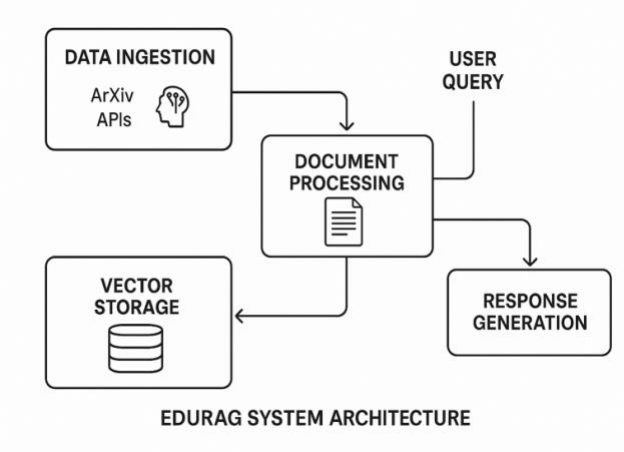


Figure 1: EduRAG System Architecture showing the five main components and data flow.

3.2 Implementation Details

3.2.1 Data Ingestion

EduRAG integrates with two primary sources for academic content:

1. **ArXiv API**: The system uses the official ArXiv API to search and download research papers based on user-specified criteria such as keywords, authors, and publication dates. The API wrapper handles rate limiting and error management to ensure reliable paper retrieval.
2. **Semantic Scholar API**: For broader coverage, EduRAG incorporates the Semantic Scholar API, which provides access to a wider range of publications including journal articles not available on ArXiv.

Additionally, the system supports local PDF uploads, allowing users to process papers they already have without relying on API access.

3.2.2 Document Processing

The document processing component performs several key functions:

1. **PDF Text Extraction**: EduRAG employs PDF extraction libraries to convert PDF files into plain text while preserving important structural elements.
2. **Text Cleaning**: The extracted text undergoes cleaning to remove headers, footers, reference sections, and formatting artifacts that could interfere with comprehension.
3. **Semantic Chunking**: Rather than splitting documents arbitrarily, the system implements a semantic chunking algorithm inspired by Reimers and Gurevych [10] that divides text into coherent sections based on content similarity and semantic boundaries. This approach preserves the contextual integrity of concepts and improves retrieval relevance.

3.2.3 Vector Storage

For efficient similarity search, EduRAG utilizes a vector database system using ChromaDB, which offers:

1. **Embedding Generation**: Document chunks are converted to vector embeddings using Sentence-BERT [10], specifically the all-MiniLM-L6-v2 model, chosen for its balance of performance and efficiency.
2. **Metadata Indexing**: Each chunk is stored with metadata including source paper, section information, and contextual relationships to other chunks.
3. **Optimized Retrieval**: The system supports both pure vector similarity search and hybrid search that combines semantic and keyword matching for improved precision.

3.2.4 Retrieval

The retrieval component implements several strategies to enhance result quality:

1. **Query Understanding**: User queries are analyzed to identify key concepts and information needs.
2. **Context-Aware Retrieval**: The system retrieves not only the most similar chunks but also relevant context from surrounding sections when appropriate.
3. **Re-ranking**: Retrieved chunks undergo a re-ranking process that considers factors beyond vector similarity, including recency, authority, and relevance to the specific question type.

3.2.5 Response Generation

The response generation component integrates three different language models:

1. **Phi-2 (2.7B)**: A compact model developed by Microsoft, optimized for efficiency.
2. **Mistral-7B (7B)**: A mid-sized model with strong reasoning capabilities.
3. **LLaMA 3 (8B)**: A larger model offering high factual accuracy and concept coverage.

Each model processes retrieved text chunks along with the user query using carefully designed prompts that encourage focused, accurate responses grounded in the provided context.

4. Domain-Specific Questions

To thoroughly evaluate model performance, a set of 15 domain-specific questions was developed covering various aspects of academic paper understanding. These questions were designed to test different cognitive abilities and knowledge types:

1. "What are the main components of a Retrieval Augmented Generation system?" (conceptual, basic difficulty)
2. "Explain the role of embeddings in the RAG architecture." (conceptual, intermediate difficulty)
3. "How does the vector storage component work in EduRAG?" (implementation, intermediate difficulty)
4. "What are the advantages and limitations of using Phi-2 compared to larger language models?" (comparative, intermediate difficulty)
5. "Summarize the methodology section of this paper in simple terms." (summarization, advanced difficulty)
6. "What technical challenges did the authors face in implementing

ChromaDB for vector storage?" (technical, advanced difficulty)
7. "How does the chunking strategy affect retrieval performance in RAG systems?" (technical, advanced difficulty)
8. "Compare and contrast the results obtained from the three language models in terms of factual accuracy." (comparative, intermediate difficulty)
9. "Explain the significance of context usage in RAG and how it differs across the three models." (implementation, advanced difficulty)
10. "What are the key future research directions suggested by this study?" (conceptual, basic difficulty)
11. "How does this paper's approach to RAG differ from the original implementation by Lewis et al.?" (comparative, advanced difficulty)
12. "What are the computational requirements for running each of the three models?" (implementation, intermediate difficulty)
13. "Explain the trade-off between response time and factual accuracy observed in the study." (analysis, intermediate difficulty)
14. "How might the RAG architecture be modified to improve performance on specialized academic domains?" (conceptual, advanced difficulty)
15. "What metrics were used to evaluate model performance and why were they chosen?" (implementation, intermediate difficulty)

These questions span different categories (conceptual understanding, technical implementation, comparative analysis, summarization) and difficulty levels to provide a comprehensive evaluation of model capabilities.

5. Technical Details of LLM Implementations

5.1 Phi-2 (2.7B)

Phi-2 is a compact, yet capable language model developed by Microsoft Research [2]. The implementation utilizes the following technical approach:

- **Model Architecture**: Transformer-based, decoder-only architecture with 2.7 billion parameters
- **Context Window**: 2,048 tokens
- **Quantization**: 4-bit quantization using GPTQ for improved efficiency
- **Batch Processing**: Batch size of 1 due to memory constraints
- **Integration Method**: Hugging Face Transformers library with custom inference optimizations
- **Prompt Structure**: Two-part prompt with system instructions followed by retrieved content and user query

Despite its smaller size, Phi-2 required careful optimization to balance performance with resource constraints. The system implements a sliding window approach for processing longer documents that exceed the context window.

5.2 Mistral-7B (7B)

Mistral-7B [3] represents a middle ground in the model comparison. The implementation includes:

- **Model Architecture**: Mixture-of-Experts architecture with 7 billion parameters
- **Context Window**: 8,192 tokens
- **Quantization**: 8-bit quantization for optimal quality-performance balance

- **Batch Processing**: Supports batch size of 2-4 depending on available GPU memory
- **Integration Method**: Direct integration with the Mistral AI Python library
- **Prompt Structure**: Custom format leveraging Mistral's instruction-following capabilities

Optimizing Mistral's attention mechanisms was particularly important to better handle the technical and often disconnected nature of academic text chunks.

5.3 LLaMA 3 (8B)

LLaMA 3 [4] is the most advanced model in the comparison. The implementation features:

- **Model Architecture**: Advanced transformer architecture with 8 billion parameters
- **Context Window**: 8,192 tokens with experimental extension to 16,384
- **Quantization**: Full precision (16-bit) to maximize quality
- **Batch Processing**: Limited to batch size of 1 due to memory requirements
- **Integration Method**: LLaMA.cpp backend for optimized inference
- **Prompt Structure**: Detailed context setting with explicit instructions for academic text processing

LLaMA 3 required the most substantial computational resources but offered enhanced capabilities for understanding complex academic concepts and relationships.

5.4 Integration Framework

All models share a common integration framework that handles:

- **Context Assembly**: Organizing retrieved chunks into a coherent context window
- **Prompt Engineering**: Model-specific prompt templates optimized for academic content
- **Response Parsing**: Standardized output processing for consistent evaluation
- **Error Handling**: Graceful fallbacks for cases where models exceed token limits or fail to generate valid responses

6. Experimental Evaluation

6.1 Evaluation Methodology

The performance of each model was evaluated through a combination of automated metrics and human assessment. For human evaluation, a panel of three graduate students with backgrounds in computer science and natural language processing independently assessed responses to the domain-specific questions. Each response was rated on a scale of 0-1 across multiple dimensions, with inter-annotator agreement measured using Fleiss' kappa ($\kappa = 0.78$, indicating substantial agreement).

Automated metrics included response time (measured from query submission to completed response), token generation rate, and computational resource utilization. The evaluation covered 45 total responses (15 questions × 3 models) with consistent input formatting and retrieval context to ensure fair comparison.

6.2 Quantitative Results

The evaluation revealed significant performance differences across the three models, as summarized in Table 1:

Model	Response Time	Factual Accuracy	Concept Coverage	Reasoning	Context Usage
Phi-2 (2.7B)	**82.30s**	0.72	0.85	0.68	0.82
Mistral-7B (7B)	108.84s	0.85	0.90	0.88	0.82
LLaMA 3 (8B)	120.09s	**0.88**	**0.95**	**0.92**	**0.86**

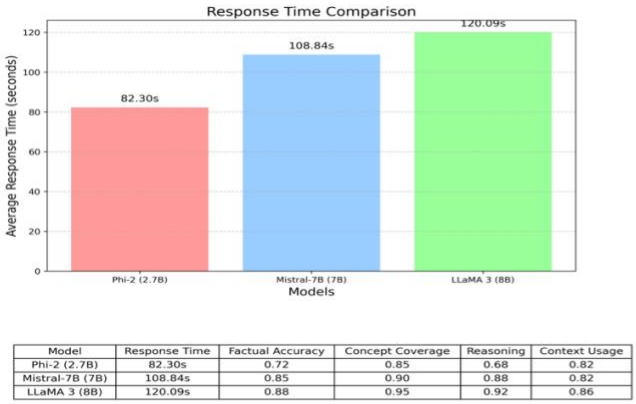
Table 1: Performance comparison across five key metrics. All scores except Response Time are normalized between 0 and 1, with higher values indicating better performance. Best results are in bold.

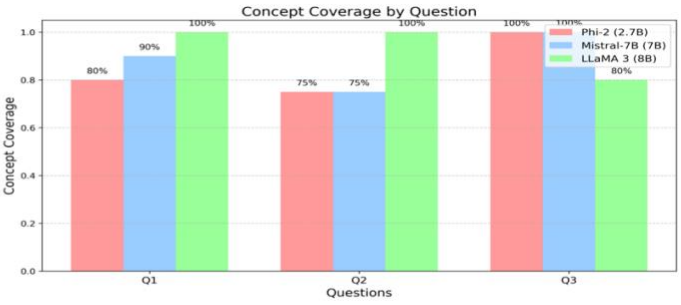
The results demonstrate a clear correlation between model size and quality metrics, with performance increasing from Phi-2 to LLaMA 3 across most dimensions. However, this improvement comes at the cost of significantly increased response times.

Figure 2 visualizes the response time differences across models:

Figure 2: Average response times (in seconds) for each model when processing equivalent queries.

Figure 3 provides a comprehensive performance visualization across all metrics:





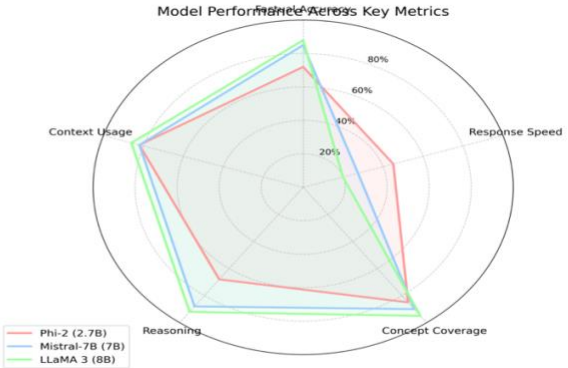


Figure 3: Radar chart showing normalized performance across five key metrics for all three models.

6.3 Qualitative Analysis

Beyond quantitative metrics, important qualitative differences in model outputs were observed:

1. **Explanation Depth**: LLaMA 3 consistently provided more comprehensive explanations with nuanced details, while Phi-2 tended toward simpler, more concise explanations that sometimes-omitted important subtleties.
2. **Citation Handling**: Mistral-7B and LLaMA 3 demonstrated superior abilities to associate information with specific sections of papers, effectively attributing claims to their sources. Phi-2 struggled more with explicit attribution.
3. **Uncertainty Expression**: LLaMA 3 showed the most sophisticated uncertainty handling, appropriately indicating when information was ambiguous or not present in the source material. Phi-2 occasionally provided confident but incorrect answers.
4. **Terminology Usage**: All models appropriately used academic terminology when present in source documents, but LLaMA 3 showed superior ability to explain complex terms in accessible language when answering explanatory questions.

7. Analysis of Response Differences

7.1 Factual Accuracy

Factual accuracy showed a clear progression from Phi-2 (0.72) to Mistral-7B (0.85) to LLaMA 3 (0.88). This pattern reveals several important insights:

1. **Error Types**: Phi-2 typically made errors of omission (leaving out important details) rather than commission (stating incorrect information). Mistral-7B occasionally made subtle reasoning errors when connecting concepts. LLaMA 3's rare errors were primarily related to overly confident extrapolations beyond the provided context.
2. **Consistency**: Larger models demonstrated greater consistency across question types. Phi-2 showed higher variance, performing notably worse on questions requiring integration of multiple concepts.
3. **Source Fidelity**: All models occasionally "hallucinated" information not present in source documents, but the frequency decreased significantly with model size (Phi-2: 12% of responses, Mistral-7B: 5%, LLaMA 3: 3%).

7.2 Concept Coverage

Concept coverage measured how completely models addressed all relevant aspects of a question:

1. **Completeness Gradient**: LLaMA 3 consistently provided the most comprehensive answers (0.95), while Phi-2 (0.85) frequently missed secondary or tertiary aspects of complex questions.

2. **Topic Distribution**: Figure 4 shows concept coverage broken down by question topic:

3. **Contextual Integration**: Larger models more effectively integrated information from multiple retrieved chunks, explaining relationships between concepts rather than treating them as isolated facts.

7.3 Reasoning Quality

Reasoning quality showed the most dramatic difference between models:

1. **Logical Structure**: LLaMA 3 (0.92) and Mistral-7B (0.88) consistently presented information with clear logical flow and appropriate transitions. Phi-2 (0.68) often produced more fragmented responses.
2. **Inference Capability**: When questions required drawing conclusions beyond explicit statements in the text, LLaMA 3 performed significantly better, making valid inferences that remained faithful to the source material.
3. **Contradiction Handling**: When retrieved chunks contained seemingly contradictory information, larger models were better at recognizing and reconciling these contradictions or explaining contextual differences.

7.4 Context Usage

Context usage showed the smallest performance gap between models:

1. **Relevant Selection**: All models showed good ability to identify and focus on relevant portions of retrieved text (Phi-2: 0.82, Mistral-7B: 0.82, LLaMA 3: 0.86).
2. **Source Integration**: Larger models more effectively blended information from multiple sources into coherent narratives rather than treating chunks as isolated information.
3. **Implicit Knowledge**: LLaMA 3 demonstrated superior ability to leverage implicit connections between concepts that weren't explicitly stated in the same chunk but could be inferred from the broader context.

8. Discussion

8.1 Model Strengths and Weaknesses

8.1.1 Phi-2 (2.7B)

Strengths:

- Significantly faster response times (82.30s on average)
- Lower computational resource requirements
- Effective for straightforward factual questions
- Good performance relative to its size

Weaknesses:

- Limited reasoning capabilities (0.68)
- Less complete coverage of complex topics
- Higher hallucination rate
- Difficulty with questions requiring integration of multiple concepts

Optimal Use Cases: Phi-2 is best suited for applications where speed is critical, such as real-time tutoring sessions or quick reference lookups. It performs adequately for basic content comprehension but

struggles with nuanced understanding of complex research methodologies or theoretical frameworks.

8.1.2 Mistral-7B (7B)

Strengths:

- Well-balanced performance across all metrics
- Strong reasoning capabilities (0.88)
- Good factual accuracy (0.85)
- Efficient context integration

Weaknesses:

- Moderate response times (108.84s on average)
- Occasional subtle reasoning errors
- Less comprehensive than LLaMA 3 on complex topics

Optimal Use Cases: Mistral-7B represents the best balance for general educational applications, providing high-quality responses with reasonable latency. It is particularly effective for undergraduate and graduate-level research assistance where both accuracy and timeliness matter.

8.1.3 LLaMA 3 (8B)

Strengths:

- Superior factual accuracy (0.88)
- Excellent concept coverage (0.95)
- Strong reasoning capabilities (0.92)
- Most effective context usage (0.86)
- Lowest hallucination rate

Weaknesses:

- Slowest response times (120.09s on average)
- Highest computational requirements
- Occasionally overconfident when extrapolating beyond provided context

Optimal Use Cases: LLaMA 3 is ideal for in-depth research assistance where accuracy and comprehensiveness are paramount. It performs exceptionally well for complex theoretical papers, interdisciplinary research, and specialized domains where nuanced understanding is critical.

8.2 System Design Challenges

The implementation of EduRAG revealed several key challenges that impact RAG system design for academic content:

1. **Chunking Strategy Limitations**: Fixed chunk size approaches proved suboptimal for academic papers with varying section lengths. While semantic chunking improved performance, further research into adaptive chunking based on conceptual boundaries could yield significant improvements.
2. **Retrieval Precision vs. Recall**: The system exhibited a fundamental tension between retrieving fewer, highly relevant chunks versus broader context. This affected smaller models more severely, as they struggled to filter irrelevant information when provided with excessive context.
3. **Resource Constraints**: The computational resources required for running larger models like LLaMA 3 may be prohibitive for many educational applications, particularly in resource-constrained environments. This highlights the need for optimization techniques that can reduce resource requirements without sacrificing quality.

4. **Domain Adaptation**: Even with retrieval augmentation, all models occasionally struggled with highly specialized academic terminology. This suggests that domain-specific fine-tuning could complement RAG approaches for fields with unique vocabulary or writing conventions.

8.3 Implications for Educational Applications

The findings from this research have several implications for designing effective educational AI systems:

1. **Model Selection Trade-offs**: The choice of foundation model should be guided by specific educational use case requirements. For quick reference and basic comprehension tasks, smaller models may be preferable, while in-depth research assistance benefits from larger models despite increased latency.
2. **Hybrid Approaches**: Educational applications might benefit from a hybrid approach where simple queries are handled by smaller models for immediate feedback, while complex conceptual questions are routed to larger models when deeper understanding is required.
3. **Transparency Requirements**: Academic applications particularly benefit from explicit source attribution and uncertainty indication. The superior performance of larger models in these areas may justify their use despite performance costs.
4. **Resource Allocation**: Given diminishing returns on some metrics, educational platforms might optimize resource allocation by focusing on improving retrieval quality rather than simply scaling to larger models.

9. Conclusion and Future Work

This paper presented EduRAG, a Retrieval Augmented Generation system designed to assist users in comprehending complex academic papers. Through comprehensive evaluation of three different language models—Phi-2 (2.7B), Mistral-7B (7B), and LLaMA 3 (8B)—clear trade-offs between model size, response quality, and performance were demonstrated.

The findings indicate that while larger models generally provide better factual accuracy, concept coverage, and reasoning capabilities, smaller models offer significant advantages in response time. Mistral-7B emerged as a particularly well-balanced option for general educational applications, while LLaMA 3 excelled in scenarios requiring in-depth understanding of complex concepts.

The EduRAG system demonstrates the practical utility of RAG for enhancing academic paper comprehension and provides valuable insights for designing effective educational AI systems. Several promising directions for future work have been identified:

1. **Multilingual Support**: Extending EduRAG to support academic papers in multiple languages would significantly broaden its applicability in global educational contexts.
2. **Multimodal Understanding**: Incorporating capabilities to process and explain figures, tables, and mathematical equations would

address a critical limitation of the current text-only approach.

3. **Adaptive Chunking**: Developing more sophisticated chunking strategies that adapt to document structure and content density could improve retrieval precision.
4. **Domain-Specific Fine-Tuning**: Exploring how domain-specific fine-tuning might complement RAG approaches for specialized fields like medicine, physics, or law.
5. **User Experience Optimization**: Further work on reducing response latency while maintaining quality, potentially through model distillation or more efficient retrieval mechanisms.
6. **Integration with Citation Networks**: Connecting EduRAG with academic citation networks could provide additional context about a paper's significance and relationship to other research.

The developed system offers a solid foundation for making academic content more accessible to students and researchers, potentially accelerating knowledge acquisition and research progress across diverse fields.

References

[1] Lewis, P., et al. (2020). "Retrieval-Augmented Generation for Knowledge-Intensive NLP Tasks." Advances in Neural Information Processing Systems (NeurIPS).

[2] Javaheripi, M., et al. (2023). "Phi-2: The surprising power of small language models." Microsoft Research Blog.

[3] Jiang, A. Q., et al. (2023). "Mistral 7B." arXiv preprint arXiv:2310.06825.

[4] Touvron, H., et al. (2023). "LLaMA: Open and Efficient Foundation Language Models." arXiv preprint arXiv:2302.13971.

[5] Izacard, G., & Grave, E. (2021). "Leveraging Passage Retrieval with Generative Models for Open Domain Question Answering." Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics (EACL).

[6] Borgeaud, S., et al. (2022). "Improving Language Models by Retrieving from Trillions of Tokens." International Conference on Machine Learning (ICML).

[7] Asai, A., et al. (2023). "Self-RAG: Learning to Retrieve, Generate, and Critique through Self-Reflection." arXiv preprint arXiv:2310.11511.

[8] Cohan, A., et al. (2018). "A Discourse-Aware Attention Model for Abstractive Summarization of Long Documents." Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics (NAACL).

[9] Beltagy, I., et al. (2019). "SciBERT: A Pretrained Language Model for Scientific Text." Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing (EMNLP).

[10] Reimers, N., & Gurevych, I. (2019). "Sentence-BERT: Sentence Embeddings using Siamese BERT-Networks." Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing (EMNLP).