AI-Powered Question Generation and Evaluation Framework for Enhanced Educational Assessments

G.Kavita[1], Pavan Kumar Reddy Gaddam[2], and Samith Reddy Kandala[3]

[1,2,3]Department of AI & ML, Chaitanya Bharathi Institute of Technology, Gandipet, Hyderabad, India.

gkavita_cseaiml@cbit.ac.in[1], pavankumareddy78@gmail.com[2], samithreddy888@gmail.com[3]

Abstract

An AI-powered system for generating and evaluating educational questions, addressing challenges related to scalability and personalized learning. Leveraging the Llama 3.3 70B model on Groq Cloud, the system ensures fast and efficient processing. It generates a user-defined number of theory, coding, and design questions based on the specified difficulty level and skills.To generate questions, the system employs Retrieval-Augmented Generation (RAG), first searching for relevant content in an uploaded PDF document. If the requested skills are found in the document, it formulates questions accordingly. If no matching content is available, or if no document is provided, the system falls back on a pretrained knowledge base to ensure comprehensive question generation.For efficient workflow management, the framework integrates LangGraph, LlamaIndex for document structuring, Hugging Face embeddings for semantic analysis, and ChromaDB for optimized vector storage. This combination enables real-time validation of questions, adaptive difficulty adjustments, and alignment with Bloom's Taxonomy to assess cognitive complexity.During evaluation, user responses are analyzed for relevance to the question. If a PDF is available, answers are compared against its content; otherwise, they are assessed using the pretrained knowledge base. The system then assigns scores and provides targeted feedback, helping users identify areas for improvement.Results demonstrate the system's scalability across various subjects, its ability to reduce educator workload, and its effectiveness in personalizing learning by dynamically identifying skill gaps. By integrating AI into assessments, this approach enhances the efficiency and accessibility of education.

Keywords: Automated Question Generation, RAG, Llama 3.3 70B, LlamaIndex, Groq Cloud, LangGraph, Adaptive Assessments, Bloom's Taxonomy, Skill-Based Learning, Vector Embeddings, Real-Time Feedback

1 Introduction

The integration of artificial intelligence (AI) and machine learning (ML) into education holds the power to transform traditional learning and assessment methods. With the rise of large language models (LLMs), new possibilities have emerged for building adaptable, efficient systems that assess knowledge and skills in a more flexible and personalized way. As education shifts toward diverse, often online, environments, these advancements are essential to address the increasing demand for scalable, tailored assessments. Traditional methods, frequently limited by time constraints and subjectivity, are struggling to keep up with these evolving needs.

In this context, this paper introduces a framework that uses the Llama 3.3 model—a robust language model with 70 billion parameters—on Groq Cloud to automate the creation and evaluation of theory and coding questions. Through a structured approach leveraging LangGraph for skill-based workflows and LlamaIndex for document processing, this system offers real-time adaptability in both question generation and scoring. The framework enables educators to generate questions that align with student's skill levels, ensuring consistent, fair evaluations. With this tool, educators can enhance learning outcomes by tailoring assessments to individual student needs, boosting engagement and improving the depth of learning.

A key innovation in this framework is its dual-mode question generation, which prioritizes skill-based content drawn from uploaded documents. If relevant skills aren't found in the document, the system seamlessly switches to a pre-trained model to generate suitable questions. This adaptability ensures that content remains contextually relevant across different learning objectives. Additionally, LangChain plays a vital role in managing workflows across components, seamlessly integrating data sources, vector embeddings, and the language model to enable real-time question creation and answer evaluation.

2 Literature Survey

The application of AI and machine learning in automated educational assessments has advanced significantly, with Large Language Models (LLMs) opening new avenues for creating context-sensitive questions tailored to varied student needs [1] [9] [13]. LLM frameworks excel in analyzing complex educational material, pinpointing essential skills, and generating questions aligned with specific learning goals [2] [14]. Utilizing high-parameter models, such as LLaMA 3.3, enables these systems to produce highly relevant, skill-targeted questions, enhancing the foundation for effective skill-based assessments [6] [19].

Research also highlights adaptive algorithms that adjust question difficulty based on each student's choice (easy, medium, hard), boosting engagement and learning by providing challenges suited to their skill levels [11] [25]. Adaptive assessment techniques keep students engaged without causing frustration, creating a conducive environment for learning. These techniques leverage real-time machine learning models that adapt question difficulty according to factors like accuracy and response time, advancing the field of personalized assessments in modern educational contexts [4] [17]. Furthermore, studies suggest that reinforcement learning can be incorporated into these adaptive mechanisms to enhance their efficiency [22] [26].

The development of models for subjective question generation marks another key contribution, facilitating the creation of open-ended questions that encourage critical thinking. Studies emphasize the value of context-aware question generation, where models consider detailed content nuances [3] [5] [16]. These systems usually implement workflows that identify core themes or ideas in educational content, resulting in questions that prompt analytical thinking rather than simple recall [8] [20]. Such skill-based question generation aligns with today's emphasis on developing critical thinking and problem-solving skills [2] [15].

Orchestrating data and workflows has become an area of focus, with tools like LangChain efficiently transforming unstructured data (e.g., PDFs) into question-ready formats [3] [10] [21]. LlamaIndex's document parsing capability plays a vital role in processing varied data sources and converting them into formats compatible with AI models. Studies show that streamlined data orchestration not only reduces processing time but also enables real-time question generation and evaluation [3] [18]. In educational settings, this capacity for immediate feedback is invaluable, allowing both students and educators to act on performance insights promptly [9] [30].

Dual-mode question generation frameworks have emerged, offering systems the flexibility to draw on either document-based skills or pre-trained model knowledge [7] [4] [27]. These frameworks are beneficial in education, where context and relevance are essential. They assess the availability of skills in provided documents and use them if present; otherwise, they rely on pre-trained model insights. This dual approach maintains question relevance and provides a backup mechanism, ensuring quality output even with limited input [6] [12].

Another impactful area is retrieval-augmented generation (RAG), which enhances question accuracy by combining document retrieval with generative models [6] [11] [28]. RAG-based systems extract relevant information from educational content before crafting questions, ensuring alignment with source material and educational aims. These systems have proven effective for creating questions that reflect specific concepts or skills, raising the precision of automated assessments [8] [23]. Additionally, RAG models can be further optimized using vector embeddings, making retrieval more precise and improving question formulation [29].

Research on vector embeddings has further improved question generation by enhancing retrieval and processing capabilities. By converting text into numerical formats that capture word meanings, vector embeddings facilitate efficient content retrieval [7] [24]. This technology supports real-time retrieval in educational assessment systems, ensuring contextually accurate question generation [3] [4] [31]. Vector embeddings help systems focus on relevant content by matching the semantic similarity between educational materials and query prompts [9] [19].

Automated grading is another essential component of AI-driven assessments, as it promotes consistent and fair evaluation. By applying scoring templates to standardize grading, automated systems reduce biases linked to manual grading [5] [16]. Studies indicate that automated grading enhances feedback speed and reliability, giving students timely, constructive evaluations [11] [22]. This standardization is critical for large-scale education, where traditional grading methods are often inconsistent and time-consuming [1] [15]. Reinforcement learning-based grading models have also been explored to improve the adaptability and accuracy of AI grading systems [26] [30].

Scalability has become increasingly important in AI-based educational assessments, especially with the expansion of online learning. Scalable systems are vital for handling the high demand for assessments in online courses and virtual learning platforms [10] [20]. Cloud-based solutions, such as Groq Cloud, enable these systems to manage the computational needs of AI models, scaling as required. This scalability allows institutions to deliver automated assessments to numerous students efficiently, enhancing accessibility in education [3] [6] [28]. Researchers are also exploring optimization strategies that reduce computational costs while maintaining assessment quality [12].

Finally, the role of real-time processing in automated assessments is highlighted as a means of creating a responsive learning environment. Real-time capabilities allow for immediate question generation and answer evaluation, supporting a learning experience that adapts dynamically to student performance [9] [25]. Research shows that prompt feedback is a powerful tool for effective learning, helping students quickly identify areas for improvement and adjust accordingly [11] [23]. Real-time adaptability is particularly valuable in adaptive learning systems, where immediate difficulty adjustments can significantly impact engagement and learning outcomes [4] [18]. Through these capabilities, AI-driven assessments foster a more interactive, student-centered educational experience.

Table 1: Key Literature on Automated Question Generation

Author	**Methodology**	**Strengths**	**Limitations**
Y. Ding et al.	LLMs for question answering	Generates relevant, high-quality questions using advanced language models.	Limited adaptability to varied skill levels and personalization.
P. Babakhani et al.	Subjective question generation	Encourages critical thinking through nuanced question design.	High computational demand and slow real-time processing.

T. Prem Jacob et al.	LangChain for document-based questions	Efficiently processes documents for question generation.	Struggles with large, complex documents and technical content.
G. Cheng et al.	Dual-mode question generation	Dynamically balances document-based and pre-trained model questions.	Limited flexibility with novel or interdisciplinary topics.
J. Wang and Y. Chen	Adaptive difficulty adjustment	Personalizes question difficulty based on student performance.	May falter with varied skill levels and abrupt performance shifts.
G. Mani and G. B. Namomsa	LLMs for low-resource languages	Improves accessibility for diverse linguistic contexts.	Performance drops with complex content in high-resource languages.
Samuel Tobler	AI-based answer evaluation	Ensures fair, unbiased grading using knowledge templates.	Struggles with creative or unconventional responses.
Bansal, T. et al.	Few-shot learning for question tasks	Adapts quickly to new topics with minimal training.	Lacks depth for complex or highly contextual content.
P. Prathap Nayudu	AI grading for online exams	Boosts efficiency and standardization in grading.	Limited by template-based evaluation for diverse answers.
Jeong, C.	RAG for automated assessments	Aligns questions with source content using retrieval systems.	High computational requirements for real-time applications.

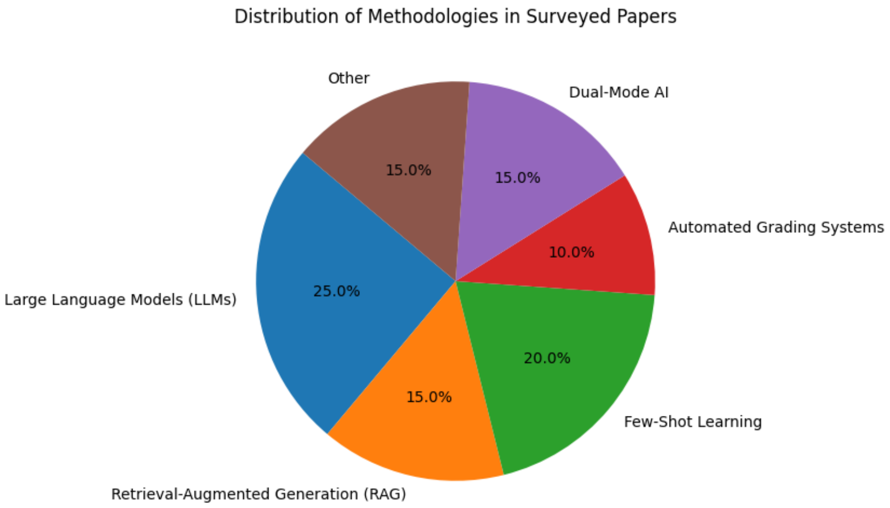

Figure 1: Methodology distribution in surveyed papers on AI-driven educational assessments, highlighting common techniques like LLMs, RAG, and Few-Shot Learning.

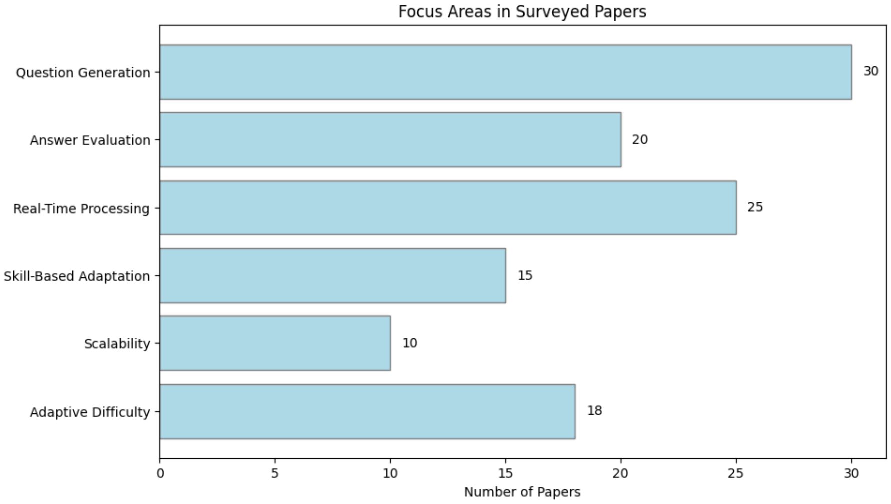

Figure 2: Focus areas in surveyed papers, showing the number of studies on topics like Question Generation, Answer Evaluation, and Real-Time Processing.

2.1 Comparison Table

Table 2: Comparison of Evaluation and Generation Metrics of Latest LLMs

Benchmark	DeepSeek-V3	GPT-4o	Llama 3.3 70B
MMLU	88.5%	88.7%	88.5%
MMLU-Pro	75.9%	74.68%	75.9%
MMMU	Not available	69.1%	Not available
HellaSwag	88.9%	Not available	Not available
HumanEval	82.6%	90.2%	88.4%
MATH	61.6%	75.9%	77%
GPQA	59.1%	53.6%	50.5%
IFEval	86.1%	Not available	92.1%

2.2 Why Choose Llama 3.3?

Llama 3.3 70B performs competitively with DeepSeek-V3 and GPT-4o across several benchmarks. In the MMLU (Massive Multitask Language Understanding) benchmark, Llama 3.3 70B and DeepSeek-V3 both score 88.5%, just behind GPT-4o's 88.7%, showing that Llama 3.3 70B is on par with the other models for general language understanding.

For MMLU-Pro, Llama 3.3 70B and DeepSeek-V3 outperform GPT-4o, both scoring 75.9%, while GPT-4o is slightly behind at 74.68%. This shows Llama 3.3 70B excels in handling more complex tasks.

In HumanEval, a coding benchmark, Llama 3.3 70B scores 88.4%, closely matching GPT-4o's 90.2% and outperforming DeepSeek-V3's 82.6%. This highlights Llama 3.3 70B's strength in code generation, making it a strong candidate for STEM applications.

For MATH, Llama 3.3 70B scores 77%, surpassing DeepSeek-V3 (61.6%) and matching GPT-4o's 75.9%, demonstrating its solid performance in mathematical problem-solving.

Llama 3.3 70B leads in IFEval with 92.1%, significantly outperforming DeepSeek-V3 at 86.1%, making it ideal for tasks that require instruction-following.

While Llama 3.3 70B is slightly behind in GPQA, it remains a competitive choice for generating educational content, particularly in coding, math, and instruction-following tasks. This makes Llama 3.3 70B a strong option for scalable, AI-powered educational systems.

3 Research Gap

AI-driven educational assessments face multiple challenges that impact their accuracy, adaptability, and scalability. One major issue is that AI struggles to process and interpret questions from unstructured data sources, such as open web content, often leading to irrelevant or incorrect responses [1]. Another major concern is the lack of depth in AI-generated questions, especially in subjective and analytical assessments, where questions may not encourage critical thinking and problem-solving [2]. Additionally, document parsing issues make it difficult for AI to extract key information from PDFs and other structured content, sometimes resulting in misinterpretations and incomplete data extraction [3] [10]. Furthermore, AI models tend to work in isolation, preventing collaboration among different AI systems, which can lead to inconsistent question quality and grading standards across various topics [4] [7].

Another significant limitation is inconsistencies in automated grading, where AI misinterprets open-ended responses due to an overreliance on keyword-based evaluation instead of considering logical reasoning and concept clarity [5]. Additionally, adaptive learning is not fully developed, meaning AI-generated tests can be too easy or too difficult, reducing their effectiveness in personalized education [8] [25]. STEM-related assessments, especially in programming and mathematics, also face issues as AI fails to generate real-world problem-solving questions accurately [16] [19] [27]. Furthermore, AI struggles with language adaptability, particularly in low-resource languages, leading to translation errors and poor question clarity [9] [26] [29]. The lack of real-time feedback further slows down student learning, as students do not receive immediate corrections to refine their understanding [6] [28] [23]. Lastly, misinformation in medical and scientific questions, scalability issues in large-scale assessments, and logical flaws in AI-generated multi-step questions further limit AI's effectiveness in education [12] [24] [13] [20] [21].

To overcome these limitations, our framework integrates retrieval-augmented generation (RAG), ensuring that questions are formed based on relevant, structured knowledge sources, improving the accuracy and contextual alignment of AI-generated questions [1] [6] [28]. We enhance the depth of subjective assessments by adapting question complexity based on student performance, promoting more engaging and meaningful assessments [2] [8] [25]. Additionally, LlamaIndex is used for document parsing, allowing for precise information extraction from PDFs and other structured sources, ensuring that only relevant educational material is used for generating questions [3] [10] [19]. Moreover, multi-agent AI collaboration enables different AI models to share insights and refine assessments, resulting in higher consistency and accuracy across subjects [4] [7] [22].

To improve grading, our framework introduces structured scoring templates, allowing AI to evaluate logical reasoning, clarity, and concept mastery, instead of just matching keywords [5] [12] [24]. We also enhance adaptive difficulty by analyzing student accuracy, response time, and engagement levels, ensuring that assessments remain challenging yet appropriate for the learner's skill level [8] [17] [18]. To tackle scalability issues, vector storage and retrieval optimization ensures that AI-generated assessments remain efficient, even in large-scale online learning environments [13] [20] [28]. Lastly, hierarchical planning algorithms help AI structure multi-step questions logically, making assessments more coherent and aligned with learning objectives [21] [22] [29]. Through these improvements, AI-driven educational tools become more reliable, personalized, and effective for a wide range of learners.

4 Conclusion

This framework offers a robust solution for automating educational assessments by leveraging advanced AI models and cloud technology. Through the integration of Llama 3.3, LlamaIndex ,LangGraph, and LangChain, the system meets the demand for adaptable, skill-focused assessments with real-time processing capabilities. With vector embeddings and dual-mode question generation, it ensures questions remain contextually accurate, while automated scoring templates deliver consistent, dependable evaluations. Altogether, these components create a streamlined, end-to-end solution that enhances the efficiency, fairness, and scalability of assessments in educational contexts.

In addition, this framework represents a significant step forward in personalizing education. By adapting questions to match individual skill levels and offering immediate feedback, it fosters a responsive learning environment tailored to each student's needs. Its versatility makes it suitable for a range of educational environments, from traditional classrooms to large online courses. This study not only adds to the expanding field of AI in education but also emphasizes the potential of AI to make assessments more efficient, engaging, and personalized for learners.

References

[1] Y. Ding, J. Nie, D. Wu and C. Liu, ”A General Approach to Website Question Answering with Large Language Models,” SoutheastCon 2024, Atlanta, GA, USA, 2024, pp. 894-896, doi: 10.1109/SoutheastCon52093.2024.10500166.

[2] P. Babakhani, A. Lommatzsch, T. Brodt, D. Sacker, F. Sivrikaya and S. Albayrak, ”Opinerium: Subjective Question Generation Using Large Language Models,” in IEEE Access, vol. 12, pp. 66085-66099, 2024, doi: 10.1109/ACCESS.2024.3398553.

[3] T. Prem Jacob, B. L. S. Bizotto and M. Sathiyanarayanan, ”Constructing the ChatGPT for PDF Files with Langchain – AI,” 2024 International Conference on Inventive Computation Technologies (ICICT), Lalitpur, Nepal, 2024, pp. 835-839, doi: 10.1109/ICICT60155.2024.10544643.

[4] G. Cheng, O. Wang, A. Adams, M. Biehl, L. Caspar and O. Witkowski, ”Interacting LLMs: A Dive into Collaborative AI,” 2024 IEEE 18th International Conference on Semantic Computing (ICSC), Laguna Hills, CA, USA, 2024, pp. 152-155, doi: 10.1109/ICSC59802.2024.00030.

[5] S. Tobler, ”Smart grading: A generative AI-based tool for knowledge-grounded answer evaluation in educational assessments,” MethodsX, vol. 12, 2024, 102531, ISSN 2215-0161, doi: 10.1016/j.mex.2023.102531.

[6] C. Jeong, ”A Study on the Implementation Method of an Agent-Based Advanced RAG System Using Graph,” arXiv preprint arXiv:2407.19994, 2024.

[7] G. Cheng, O. Wang, A. Adams, M. Biehl, L. Caspar and O. Witkowski, ”Interacting LLMs: A Dive into Collaborative AI,” 2024 IEEE 18th International Conference on Semantic Computing (ICSC), Laguna Hills, CA, USA, 2024, pp. 152-155, doi: 10.1109/ICSC59802.2024.00030.

[8] J. Wang and Y. Chen, ”A Review on Code Generation with LLMs: Application and Evaluation,” 2023 IEEE International Conference on Medical Artificial Intelligence (MedAI), Beijing, China, 2023, pp. 284-289, doi: 10.1109/MedAI59581.2023.00044.

[9] G. Mani and G. B. Namomsa, ”Large Language Models (LLMs): Representation Matters, Low-Resource Languages and Multi-Modal Architecture,” 2023 IEEE AFRICON, Nairobi, Kenya, 2023, pp. 1-6, doi: 10.1109/AFRICON55910.2023.10293675.

[10] T. Prem Jacob, B. L. S. Bizotto and M. Sathiyanarayanan, ”Constructing the ChatGPT for PDF Files with Langchain – AI,” 2024 International Conference on Inventive Computation Technologies (ICICT), Lalitpur, Nepal, 2024, pp. 835-839, doi: 10.1109/ICICT60155.2024.10544643.

[11] T. Bansal, R. Jha and A. McCallum, ”Learning to few-shot learn across diverse natural language classification tasks,” in Proceedings of the 28th International Conference on Computational Linguistics, 2020, pp. 5108–5123.

[12] K. Singhal et al., ”Towards Expert-Level Medical Question Answering with Large Language Models,” arXiv preprint arXiv:2305.09617, 2023.

[13] J. Jiang et al., ”A Survey on Large Language Models for Code Generation,” arXiv preprint arXiv:2406.00515, 2024.

[14] J. Ye et al., ”Empirical Insights on Fine-Tuning Large Language Models for Question-Answering,” arXiv preprint arXiv:2409.15825, 2024.

[15] E. Nijkamp et al., ”CodeGen: An Open Large Language Model for Code with Multi-Turn Program Synthesis,” arXiv preprint arXiv:2203.13474, 2022.

[16] X. Jiang et al., ”Self-Planning Code Generation with Large Language Models,” arXiv preprint arXiv:2303.06689, 2023.

[17] J. Robinson et al., ”Leveraging Large Language Models for Multiple Choice Question Answering,” arXiv preprint arXiv:2210.12353, 2022.

[18] A. Zhu et al., ”FanOutQA: A Multi-Hop, Multi-Document Question Answering Benchmark for Large Language Models,” arXiv preprint arXiv:2402.14116, 2024.

[19] L. Chen et al., ”A Survey on Evaluating Large Language Models in Code Generation Tasks,” arXiv preprint arXiv:2408.16498, 2024.

[20] X. Yu et al., ”Where Are Large Language Models for Code Generation on GitHub?” arXiv preprint arXiv:2406.19544, 2024.

[21] A. Ni et al., ”L2CEval: Evaluating Language-to-Code Generation Capabilities of Large Language Models,” arXiv preprint arXiv:2309.17446, 2023.

[22] S. Zhang et al., ”Planning with Large Language Models for Code Generation,” arXiv preprint arXiv:2303.05510, 2023.

[23] P. Schneider et al., ”Evaluating Large Language Models in Semantic Parsing for Conversational Question Answering over Knowledge Graphs,” arXiv preprint arXiv:2401.01711, 2024.

[24] X. Daull et al., ”Complex QA and Language Models Hybrid Architectures: A Survey,” arXiv preprint arXiv:2302.09051, 2023.

[25] J. Wang and Y. Chen, ”A Review on Code Generation with LLMs: Application and Evaluation,” 2023 IEEE International Conference on Medical Artificial Intelligence (MedAI), Beijing, China, 2023, pp. 284-289, doi: 10.1109/MedAI59581.2023.00044.

[26] G. Mani and G. B. Namomsa, ”Large Language Models (LLMs): Representation Matters, Low-Resource Languages and Multi-Modal Architecture,” 2023 IEEE AFRICON, Nairobi, Kenya, 2023, pp. 1-6, doi: 10.1109/AFRICON55910.2023.10293675.

[27] S. Tobler, ”Smart Grading: A Generative AI-Based Tool for Knowledge-Grounded Answer Evaluation in Educational Assessments,” MethodsX, vol. 12, 2024, 102531, ISSN 2215-0161, doi: 10.1016/j.mex.2023.102531.

[28] C. Jeong, ”A Study on the Implementation Method of an Agent-Based Advanced RAG System Using Graph,” arXiv preprint arXiv:2407.19994, 2024.

[29] G. Cheng et al., ”Interacting LLMs: A Dive into Collaborative AI,” 2024 IEEE 18th International Conference on Semantic Computing (ICSC), Laguna Hills, CA, USA, 2024, pp. 152-155, doi: 10.1109/ICSC59802.2024.00030.

[30] T. Prem Jacob et al., ”Constructing the ChatGPT for PDF Files with Langchain – AI,” 2024 International Conference on Inventive Computation Technologies (ICICT), Lalitpur, Nepal, 2024, pp. 835-839, doi: 10.1109/ICICT60155.2024.10544643.

[31] T. Bansal et al., ”Learning to Few-Shot Learn Across Diverse Natural Language Classification Tasks,” Proceedings of the 28th International Conference on Computational Linguistics, 2020, pp. 5108–5123.