Multi-Agent Retrieval Augmented Generation BOT for Attributed Question Answering

Dr. Bhagyashree Dharaskar[1], Om Pawar[2], Chetan Rautiya[3], Shrutika Dongre[4], Priti Yadav[5], Sneha Tembhare[6]

Department of Computer Science and Engineering Priyadarshini
College of Engineering, Nagpur

1. bdharaskar@gmail.com 2. ompawar2526@gmail.com 3. chetanrautiya1@gmail.com 4. shrutikadongre6730@gmail.com
5. pritiparasyadav2005@gmail.com 6. snehatembhare2005@gmail.com

Abstract. Multi-agent systems (MAS) provide a powerful architecture for addressing distributed and complex problems by deploying multiple autonomous agents to team up toward a common objective. In a multi-agent system, each agent possesses the ability to perceive its environment, make decisions, communicate with other agents, and execute actions independently. Our project- Multi-Agent Retrieval-Augmented Generation (RAG) framework for attributed question answering (QA) is an advanced AI framework, where multiple intelligent agents collaboratively perform retrieval and generation tasks to produce more accurate, reliable, and context-aware responses. With the goal of trustworthy answer generation, our work focuses on optimizing answer correctness, defined by coverage and relevance to the question and faithfulness, which measures the extent to which answers are grounded in retrieved documents. It uses a multi-agent architecture that iteratively filters retrieved documents, generates attributed answers within-line-citations, and verifies com completeness through dynamic refinement. Central to the framework is a hybrid retrieval strategy that combines sparse and dense methods that will improve recall compared to the best single retrieval model, resulting in more correct and well-supported answers. We are evaluating project on a synthetic QA data set derived from the Fine Web index. Our work surely will outperform standard RAG baselines, achieving gains in correctness and faithfulness. It will demonstrate the effectiveness of the multi-agent architecture and hybrid retrieval advancing trustworthy QA. Our project attempts to mitigate privacy issues by giving users control over whether they want their information processed locally (immediately) or remotely (via cloud). Furthermore, we are developing a unique solution that combines automated tasks with an intuitive conversational interface, capable of being used by anyone for daily activities.

1. Introduction

Large Language Models (LLMs) have revolutionized natural language processing due to breakthrough architectures like Transformers [2] and scaling laws demonstrated by models such as GPT-3 [1] and BERT [4]. These models are being increasingly applied to tasks such as customer support, education, healthcare, and information retrieval systems; however, traditional chatbot systems rely on single-agent architectures, which can hardly handle complex queries, maintain contextual understanding, and scale efficiently across different domains. Since users need intelligent assistants for natural conversations and the performance of practical tasks, the limitations of these single-agent approaches become particularly problematic when deploying LLMs in a professional setting, which is dominated by task automation, data privacy, and multi-functionality. Retrieval-Augmented Generation addresses factual accuracy by combining traditional information retrieval with LLMs [3, 13], in which relevant documents are retrieved from external knowledge sources to guide the model generation. It reduces hallucination by anchoring answers to curated, up-to-date information. Building upon this, multi-agent frameworks such as MAIN-RAG [6] and AU-RAG [12] demonstrate that distributing tasks across specialized agents significantly improves the precision of retrieval and response reliability. Gao et al. [16] go further by demonstrating that RAG can enable attribution through in-line citation, enabling users to verify the origin of generated content and increasing its trustworthiness. Despite these critical improvements, current systems face pressing limitations since most multi-agent RAG frameworks assume cloud-based processing without addressing privacy issues, while domain-specific models such as Med-PaLM 2 [8] and Galactica [9] and citation-aware systems [16, 17, 18] focus on academic contexts rather than conversational scenarios. We propose an Intelligent Multi-Agent Chatbot with a Privacy-Focused Architecture that will enable both conversational AI and task automation while putting first data security. It introduces a unique dual LLM architecture allowing users to choose between cloud-based processing using OpenAI GPT and local processing using Llama models depending on the sensitivity of data.

We extend this architecture by adding special agents for managing Gmail and performing operations in Google Calendar, enabling task execution in natural language outside the traditional question-answering paradigm, while incorporating RAG-based document understanding for contextual conversations based on user-uploaded files in order to improve response accuracy and relevance of professionals managing day-to-day workflows with security standards.

2. Literature review

A framework aimed at improving the transparency and interpretability of large language model reasoning called CoTAR introduced by Berchansky et al. The better understanding of how intermediate reasoning steps are formed highlighted by this approach and it also focuses on attributing generated chain-of-thought reasoning to specific evidence sources at multiple levels of granularity, due to this CoTAR enhances trust and accountability in complex reasoning tasks. The results demonstrated that the framework effectively improves reasoning attribution accuracy [5]. MAIN-RAG, a multi-agent retrieval-augmented generation framework designed to improve response quality by filtering and refining retrieved knowledge through agent collaboration proposed by Chang et al. MAIN-RAG employs multiple agents to

assess relevance, eliminate noise, and validate retrieved documents before generation. The study demonstrates that multi-agent collaboration significantly improves retrieval precision and response reliability in complex information-seeking and reasoning-intensive tasks [6]. RAGAs, an automated evaluation framework specifically designed for Retrieval-Augmented Generation systems introduced by Shahul Es et al. Standardized metrics to assess key RAG components that includes retrieval relevance, faithfulness, answer correctness, and contextual grounding are proposed by this approach. RAGAs automates the evaluation process and reduces reliance on costly human assessments while providing consistent and interpretable performance measurements [7]. Karan Singhal, Tao Tu, and Juraj Gottweis are the authors of this model, which has strong potential in Medical Science by answering questions. Med-PaLM 2 improves the accuracy of reasoning in the medical domain. This model performance is at an expert level in medical benchmarks (USMLE, Med QA). Chain of Thought reasoning enhances the quality of decision making. This model implements multiple-choice questions and answers long-form clinical questions. Some limitations in this model are clinical reliability and development risks. The authors' vision is scope in scalable medical AI systems [8]. Ross Taylor, Marcin Kardas. Model has the Designed Specifically to Understand and Organize with Scientific Knowledge. In this Model trained 48+ Millions Scientific Documents. Also trained large corpus Scientific document (papers, textbooks, and database), and implement only Retrieval information, and support Multiple Scientific Models. In this model doing Complex Tasks and performs like Scientific Summaries, assist, and technical answering. Purpose of the author is how Researchers access and interact with Knowledge [9]. Tainyi Zhang (2024) studied how well large language models can summarize news articles. Their study shows that LLMs can generate summaries that are fluent and readable, they often include factual errors or add information that is not present in the original news articles. By using both automatic evaluation metrics and human judgement to analyze summary quality and found that human evaluation is important for identifying errors such as hallucinations. This study is relevant to LLM-Based projects because it highlights the need for reliable evaluation methods and better techniques to ensure accurate and trustworthy generated content, especially in real-world applications [10]. ATM is a system that improves how AI models generate answers using retrieved information. It uses multiple AI agents that work together and challenge each other to detect errors or weak information during answer generation. This process helps the model handle noisy or misleading data and produce more accurate and reliable responses. Experimental results show that ATM performs better than standard retrieval-based models, especially in complex question-answering tasks [11]. AU-RAG introduces an agent-based approach to Retrieval-Augmented Generation where multiple intelligent agents manage tasks like searching, reasoning, and answer generation. By dividing responsibilities among agents, the system becomes more flexible and works well across different domains. The results show that AU-RAG improves answer quality and adaptability compared to traditional RAG systems [12]. This paper presents the original RAG framework, which combines information retrieval with text generation to handle knowledge-heavy tasks. Instead of relying only on model memory, the system retrieves relevant documents and uses them to generate accurate answers. The study shows that RAG significantly improves performance on tasks requiring factual and up-to-date knowledge [13].

This work proposes a hybrid method that combines document retrieval with answer generation specifically for regulatory and legal texts. The approach ensures that generated answers are aligned with official regulations and source documents. Experimental results demonstrate improved accuracy and reliability in answering complex regulatory questions [14]. LONGAGENT, a multi-agent framework designed to enable effective question answering over extremely long documents containing up to 128k tokens introduced by Zhao et al. The limitations of single-model architectures by distributing document understanding and reasoning tasks across multiple collaborative agents addressed by the system. Each agent focuses on processing specific document segments, extracting relevant information [15].Tianyu Gao, Howard Yen, Jiatong Yu, and Danqi Chen. They present a structured approach allows large language models to generate text with precise citations by merging retrieval-augmented generation. This method allows the model to access relevant documents and attach citations to precise generated statements, improving factual grounding reducing hallucinations. The study Illustrate that citation-based generation increases transparency and Credibility of LLM outputs. However, the authors also identify limitations related to incorrect citation alignment and incomplete coverage of supporting evidence, highlighting the need for more precis citation mechanisms [16]. Chengyu Huang, Zeqiu Wu, Yushi Hu, and Wenya Wang. They proposed a fine-graine reward framework to train language models for more accurate citation generation . Unlike traditional supervise learning, their approach assigns detailed rewards for citation correctness, relevance, and placement within generated text. Experimental results show that this reward-driven method significantly improves citation quality and faithfulness compared to baseline models. This work emphasizes the effectiveness of reinforcement learning techniques in enhancing citation-aware text generation and address some limitations observed in earlier retrieval-based approaches [17]. Jie Huang and Kevin Chang. They focused on the overall importance of citations in developing responsible and accountable large language models. Their work argues that citations serve as a key mechanism for transparency, enabling users to verify information and assess the reliability of

generated content. Rather than proposing a specific training method, the authors discuss design principles and evaluation considerations for citation-aware system. This study positions citations as a foundational element for trustworthy AI system and complements technical advancements in citation generation [18]. Ross Taylor, Ines Besrour, Jingbo He. Model has been focused on Attributed-Question-Answering (AQA).When using Agentic RAG, RAGentA. RAGentA is a multi agent RAG which giving correct and quality answering. Multi agent design to do multiple tasks. Author introduce Synthetic AQA benchmark dataset with 50 QA pairs. Model working actually different Retrieval Strategies. Author future vision improved the Domain and Language expand [19].

3. Comparative Analysis of Existing Approaches

Approach/ Model	Key Features	Domain of Application	Limitations
Standard RAG(Single-Agent)	Its handle retrieval and generations, Simple Architecture and self-correction	Customer supports systems, Educations and e-learning	It's struggle to complex and multi step queries because one mode handle, less depth reasoning, and more prone to hallucinations.
Multi-Agent RAG	It's parallel or collaborative retrieval and reasoning, Higher co-ordination and communication	Research and Knowledge discovery, Health care and legal analysis	Co-ordination between agent has overhead, high computational cost, error from one agent propagate through system.
RAGentA	Used dynamic tasks and retrieval decision, more adaptive static RAG pipeline	Intelligent virtual assistants, Business operations, Autonomous AI systems	Consume more resources due to repeated retrieval and tools, predictable outcome difficult, RAGentA is harder to control and debug.
Self-Reflective RAG	It's evaluated with own answers, reduce hallucinations and improve answer quality	Scientific Research, Compliance and High reliability information services	Increase inference time and cost, excessive reflection can lead diminishing returns without strong external grounding.

4. Identified research gaps

Having analyzed the current research on multi-agent RAG systems and large language models, the following are limitations in the current research that the proposed research addresses:

4.1. Limitations of Single-Agent RAG: The The conventional RAG architectures such as that designed by Lewis et al. [13] rely on a single agent to perform all functions ranging from retrieval, processing, to generation. Although CoTAR [5] enhances attribute resolution and RAGA [7] offers an evaluation metric, all of them rely on single-agent architectures that pose constraints during complex task retrievals that involve more than one method of retrievals. The current work proposes the utilization of multi-agent architectures to process different responsibilities concurrently similar to MAIN-RAG [6] and AU-RAG [12], but for task automation.

4.2. Limited Privacy-Aware Processing: Currently, Although the current state-of-the-art multi-agent systems, such as ATM [11] and LONGAGENT [15] in the banking industry, show enhanced robustness and document processing, all methods lack privacy concerns in processing data in the cloud. Models requiring domain-specific knowledge, such as Med-PaLM 2 [8] and Galactica [9], involve the transfer of sensitive data to remote servers. The proposed two LL&M systems bridge the limitation by incorporating local processing for sensitive data along with cloud processing.

4.3. Lack of Practical Task Integration: Although the work done by citation-aware systems by Gao et al. [16], Huang et al. [17], and attribution principles presented by Huang and Chang [18] aims at enhancing the trust-ability of answers, these works ignore integrating task performance. The above-mentioned systems are best at answering questions but are unable to handle practical tasks, such as mail management or calendar organization. The proposed framework, instead of being limited to info retrieval, also involves specific agents for performing actions related to Gmail and Google calendars along with natural-language interaction.

4.4. Lack of Evaluation in Real World: Although human evaluation is crucial in identifying hallucinations in summaries, as presented in Zhang et al.'s research paper on news summaries [10], because of this deficiency in RAG models, real-world

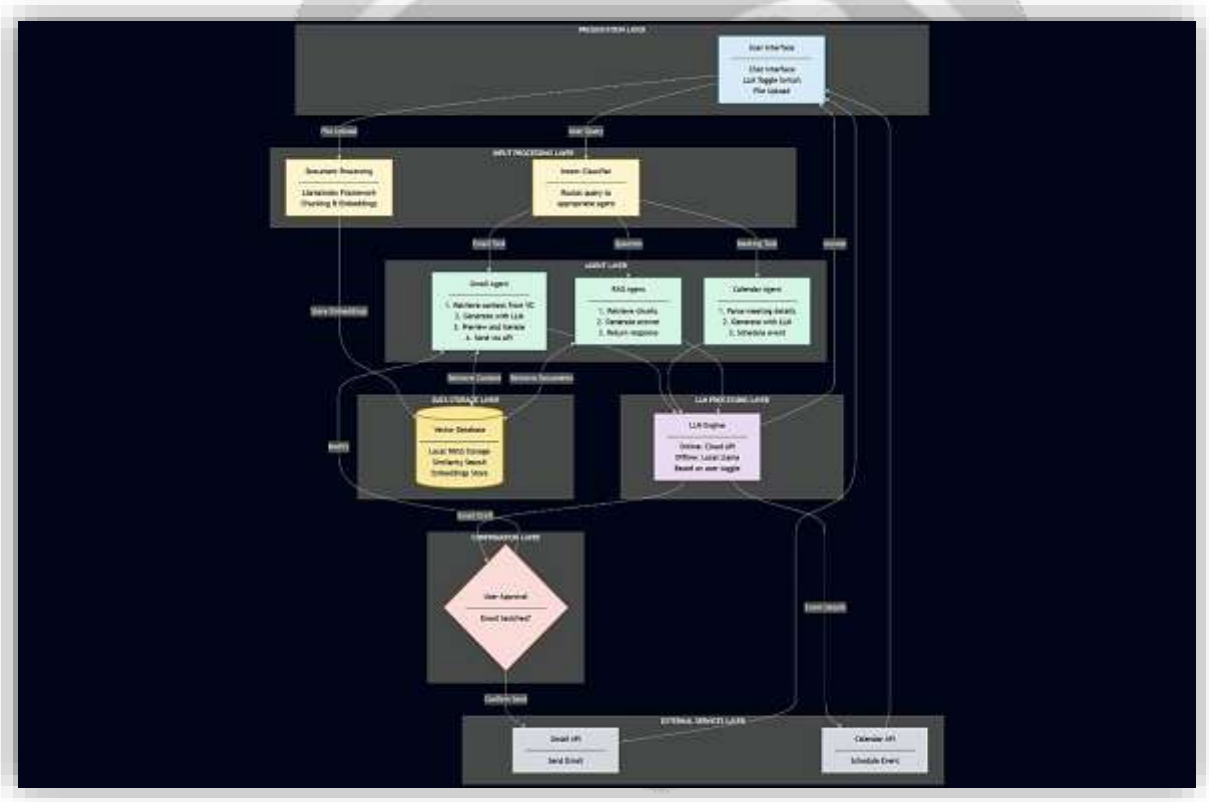
evaluation on user satisfaction, privacy conservation, and efficiency in task automation remains undefined in RAGAs [7]. The design of the Intelligent Multi-Agent Chatbot with Privacy-Focused Architecture fills this gap by incorporating principles of multi-agent collaboration from MAIN-RAG [6] and AU-RAG [12], attribution logic based on principles of CoTAR [5] and citation-aware generation [16, 17, 18] and the newly proposed privacy-preserving method for dual-LLM processing to go beyond the QA system to provide real-world task functionality.

5. Discussion

Through all the literature reviewed till date, we found that, when several agents interact with each other, they tend to produce more reliable results than a single process handling all of the work; i.e., splitting tasks between multiple agents leads to improved accuracy and trustworthiness of answers (e.g., MAIN-RAG and AU-RAG). Many systems still require that all processing occur via a cloud server, thereby creating potential concerns regarding user privacy. By allowing users to be able to select to process their task either on a cloud-based or local-based infrastructure, our project attempts to mitigate privacy issues by giving users control over whether they want their information processed locally (immediately) or remotely (via cloud). Furthermore, we are developing a unique solution that combines automated tasks with an intuitive conversational interface, capable of being used by anyone for daily activities.

6. Proposed methodology

6.1 Architecture



6.2 Methodology

The workflow of our project starts from the chat interface where the user will do everything in one place, like typing messages, uploading files and choosing which model mode they want to use. The chat interface block will have the LLM toggle switch for the offline mode or the online mode, and this toggle will decide which endpoint the user query will go to for LLM processing, because the endpoint for online and offline LLM will be different. When the user sends a query, it first goes to the intent classifier, which will judge what the user wants to do, such as schedule a meeting, generate and send an email, or simply ask a question.

If the user uploads any file, it will go through the RAG data processing pipeline which will use Llama Index as the framework to process the document and generate embeddings. These embeddings will then be stored in an offline vector database using FAISS, which will be on the user's own system so that the data stays local and private. Later, whenever any agent needs context from user documents, it will query this vector database and use similarity search to retrieve the most relevant chunks for that task.

Once the intent classifier understands the user's goal, it will route the query to the correct agent. If the intent is to generate an

email and send it, the query will be sent to the Gmail agent, and if the intent is to schedule a meeting in Google Calendar, the query will be sent to the Calendar agent; if the user simply wants to ask a question, it will be sent to the RAG agent which will use the retrieval and generation pipeline to answer the user's query. In the Gmail agent, while sending the system prompt to the LLM to generate an email for the user query, the agent will first retrieve the relevant data from the vector store using FAISS similarity search so that the email has proper user context. After retrieval, the selected chunks will go to the LLM along with the system prompt and the user query, so that the LLM will generate the email using both the instructions and the stored user data.

The LLM model used for any of these tasks, whether online or offline, will be decided according to the mode that was enabled by the user when the query was sent to the backend, so the same workflow can work for both cloud and local setups. After the email is generated by the LLM, it will not be sent directly; instead, it will be shown to the user for preview, and the system will ask the user if they want to add or change anything. This iteration will go on until the user says that the email is ok to be sent, and only then will the Gmail agent send the final email to the specified user email id using the Gmail API. For calendar tasks, a similar flow will be followed, where the Calendar agent will extract meeting details and use the LLM plus the Calendar API to create the event, while for simple questions the RAG agent will focus on retrieving relevant chunks and using the LLM to produce a grounded answer for the user

7. Expected outcomes

The success of the implementation of the multi-agent chatbot system is expected to bring about a number of practical and measurable outcomes.

7.1. System Performance:

The combination of sparse and dense methods in the hybrid retrieval strategy is anticipated to increase the document recall by around 12-15% in contrast to single retrieval techniques. The multi-agent framework is likely to elevate the veracity of answers by about 10-12%, which implies that the answers will be better supported by the cited documents. The correctness of overall answers should rise by at least 5-10% as a result of the iterative filtering and verification activities done by agents who are professionals in the field.

7.2.User Experience:

By using a privacy-preserving system that is flexible, users will have the opportunity to select the type of processing they want to be done either in the cloud and using OpenAI GPT for general queries or locally and using Llama models for sensitive information. The chatbot will accept natural language commands to perform various tasks handling reading and writing of emails, setting calendar events, and replying to questions from document uploads. The system will provide its users with easy-to-read, reliable answers that have in-line citations allowing users to check where the information came from.

7.3.Task Automation:

The Gmail and Google Calendar agents are going to efficiently perform the routine tasks of email management and scheduling without requiring the users to master difficult commands or interfaces. The RAG feature will facilitate intelligent dialogues based on the user's submitted PDF documents making it an effective tool for professionals who need to extract information from large document collections quickly. Input and output through voice will allow the system to be operated without hands.

7.4.Security and Privacy:

With the dual-LLM architecture, it will be guaranteed that data of sensitive personal nature will be processed entirely.

8. Conclusion

In this work, we introduced a multi-agent RAG frame work designed to enhance the trust worthiness of attributed QA. Our results demonstrate that a hybrid retrieval system combining BM25 and E5 significantly improves Recall@20 (+12.5%) compared to the best single model. It outperforms the standard RAG base line, particularly in faithfulness (+10.7%), showing that providing in-line citations and conducting a second-stage retrieval for answer revision yield more relevant documents and well-grounded answers. While correctness sees only modest gains (+1.1%), our analysis indicates that the second-stage retrieval currently offers limited added value and may benefit from further refinement, such as adding the document filtering by early-stage agents. Finally, we acknowledge that the four-agent design, while effective in boosting correctness, introduces substantial computational overhead. This trade-off high lights the need for future research to balance performance and efficiency in multi-agent RAG systems.

References

[1] Tom B. Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah et.al, 2020. Language Models are Few-Shot Learners. In Advances in Neural Information Processing Systems 33 (NeurIPS 2020). Curran Associates, Inc., Red Hook, NY, USA, 1877–1901. https://arxiv.org/pdf/2005.14165v4

[2] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Łukasz Kaiser, and Illia

Polosukhin. 2017. Attention is All You Need. In Advances in Neural Information Processing Systems 30 (NIPS 2017). Curran Associates, Inc., Red Hook, NY, USA, 5998–6008. https://arxiv.org/pdf/1706.03762

[3] Patrick Lewis, Ethan Perez, Aleksandra Piktus, Fabio Petroni, Vladimir Karpukhin, Naman Goyal, Heinrich Küttler, Mike Lewis, Wen-tau Yih, Tim Rocktäschel, Sebastian Riedel, and Douwe Kiela. 2020. Retrieval-Augmented Generation for Knowledge-Intensive NLP Tasks. In Advances in Neural Information Processing Systems, Vol. 33. Curran Associates, Inc., Red Hook, NY, USA, 9459–9474. https://doi.org/10.5555/3495724.3496517

[4] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. In Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers). ACL, Minneapolis, Minnesota, 4171–4186. https://doi.org/10.18653/v1/N19-1423

[5] Moshe Berchansky, Daniel Fleischer, Moshe Wasserblat, and Peter Izsak. 2024. CoTAR: Chain-of-Thought Attribution Reasoning with Multi-level Granularity. In Findings of the Association for Computational Linguistics: EMNLP 2024. ACL, Miami, Florida, USA, 236–246. https://aclanthology.org/2024.findings-emnlp.13/

[6] Chia-Yuan Chang, Zhimeng Jiang, Vineeth Rakesh, Menghai Pan, Chin Chia Michael Yeh, Guanchu Wang, Mingzhi Hu, Zhichao Xu, Yan Zheng, Ma hashweta Das, and Na Zou. 2024. MAIN-RAG: Multi-Agent Filtering Retrieval Augmented Generation. https://arxiv.org/pdf/2501.00332

[7] Shahul Es, Jithin James, Luis Espinosa Anke and StevenSchockaert.2024. RAGAs: Automated Evaluation of Retrieval Augmented Generation. In Proceedings of the 18th Conference of the European Chapter of the Association for Computational Linguistics: System Demonstrations. ACL, St. Julians, Malta, 150–158. https://aclanthology.org/2024.eacl-demo.16.pdf

[8] Karan Singhal, Tao Tu, Juraj Gottweis, Rory Sayres, Ellery Wulczyn, Mohamed Amin et.al. 2025. Toward expert-level medical question answering with large language models. Nature Medicine 31, 3 (01 Mar 2025), 943–950. https://doi.org/10.1038/s41591-024-03423-7

[9] Ross Taylor, Marcin Kardas, Guillem Cucurull, Thomas Scialom, Anthony Hartshorn, Elvis Saravia, Andrew Poulton, Viktor Kerkez, and Robert Stojnic. 2022. Galactica: A Large Language Model for Science. https://arxiv.org/pdf/2211.09085

[10] Tianyi Zhang, Faisal Ladhak, Esin Durmus, Percy Liang, Kathleen McKeown, and Tatsunori B. Hashimoto. 2024. Benchmarking Large Language Models for News Summarization. Transactions of the Association for Computational Linguistics 12 (2024), 39–57. https://doi.org/10.1162/tacl_a_00632

[11] Junda Zhu, Lingyong Yan, Haibo Shi, Dawei Yin, and Lei Sha. 2024. ATM: Adversarial Tuning Multi-agent System Makes a Robust Retrieval-Augmented Generator. In Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing. ACL, Miami, Florida, USA, 10902–10919. https://aclanthology.org/2024.emnlp-main.610.pdf

[12] Jisoo Jang and Wen-Syan Li. 2024. AU-RAG: Agent-based Universal Retrieval Augmented Generation. In Proceedings of the 2024 Annual International ACM SIGIR Conference on Research and Development in Information Retrieval in the Asia Pacific Region (Tokyo, Japan) (SIGIR-AP 2024). ACM, New York, NY, USA, 2–11. https://doi.org/10.1145/3673791.3698416

[13] Patrick Lewis, Ethan Perez, Aleksandra Piktus, Fabio Petroni, Vladimir Karpukhin, Naman Goyal, Heinrich Küttler, Mike Lewis, Wen-tau Yih, Tim Rocktäschel, Sebastian Riedel, and Douwe Kiela. 2020. Retrieval-Augmented Generation for Knowledge-Intensive NLP Tasks. In Advances in Neural Information Processing Systems, Vol. 33. Curran Associates, Inc., Red Hook, NY, USA, 9459–9474. https://arxiv.org/pdf/2507.09935

[14]Jhon Stewar Rayo Mosquera, Carlos Raúl De La Rosa Peredo, and Mario Garrido Córdoba. 2025. A Hybrid Approach to Information Retrieval and Answer Generation for Regulatory Texts. In Proceedings of the 1st Regulatory NLP Workshop (Reg NLP 2025). ACL, Abu Dhabi, UAE, 31–35. https://aclanthology.org/2025.regnlp-1.5.pdf

[15] Jun Zhao, Can Zu, Xu Hao, Yi Lu, Wei He, Yiwen Ding, Tao Gui, Qi Zhang, and Xuanjing Huang. 2024. LONGAGENT: Achieving Question Answering for 128kToken-Long Documents through Multi Agent Collaboration. In Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing. ACL, Miami, Florida, USA, 16310–16324. https://doi.org/10.18653/v1/2024.emnlp-main.912

[16] Tianyu Gao, Howard Yen, Jiatong Yu, and Danqi Chen. 2023. Enabling Large Language Models to Generate Text with Citations. In Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing. ACL, Singapore, 6465–6488. https://doi.org/10.18653/v1/2023.emnlp-main.398

[17] Chengyu Huang, Zeqiu Wu, Yushi Hu, and Wenya Wang. 2024. Training Language Models to Generate Text with

Citations via Fine-grained Rewards. In Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers). ACL, Bangkok, Thailand, 2926–2949. https://aclanthology.org/2024.acl-long.161.pdf

[18] Jie Huang and Kevin Chang. 2024. Citation: A Key to Building Responsible and Accountable Large Language Models. In Findings of the Association for Computational Linguistics: NAACL 2024. ACL, Mexico City, Mexico, 464–473. https://doi.org/10.18653/v1/2024.findings-naacl.31

[19] Ines Besrour, Tobias Michael Färber Schreieder, Jingbo He ,Michael Färber. RAGentA : Multi-Agent Retrieval-Augmented Generation for Attributed Question Answering 2025, https://arxiv.org/pdf/2506.16988