International Conference on Science, Engineering & Management Trends

International Journal of Science, Engineering and Technology
ISSN: 2348-4098, P-ISSN: 2395-4752

Personal AI Research Agent using RAG and MCP

Mital Kadu[1], Abhilasha Bhagat[2], Vinay Alapure[3], Pushkar Borse[4], Shruti Deshmukh[5], Snehalata Gujar[6]

[1,2]Assistant Professor, Dept. of AI&DS, Dr. DY Patil Institute of Engineering, Management and Research, Maharashtra, Pune, India

[3,4]UG Scholar, Dept. of AI&DS, Dr D.Y Patil Institute of Engineering Management and Research, Maharashtra, India

[5,6]Dept. of AI&DS, Dr D.Y Patil Institute of Engineering Management and Research, Maharashtra, India

Abstract- The rapid expansion of academic literature and digital research resources has significantly increased the complexity of modern research activities. Researchers are required to navigate large volumes of documents across multiple platforms, often leading to fragmented workflows and difficulties in maintaining contextual continuity over extended research periods. While Large Language Models (LLMs) have improved natural language interaction with research content, standalone LLM-based approaches often suffer from limitations such as hallucinations, weak factual grounding, and lack of persistent memory, reducing their effectiveness in document-centric research tasks. Recent advancements have introduced hybrid frameworks that augment LLMs with external retrieval and orchestration mechanisms to address these challenges. This review paper examines key developments in Retrieval-Augmented Generation (RAG) and orchestration frameworks such as the Model Context Protocol (MCP), focusing on their application in AI-assisted research systems. RAG enhances response reliability by grounding language model outputs in relevant source documents, while MCP enables structured coordination between LLMs and external tools, including web search and summarization services. The review also highlights the growing importance of persistent memory mechanisms for supporting long-term research continuity and cumulative knowledge building. By synthesizing findings from recent studies, this paper identifies common architectural patterns and design principles adopted in modern AI-driven research assistants. The analysis discusses the strengths and limitations of existing approaches with respect to retrieval quality, context management, scalability, and security. Additionally, open challenges related to standardized evaluation, long-term memory management, and secure tool orchestration are highlighted. Overall, this review provides a consolidated perspective on current practices and emerging trends in AI- assisted research systems, offering insights that can guide future research and development in this rapidly evolving field.

Keywords- Large Language Models (LLMs), Retrieval-Augmented Generation (RAG), Model Context Protocol (MCP), Research Automation, Vector Databases and Persistent Memory etc.

International Journal of Science, Engineering and Technology
ISSN: 2348-4098, P-ISSN: 2395-4752

I. INTRODUCTION

The rapid growth of scholarly publications, technical reports, and online knowledge resources has made research increasingly complex and data intensive. Researchers today must navigate large collections of documents across multiple platforms, which often leads to fragmented workflows and difficulty in maintaining contextual understanding over time. While digital search engines and reference management tools provide basic support, they offer limited assistance in reasoning across documents or preserving research context during long-term studies. Recent advances in Large Language Models (LLMs) have introduced new possibilities for research assistance through natural language interaction. LLMs can summarize content, answer questions, and support exploratory analysis; however, when used in isolation, they are prone to hallucinations and rely heavily on static training data.

To address these limitations, Retrieval-Augmented Generation (RAG) has emerged as an effective approach that combines language models with external document retrieval, enabling responses to be grounded in relevant source material. As a result, RAG-based systems have gained significant attention in academic and industrial research settings. At the same time, the growing complexity of research workflows has led to increased interest in tool-augmented language models and orchestration frameworks. Mechanisms such as the Model Context Protocol (MCP) aim to enable structured interaction between LLMs and external tools, supporting tasks such as web search, document management, and summarization. While these approaches improve system flexibility and capability, they also introduce challenges related to scalability, evaluation, and security.

This review paper surveys recent developments in Retrieval-Augmented Generation, tool-augmented LLMs, and orchestration frameworks such as MCP. It examines their core concepts, applications, limitations, and emerging research directions, with the goal of providing a clear understanding of how these technologies are shaping modern AI-assisted research workflows.

II. BACKGROUND AND CURRENT STATE OF THE FIELD

The evolution of AI-driven research systems marks a significant shift from static, query-based tools to intelligent, context-aware assistants capable of reasoning and retrieval. Traditional AI models lacked memory persistence, processing each query in isolation and losing continuity between sessions—an essential element for research workflows. The introduction of Retrieval Augmented Generation (RAG) addressed this limitation by grounding responses in external knowledge sources, improving factual accuracy and citation traceability. However, early RAG frameworks were still static and unable to dynamically orchestrate tools or adapt to evolving research contexts, limiting their effectiveness in complex, multi-domain applications.

The emergence of the Model Context Protocol (MCP) brought further advancement by enabling structured coordination between multiple AI tools and agents. MCP allows language models to invoke external utilities—such as summarizers, web searches, and citation checkers— The exponential growth of academic publications, technical reports, and digital knowledge repositories has significantly increased the complexity of modern research workflows. Researchers are often required to explore large volumes of heterogeneous documents across multiple platforms while maintaining contextual understanding over extended periods. Although traditional search engines and reference management tools assist in locating relevant material, they offer limited support for synthesizing information, reasoning across documents, or preserving research context during long-term investigations. These limitations have created a strong demand for intelligent, context-aware research assistance systems.

International Journal of Science, Engineering and Technology
ISSN: 2348-4098, P-ISSN: 2395-4752

To overcome these challenges, Retrieval-Augmented Generation (RAG) has emerged as a prominent paradigm that integrates external document retrieval with generative language models. RAG systems retrieve semantically relevant passages from knowledge sources and use them to ground model responses, thereby improving factual accuracy and traceability. Early RAG architectures demonstrated strong performance in knowledge-intensive NLP tasks, and subsequent studies have extended RAG to scholarly literature exploration, industrial applications, and domain-specific decision support systems. Despite these advances, research has shown that RAG performance is highly dependent on retrieval quality, and errors in retrieval can significantly affect downstream generation.

As RAG-based systems evolved, increasing attention has been paid to evaluation, robustness, and domain adaptation. Recent work has proposed metrics and methodologies to assess retrieval effectiveness and analyze its impact on generative performance, highlighting the need for reliable and interpretable retrieval mechanisms. Other studies have explored robust fine-tuning strategies to enhance the stability of retriever–generator pipelines under noisy or incomplete retrieval conditions. Domain-specific applications, such as manufacturing quality control, further demonstrate the adaptability of RAG while emphasizing the importance of specialized knowledge sources.

Beyond retrieval, the growing complexity of research workflows has driven the development of tool-augmented LLMs, which enable language models to interact with external tools such as search engines, databases, and summarization services. Surveys in this area have categorized frameworks that allow LLMs to dynamically invoke tools based on user intent, extending their capabilities beyond static text generation. While these approaches enhance flexibility and task coverage, they introduce challenges related to tool coordination, prompt scalability, and contextual management.

To address orchestration and context management challenges, recent research has introduced the Model Context Protocol (MCP) as a structured framework for coordinating interactions between LLMs and external tools. MCP enables standardized context exchange, tool invocation, and control flow, supporting more modular and extensible AI systems. Studies analyzing MCP have highlighted both its potential and its limitations, including security concerns, computational overhead, and the lack of standardized evaluation benchmarks. Emerging approaches that integrate RAG with MCP aim to reduce prompt complexity by dynamically retrieving relevant tool contexts and managing interaction flow more efficiently.

Despite significant progress, existing RAG and MCP- based systems largely operate with limited long-term memory and session persistence, restricting their effectiveness in extended research scenarios. Most systems treat interactions as short-lived, failing to preserve evolving research context, intermediate insights, or cumulative knowledge across sessions. Consequently, current research increasingly emphasizes the need for integrated architectures that combine retrieval accuracy, tool orchestration, and persistent contextual memory to support continuous, document-centric research workflows. This evolving landscape reflects a shift toward intelligent research agents capable of acting not merely as conversational interfaces, but as sustained cognitive collaborators in the research process.

III. OBJECTIVES OF THE REVIEW

The objective of this review is to examine recent advancements in Large Language Models (LLMs), Retrieval-Augmented Generation (RAG), and orchestration frameworks such as the Model Context Protocol (MCP) in the context of AI-assisted research systems. The review aims to provide a structured understanding of how these technologies address challenges related to factual grounding, contextual continuity, and tool integration in research workflows. By analyzing existing approaches, this paper seeks to identify common architectural patterns, strengths, and limitations across current systems. In

International Journal of Science, Engineering and Technology
ISSN: 2348-4098, P-ISSN: 2395-4752

addition, the review highlights open challenges related to retrieval quality, scalability, memory management, and security, while synthesizing insights that can guide future research and the development of more reliable and effective AI-based research assistance tools.

IV. LITERATURE REVIEW

Sr. No	Paper Title	Journal	Author & Year	Methodology
1.	Retrieval- Augmented Generation for Knowledge-Intensive NLP Tasks	NeurIPS	Lewis et al., 2020	Introduces RAG by combining dense retrieval with sequence- to-sequence generation for knowledge-grounded NLP.
2.	Retrieval- Augmented Generation for Large Language Models: A Survey	arXiv preprint	Gao et al., 2023	Surveys RAG architecture, retrieval strategies, and evaluation methods for large language models. .
3.	LLM with Tools: A Comprehensiv e Survey	arXiv	Han et al., 2024	Surveys tool- augmented LLMs, including retrieval systems and external APIs.
4.	Evaluating Retrieval Quality in Retrieval- Augmented Generation	CIKM	Zhang et al., 2024	Proposes metrics to evaluate retrieval quality and its impact on RAG performance.
5.	RAG in Manufacturin g Quality Control	Advanced Engineering Informatics	Liu et al., 2025	Applies RAG techniques to manufacturing quality control using domain- specific document retrieval.
6.	RAG4DS: A Lifecycle View of Retrieval- Augmented Generation for Data Spaces	IEEE Access	Gupta & Mehta, 2025	Presents a lifecycle-based framework covering design, deployment RAG systems.
7.	Model Context Protocol (MCP): Landscape, Security Threats, and Future Research Directions	arXiv preprint	Hou et al., 2025	Analyzes MCP architecture, identifies security risks, and outlines future research challenges.
8.	Exploring AI- Driven Approaches for Unstructured Document Analysis and Future Horizons	Journal of Big Data	Mahad -evkar et al., 2024	Reviews AI and NLP techniques for analyzing and extracting information from unstructured documents.
9.	RAG-MCP: Mitigating Prompt Bloat in LLM Tool Selection via Retrieval-Augmented Generation	arXiv preprint	Gan & Sun, 2025	Proposes a RAG-based approach to dynamically retrieve tool contexts and reduce prompt size.
10.	Retrieval- Augmented Generation for Scholarly Literature Exploration	arXiv preprint	Liu et al., 2024	Develops a RAG pipeline for semantic search and contextual generation over scholarly literature.

V. METHODOLOGY OF THE REVIEW

This review focuses on examining recent research developments related to Large Language Models (LLMs), Retrieval-Augmented Generation (RAG), and orchestration frameworks such as the Model Context Protocol (MCP), with particular emphasis on their role in AI-assisted research systems. The discussion is centered on understanding how these technologies are designed, combined, and applied to address challenges such as factual grounding, contextual continuity, tool coordination, and long-term research support. The scope of the review primarily covers studies published between 2023 and 2025 to reflect the most recent advancements in this rapidly evolving area.

The reviewed literature is analyzed in a structured yet descriptive manner to identify common design patterns and architectural trends. Rather than focusing solely on individual models or performance metrics, the review emphasizes how different components— such as retrieval mechanisms, language generation models, orchestration layers, and memory modules— interact within broader research assistance frameworks. To facilitate clarity, existing works are grouped into thematic categories, including retrieval- based generation systems, tool-augmented and multi-agent frameworks, and

International Journal of Science, Engineering and Technology
ISSN: 2348-4098, P-ISSN: 2395-4752

memory-enabled research assistants. This categorization helps highlight similarities and differences in design choices across studies.

A comparative perspective is adopted to examine how these approaches address key aspects such as retrieval effectiveness, orchestration flexibility, and support for persistent context. Based on this synthesis, a reference architectural view is presented to illustrate a commonly observed structure in AI-assisted research systems. This conceptual architecture typically includes a user interaction layer for document upload and querying, a retrieval layer based on RAG for accessing relevant information, a coordination layer that leverages MCP for managing tools and context flow, a memory layer for maintaining long-term research continuity, and a generation layer where LLMs produce context-aware responses.

Overall, this methodology enables a coherent synthesis of existing research without proposing a new system or implementation. By focusing on commonly observed architectural patterns and design choices reported in the literature, the review highlights how components such as language models, retrieval mechanisms, orchestration frameworks, and memory systems are typically integrated. Attention is given to strengths and limitations related to factual grounding, contextual continuity, scalability, and system complexity, rather than relying solely on performance comparisons. This approach consolidates diverse research efforts into a clear overview of current practices in AI-assisted research systems. Additionally, the analysis helps identify gaps in existing work, particularly in areas such as long-term memory management, standardized evaluation, and secure tool orchestration, thereby informing future research directions in AI-driven research assistance technologies.

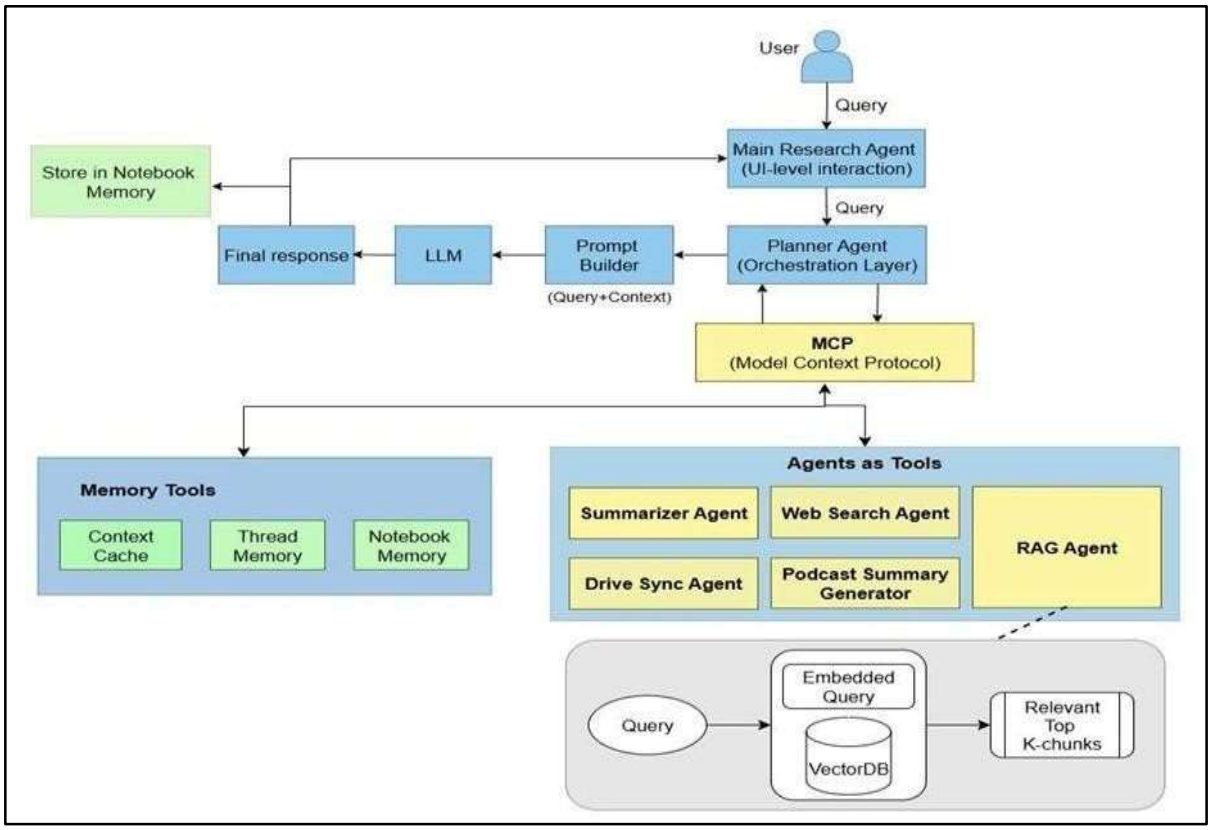

Figure 1: System Architecture

International Journal of Science, Engineering and Technology
ISSN: 2348-4098, P-ISSN: 2395-4752

VI. CRITICAL ANALYSIS & SYNTHESIS- DISCUSSION

The development of AI-driven research systems reflects a remarkable evolution from static question-answering tools to intelligent, context-aware assistants capable of reasoning, retrieval, and orchestration. Traditional AI models, while effective in generating human-like text, suffered from a critical limitation lack of memory persistence. They processed each user query in isolation, disregarding previous context or accumulated knowledge. Consequently, their responses were often repetitive, fragmented, and unsuitable for research environments where continuity and context retention are vital. Retrieval-Augmented Generation (RAG) emerged as a transformative step in addressing these deficiencies by combining language models with external knowledge retrieval.

It improved factual grounding, reduced hallucinations, and enabled traceable outputs; however, RAG-based systems still lacked dynamic adaptability and tool coordination, restricting their effectiveness in multi- domain or evolving research tasks.

The advent of the Model Context Protocol (MCP) significantly advanced this field by enabling structured interaction between multiple AI tools and agents. MCP serves as a communication and orchestration layer that allows LLMs to invoke specialized utilities—such as search, summarization, and citation management— based on user intent. This capability marked a shift from passive information retrieval to active cognitive orchestration, where AI systems can autonomously determine which tools to use and when. Despite this progress, existing MCP-based architectures remain primarily experimental, often constrained by high computational overhead, limited real-time responsiveness, and lack of standardized evaluation benchmarks. Many studies also overlook essential aspects like data security, explainability, and integration with persistent memory—all crucial for academic research and enterprise-level deployments.

A deeper analysis of the existing literature reveals that context continuity and personalization remain underexplored dimensions in current RAG-MCP frameworks. While these systems enhance the accuracy and relevance of responses, they rarely preserve the long-term memory required for ongoing research projects. This shortcoming disrupts the coherence of literature reviews, hypothesis development, and iterative experimentation, forcing researchers to repeatedly reintroduce context. Furthermore, the lack of adaptive retrieval refinement limits the ability of current systems to evolve with user needs or to adapt retrieval strategies over time. The absence of a unified, persistent knowledge base also increases the risk of redundant computations and loss of valuable intermediate insights, making scalability a major challenge for broader implementation.

A synthesis of existing studies shows that long-term context continuity and personalization remain insufficiently addressed in current RAG–MCP frameworks. While these systems improve response accuracy and relevance, most do not preserve research context across sessions, requiring users to repeatedly reintroduce background information and disrupting tasks such as literature review and iterative analysis. Limited support for adaptive retrieval further restricts system evolution, while the absence of unified, persistent knowledge bases raises scalability concerns in real-world research settings.

Overall, the literature indicates that effective AI- assisted research systems require a more holistic integration of retrieval, orchestration, and persistent memory. Although retrieval improves factual grounding and orchestration enhances flexibility, their benefits remain limited without coherent context management. Addressing these gaps is crucial for transitioning from experimental prototypes to reliable, scalable, and user-centric AI-driven research assistance technologies.

VII. FUTURE DIRECTIONS / RESEARCH GAPS

The next generation of intelligent research agents will benefit from the integration of multimodal and cross- domain data understanding. Incorporating text, audio, video, and image-based knowledge will enable AI systems to process and reason over diverse academic materials such as research presentations, scientific figures, and recorded lectures. Additionally, embedding adaptive RAG pipelines that can dynamically refine retrieval strategies based on user intent or past interactions will significantly enhance contextual accuracy. The adoption of federated learning and edge AI architectures also holds promise, allowing decentralized systems to operate on local research data while preserving privacy and reducing dependence on centralized cloud servers. These improvements can transform AI research agents into autonomous collaborators capable of operating efficiently even in bandwidth-limited or privacy- sensitive environments.

Moreover, future systems must address the ethical, interpretability, and trust dimensions of AI- assisted research. Persistent memory and tool orchestration raise legitimate concerns about data governance, bias propagation, and intellectual property protection. Integrating transparent explainability mechanisms, such as citation tracing and confidence scoring, will enhance user trust and accountability. Additionally, building collaborative multi-agent networks where several research agents share structured memory while maintaining user-specific privacy could revolutionize team-based research workflows. Finally, the fusion of knowledge graphs, contextual ontologies, and reinforcement learning-based optimization may enable systems to not only retrieve information but also reason, hypothesize, and self- improve—paving the way for autonomous, adaptive, and ethical AI research ecosystems.

VIII. CONCLUSION

This review paper examined recent research on Large Language Models, Retrieval-Augmented Generation, and orchestration frameworks such as the Model Context Protocol in the context of AI-assisted research systems. The analysis highlights how standalone language models, despite their strong natural language capabilities, face limitations related to factual grounding, contextual continuity, and long- term research support. Approaches that integrate retrieval mechanisms and external tool coordination have emerged as effective solutions to address these challenges.

The review shows that Retrieval-Augmented Generation improves response reliability by grounding outputs in relevant documents, while tool- augmented frameworks and MCP enable more flexible and dynamic research workflows. However, the literature also reveals persistent challenges, including dependency on retrieval quality, increased system complexity, limited standardization in evaluation, and emerging security concerns related to context and tool manipulation. Additionally, long-term memory management remains an underexplored area in many existing systems.

Overall, the findings indicate that effective AI-assisted research systems require a balanced integration of language understanding, retrieval accuracy, orchestration control, and security awareness. While significant progress has been made, further research is needed to address open challenges and bridge the gap between experimental systems and practical research tools. By consolidating current knowledge and identifying research gaps, this review aims to support future work toward more reliable, scalable, and trustworthy AI-driven research assistance technologies.

Acknowledgements
The authors would like to sincerely thank their project guide and faculty members for their guidance, support, and valuable feedback throughout the development of this review paper. Their insights and suggestions were instrumental in improving the clarity and quality of the work. The authors are also

grateful to the department for providing a supportive academic environment and the necessary resources to complete this study. In addition, appreciation is extended to the researchers and scholars whose published work has been reviewed and cited in this paper, as their contributions greatly enriched this review.

REFERENCES

1. Lewis, P., Perez, E., Piktus, A., Petroni, F., Karpukhin, V., Goyal, N., & Stoyanov, Y. (2020). Retrieval-augmented generation for knowledge-intensive NLP tasks. Advances in Neural Information Processing Systems, 33, 9459–9474.
2. Gao, L., Zhang, J., Han, S., & Liu, J. (2023). Retrieval- augmented generation for large language models: A survey. arXiv preprint, arXiv:2312.10997.
3. Han, J., Lee, M., & Cho, K. (2024). LLM with tools: A comprehensive survey. arXiv preprint, arXiv:2409.18807.
4. Zhang, C., Chen, Y., Wang, X., & Li, T. (2024). Evaluating retrieval quality in retrieval-augmented generation. Proceedings of the ACM International Conference on Information and Knowledge Management (CIKM), 201–210.
5. Liu, Y., Chen, H., & Zhao, M. (2025). RAG in manufacturing quality control. Advanced Engineering Informatics, 59, 103007.
6. Gupta, R. K., & Mehta, A. (2025). RAG4DS: A lifecycle view of retrieval-augmented generation for data spaces. IEEE Access, 13, 45102–45115.
7. Hou, X., Zhao, Y., Wang, S., & Wang, H. (2025). Model context protocol (MCP): Landscape, security threats, and future research directions. arXiv preprint, arXiv:2506.13538.
8. Mahadevkar, S. V., Patil, S., Kotecha, K., Soong, L. W., & Choudhury, T. (2024). Exploring AI-driven approaches for unstructured document analysis and future horizons. Journal of Big Data, 11(92), 1–22.
9. Gan, T., & Sun, Q. (2025). RAG-MCP: Mitigating prompt bloat in LLM tool selection via retrieval-augmented generation. arXiv preprint, arXiv:2505.03275.
10. Liu, X., Zhang, Y., & Wang, H. (2024). Retrieval- augmented generation for scholarly literature exploration. arXiv preprint, arXiv:2402.07872.
11. Han, J., Lee, M., & Cho, K. (2025). Sculptor: Empowering LLMs with cognitive agency via active context management. arXiv preprint, arXiv:2508.04664.
12. Tan, D., Wu, S., & Li, Y. (2025). Robust fine-tuning for retrieval-augmented generation. Proceedings of the 2025 ACM Web Conference, 89–97.