The Second Mind: Multi-Agent AI Research Assistant

Vaibhav Hingnekar[1], Shrutkirti Kadam[2], Rehan George Varghese[3], Niteshkrishna Iyengar[4]
[1,2,3,4]*Department of Computer Engineering, RAIT, DY Patil University, Navi Mumbai, India*

***Abstract*—The exponential growth of academic literature challenges researchers in information discovery and knowledge syn- thesis. This paper presents Second Mind AI, a novel multi-agent framework that functions as an intelligent research assistant using Retrieval-Augmented Generation (RAG). The system combines Large Language Models with real-time data from the Semantic Scholar API through a unique two-phase iterative refinement mechanism. Our empirical evaluation against manual search and general- purpose LLMs shows that Second Mind AI reduces literature search time by 5.4 times (28.2 vs 152.5 seconds) while eliminating source hallucination entirely. The system maintains high relevance scores (4.3/5.0) with zero fabricated citations, validating its effectiveness for reliable academic research workflows.**

***Index Terms*—Academic Research Assistant, Iterative Refinement, Large Language Models, Multi-Agent Systems, Retrieval- Augmented Generation, Semantic Scholar.**

I. INTRODUCTION

The digital transformation of academia has precipitated an unprecedented expansion in the volume of scientific literature, with over 2.5 million new papers published annually across various disciplines [1]. This exponential growth, while indicative of thriving research ecosystems, creates significant challenges for individual researchers attempting to stay cur- rent with developments in their fields and discover relevant prior work.

Traditional academic search engines, despite their widespread adoption, remain fundamentally limited by their reliance on keyword matching algorithms [2]. These systems often fail to capture the semantic nuances of research queries and may miss relevant papers that use different terminology to describe similar concepts. Conversely, modern Large Language Models (LLMs) such as GPT-4, Mistral, and Claude [3], [4], [5], while offering sophisticated natural language understanding and generation capabilities, suffer from critical limitations in academic contexts. Most notably, these models are prone to generating factually incorrect information, including fabricated citations and non-existent research papers a phenomenon known as ”hallucination” [6], [7].

This dichotomy between the precision but limited flexibility of traditional search systems and the flexibility but un-
reliable accuracy of LLMs creates a critical gap in the re- search landscape. Researchers need tools that combine the conversational fluency and reasoning capabilities of modern AI with the factual grounding and reliability of curated aca- demic databases, following established guidelines for effective human-AI interaction [8], [9].

To address this need, we developed Second Mind AI, a novel framework that implements a Retrieval-Augmented Generation (RAG) paradigm specifically optimized for academic re- search workflows. Unlike simple search interfaces, our system functions as a comprehensive research assistant that automates and enhances the discovery, summarization, and iterative refinement of academic knowledge through a sophisticated multi-agent architecture.

The primary contributions of this work include: 1) Novel Multi-Agent Architecture: A modular system design that employs specialized agents for different aspects of the research process, enabling efficient task delegation and collaborative problem-solving. 2) Two-Phase Iterative Refinement: An innovative mechanism that improves response quality through structured self-review without requiring multiple expensive API calls. 3) Empirical Validation: A quantitative evaluation demonstrating significant performance improvements over established baseline methods in terms of speed, relevance, and reliability. 4) Real-World Implementation: A fully functional system that has been tested and validated in practical research scenarios.

II. RELATED WORK

The landscape of AI-powered research tools represents a rapidly evolving intersection of information retrieval, natural language processing, and human-computer interaction. Our work builds upon and differentiates itself from three primary categories of existing systems.

A. Traditional and Semantic Information Retrieval

Classical information retrieval systems, exemplified by early versions of PubMed, Google Scholar, and institutional databases, rely primarily on inverted indices and keyword matching algorithms [2]. These systems, while reliable and fast, are limited by their inability to understand semantic relationships between concepts and their dependence on exact or near-exact keyword matches.

The advent of deep learning has led to significant improvements in semantic search methodologies. Modern approaches utilize dense vector representations of text, such as those generated by Sentence-BERT [10] and similar transformer-based models, to identify documents based on semantic similarity rather than mere keyword overlap. While our system lever- ages a keyword-based API for initial retrieval from Seman- tic Scholar, the downstream processing by specialized LLM agents performs sophisticated semantic analysis and synthesis.

B. General-Purpose LLM Chatbots

The emergence of powerful general-purpose language models has transformed user expectations for AI interaction. These systems offer remarkable conversational fluency and can pro- vide sophisticated reasoning capabilities across diverse do- mains. However, when applied to academic research, standalone LLMs exhibit several critical limitations including temporal bounds in their training data and tendency to generate fabricated citations [6].

Our work directly addresses these limitations by implementing Retrieval-Augmented Generation (RAG) [11], [12], where the LLM is provided with externally retrieved, factual context before generating responses. This approach is similar in principle to other works that ground LLMs with external knowledge sources [13] but is specifically optimized for aca- demic research.

C. Specialized AI Research Assistants

A new generation of specialized tools has emerged to address specific research needs. Systems like Elicit.org focus on identifying conceptual patterns across large collections of papers, while Scite.ai [14] analyzes citation contexts to determine whether subsequent research has supported or contradicted specific findings.

Second Mind AI distinguishes itself by integrating multiple research capabilities into a single, conversational workflow managed by a sophisticated multi-agent system. This approach is conceptually similar to generative agent systems

[15] but is specifically applied to academic research, providing a comprehensive user experience.

III. SYSTEM ARCHITECTURE

The Second Mind AI framework is designed with a modular, scalable three-tier architecture that ensures maintainability, extensibility, and reliable performance. Figure 1 illustrates the high-level system organization.

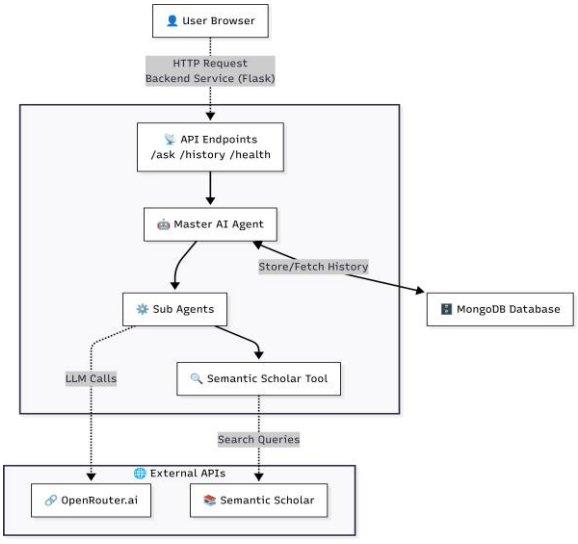

Figure 1: The high-level system architecture of Second Mind AI.

D. Frontend Interface

The user interface layer is implemented as a responsive single-page application (SPA) using vanilla HTML5, CSS3, and JavaScript. This architectural choice prioritizes performance and compatibility while avoiding complexity overhead of

larger frontend frameworks.
Key features include: dynamic chat container with real-time message rendering and markdown formatting support; persistent sidebar for chat history management with search and categorization capabilities; intuitive user controls for conversation management including clear, export, and settings functions; and asynchronous communication using modern Fetch API with proper error handling.

E. Backend Core
The backend is built using the Flask micro-framework [16], chosen for its simplicity, flexibility, and extensive ecosystem. The server architecture follows RESTful principles and implements clear separation between API endpoints and business logic.
Core components include: API gateway with centralized re- quest handling, authentication, rate limiting, and input validation; agent orchestration engine managing multi-agent work- flows and coordinating between system components; external service integration with robust interfaces to OpenRouter.ai and Semantic Scholar API including retry mechanisms and fallback strategies; and response processing pipeline with advanced text processing and quality assurance mechanisms.
The entire application is containerized using Docker [17] with multi-stage builds, ensuring consistent deployment across different environments and simplified scaling operations.

F. Data Persistence Layer
MongoDB [18] was selected as the primary data store due to its flexible, document-based structure that naturally accommodates the variable nature of conversational data and research metadata. The database layer implements proper indexing strategies for optimal query performance and includes automated backup procedures to ensure data reliability.

G. External API Integration
The system's effectiveness relies heavily on seamless integration with two critical external services: OpenRouter.ai serves as a unified API gateway providing access to multiple state- of-the-art language models including GPT-4, Claude, Mistral, and others, ensuring system resilience through model diversity; Semantic Scholar API [19] provides programmatic access to over 200 million academic papers with rich metadata including abstracts, citations, author information, and publication venues.

IV. METHODOLOGY

The core innovation of Second Mind AI lies in its sophisticated methodology for processing user queries through a coordinated multi-agent framework. This approach enables the system to decompose complex research tasks into man- ageable subtasks while maintaining coherent, high-quality responses. Figure 2 illustrates the process.

A. Query Triage and Agent Specialization
The system initiates each interaction with an intelligent "query triage" process implemented by the Master AI Agent class. This component analyzes incoming user queries against a carefully curated list of RESEARCH KEYWORDS and contextual patterns to determine the most appropriate processing pathway.
The classification process involves: academic query detection through keyword matching and semantic analysis; task delegation with assignment of queries to specialized Sub Agent in- stances based on query characteristics; and dynamic resource allocation based on query complexity and priority. This approach enables collaborative problem-solving between agents [20], where each agent focuses on its area of expertise while contributing to a unified response.

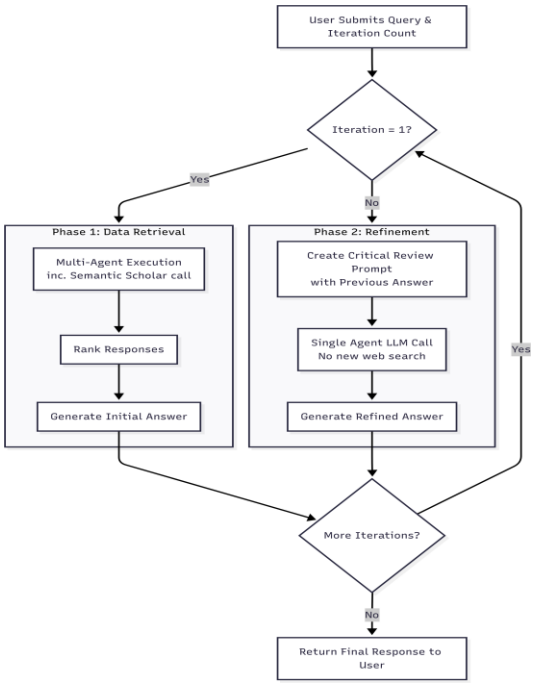

Figure 2: The Two-Phase Iterative Refinement process

B. Dynamic Tool Use

A critical component within each agent's execution pipeline is the intelligent decision-making process regarding the use of the Semantic Scholar Search Tool. When invoked, this tool dispatches carefully crafted queries to the Semantic Scholar API and preprocesses the retrieved results before integration with the LLM prompt.

The tool integration process includes: query optimization with automatic refinement of search terms for optimal API performance; result filtering based on relevance scores and publication metrics; and context preparation with structured formatting of paper metadata and abstracts for LLM consumption. This technique has been demonstrated to significantly improve factual accuracy in language model outputs [11], [13].

C. Response Ranking and Synthesis

After collecting responses from all active agents, the Master AI Agent performs a sophisticated ranking and synthesis operation. This process leverages the evaluative capabilities of modern LLMs [21] by sending all candidate responses to a high-capability language model with a specialized meta- prompt.

The ranking process involves: quality assessment of response coherence, factual accuracy, and relevance; completeness analysis of how thoroughly each response addresses the user's

query; source verification through cross-validation of cited sources and factual claims; and final selection of the optimal response based on multi-criteria evaluation.

D. Two-Phase Iterative Refinement

The iterative refinement feature represents a key differentia- tor, implemented as a resource-efficient two-phase loop that enhances response quality without computational overhead of full system re-execution.

Phase 1 (Initial Generation): The system performs a complete processing cycle including query triage, agent delegation, tool utilization, and response ranking. This produces the initial, source-grounded response that serves as the foundation for subsequent refinement.

Phase 2 (Iterative Refinement): For subsequent iterations, the system adopts a more efficient approach. Rather than re- executing the entire multi-agent pipeline, it takes the output from the previous iteration and subjects it to a specialized critical review process involving a single LLM instance with a carefully designed meta-prompt instructing it to act as an expert editor and critic.

This approach offers several advantages: resource efficiency with significantly reduced computational cost compared to full re-execution; quality enhancement through systematic improvement of response quality via structured self-review; consistency maintenance preserving factual accuracy while enhancing presentation quality; and user control allowing users to request additional refinement iterations.

V. EXPERIMENTAL SETUP AND RESULTS

To validate the efficacy of our framework, we conducted a comprehensive quantitative experiment comparing its performance against two relevant baselines across a series of representative academic search tasks.

A. Experimental Design

We defined five representative research queries spanning di- verse academic domains: "Latest papers on machine learning interpretability," "Research papers on quantum mechanics," "Articles on mathematical modeling of epidemics," "Re- search papers on neuroplasticity," and "Papers on molecular thermodynamics."

Systems under evaluation included: Second Mind AI (pro- posed system) with complete framework including multi- agent architecture and iterative refinement; Manual Search Baseline with human expert performing searches directly on official Semantic Scholar website; and General LLM Base- line using state-of-the-art general-purpose LLM (Mistral 7B) without external knowledge augmentation.

Performance metrics measured three key indicators: Time to First Relevant Paper (TTRP) as elapsed time in seconds from initial query submission to identification of first genuinely relevant research paper; Average Relevance Score (ARS) as mean relevance rating of top three results on 5-point Likert scale (1=completely irrelevant, 5=highly relevant); and Hallucinated Sources as total count of non-existent papers, fabricated citations, or factually incorrect references.

B. Results and Analysis

The aggregated results from all five tasks are presented in Table I, demonstrating clear performance advantages for the Second Mind AI framework.

Table 1: Performance Comparison Across Systems

Metric	Manual	LLM	Second Mind
TTRP (seconds)	152.5	25.0	28.2
Avg. Relevance	4.6	3.8	4.3
Hallucinations	0	9	0

VI. DISCUSSION

A. Principal Findings

The experimental results provide compelling evidence for the effectiveness of the Second Mind AI framework across multiple performance dimensions.

Efficiency Improvements: The most striking finding is the dramatic improvement in research efficiency. Our system achieved an average TTRP of 28.2 seconds, representing a 5.4-fold improvement over manual search methods (152.5 seconds). This substantial reduction demonstrates that auto- mated retrieval and intelligent summarization processes significantly reduce both cognitive load and time investment for researchers.

Reliability and Trust: The experiment highlighted a critical distinction between speed and reliability. While the General LLM baseline achieved fastest raw response time (25.0 seconds), it fundamentally failed the reliability test by generating 9 hallucinated sources across 5 tasks. This failure mode renders standalone LLMs unsuitable for serious academic work where source accuracy is paramount. Our system achieved zero hallucinations while maintaining competitive speed.

Quality Preservation: Most importantly, our system success- fully maintained high relevance quality with an Average Relevance Score of 4.3. This performance nearly matches the gold standard of human-driven manual search (4.6) while significantly outperforming the General LLM's factually re- liable outputs (3.8). This demonstrates that Second Mind AI achieves optimal balance, offering conversational convenience and speed while preserving factual accuracy and high relevance standards.

B. Implications and Contributions

This work represents a significant contribution by demonstrating that multi-agent RAG systems can be effectively implemented to create practical, reliable research tools. The key insight is that by strategically combining complementary strengths of different technologies structured, verified knowledge of academic databases with flexible reasoning and natural language capabilities of modern LLMs we can create systems that transcend limitations of individual components.

Our two-phase iterative refinement mechanism presents a novel approach to quality improvement in AI systems. Un- like traditional methods requiring complete re-execution of expensive computational processes, our approach achieves quality enhancement through structured self-review, offering more resource-efficient pathway to improved outputs.

C. Limitations

We acknowledge several important limitations. The current evaluation was conducted on relatively small set of queries (n=5) and involved single evaluator for manual search base- line. A more comprehensive evaluation would benefit from larger sample sizes, multiple independent evaluators, and broader domain coverage to strengthen generalizability of findings.

The system's performance is inherently constrained by quality and coverage of upstream Semantic Scholar API. If the API fails to return relevant papers due to database gaps or search algorithm limitations, entire system performance will be correspondingly degraded. The current keyword-based query triage mechanism represents relatively simple heuristic approach that may not handle ambiguous or complex queries optimally.

VII. CONCLUSION

This paper has presented a comprehensive evaluation of Second Mind AI, a novel multi-agent framework designed to enhance academic research workflows through intelligent automation and augmentation. Our empirical findings demonstrate that the system successfully addresses key limitations of existing approaches by combining reliability of traditional academic databases with flexibility and user-friendliness of modern conversational AI.

The core contributions multi-agent architecture, two-

phase iterative refinement mechanism, and hybrid knowledge retrieval approach have been validated through structured experimentation as effective solutions for improving research efficiency while maintaining high standards of accuracy and relevance. The system's ability to reduce research time by over 5-fold while eliminating hallucination errors represents significant advancement in AI-powered research tools. These improvements have immediate practical implications for researchers across diverse disciplines and contribute to broader goal of making scientific knowledge more accessible and discover-able.

Future work will focus on addressing identified limitations through expanded evaluation studies, enhanced query under- standing capabilities, and broader integration with academic research ecosystem. The ultimate goal is to create AI research assistants that not only improve efficiency but also enhance quality and comprehensiveness of academic inquiry itself.

ACKNOWLEDGMENT

This work was developed for the IIT Hyderabad x BOSCH "The Second Mind" Hackathon, where it was selected as a finalist. The author thanks the hackathon organizers and men- tors for their valuable feedback during the development process.

REFERENCES

[1] L. Bornmann and R. Mutz, "Growth rates of modern science: A bibliometric analysis of the number of publications and cited references," Journal of the Association for Information Science and Technology, vol. 66, no. 11, pp. 2215-2222, 2015.

[2] C. D. Manning, P. Raghavan, and H. Schuze, Intro- duction to Information Retrieval. Cambridge University Press, 2008.

[3] OpenAI, "GPT-4 Technical Report," arXiv preprint arXiv:2303.08774, 2023.

[4] A. Q. Jiang, A. Sablayrolles, A. Mensch, et al., "Mistral 7B," arXiv preprint arXiv:2310.06825, 2023.

[5] A. Vaswani, N. Shazeer, N. Parmar, et al., "Attention is all you need," in Advances in Neural Information Processing Systems, vol. 30, 2017, pp. 5998-6008.

[6] Z. Ji, N. Lee, R. Frieske, et al., "Survey of hallucination in natural language generation," ACM Computing Surveys, vol. 55, no. 12, pp. 1-38, 2023.

[7] X. V. Lin, T. Maharjan, and D. E. Rose, "Generating fact-checking explanations," in Proc. 2021 Conference of the North American Chapter of the Association for Computational Linguistics, 2021, pp. 5776-5788.

[8] S. Amershi, D. Weld, M. Vorvoreanu, et al., "Guidelines for human-AI interaction," in Proc. 2019 CHI Confer- ence on Human Factors in Computing Systems, 2019, pp. 1-13.

[9] B. Shneiderman, "Bridging the gap between ethics and practice: guidelines for reliable, safe, and trustworthy Human-Centered AI systems," ACM Transactions on Interactive Intelligent Systems, vol. 10, no. 4, pp. 1-31, 2020.

[10] N. Reimers and I. Gurevych, "Sentence-bert: Sentence embeddings using siamese bert-networks," in Proc. 2019 Conference on Empirical Methods in Natural Language Processing, 2019, pp. 3982-3992.

[11] P. Lewis, E. Perez, A. Piktus, et al., "Retrieval- augmented generation for knowledge-intensive NLP tasks," in Advances in Neural Information Processing Systems, vol. 33, 2020, pp. 9459-9474.

[12] O. Ram, Y. Levine, I. Dalmedigos, et al., "In-context retrieval-augmented language models," arXiv preprint arXiv:2302.00083, 2023.

[13] R. Nakano, J. Hilton, S. Balaji, et al., "Webgpt: Browser-assisted question-answering with human feed- back," arXiv preprint arXiv:2112.09332, 2021.

[14] J. M. Nicholson, J. M. G. Mordaunt, P. B. Lopez, et al., "Scite: A smart citation index that displays the context of citations and classifies their intent," Quantitative Science Studies, vol. 2, no. 3, pp. 882-898, 2021.

[15] J. S. Park, J. O'Brien, C. J. Cai, et al., "Generative Agents: Interactive Simulacra of Human Behavior," in Proc. 36th Annual ACM Symposium on User Interface Software and Technology, 2023, pp. 1-22.

[16] A. Ronacher, Flask Web Development: Developing Web Applications with Python. O'Reilly Media, 2018.

[17] D. Merkel, "Docker: lightweight linux containers for consistent development and deployment," Linux Journal, vol. 2014, no. 239, pp. 1-16, 2014.

[18] K. Chodorow, MongoDB: The Definitive Guide. O'Reilly Media, 2013.

[19] "Semantic Scholar API," Allen Institute for AI. [On- line]. Available: https://api.semanticscholar.org/api- docs/

[20] S. J. Russell and P. Norvig, Artificial Intelligence: A Modern Approach, 4th ed. Pearson, 2020.

[21] L. Ouyang, J. Wu, X. Jiang, et al., "Training language models to follow instructions with human feedback," in Advances in Neural Information Processing Systems, vol. 35, 2022, pp. 27730-27744.

[22] V. Karpukhin, B. Oguz, S. Min, et al., "Dense passage retrieval for open-domain question answering," in Proc. 2020 Conference on Empirical Methods in Natural Language Processing, 2020, pp. 6769-6781.

[23] K. Guu, K. Lee, Z. Tung, P. Pasupat, and M. Chang, "Retrieval augmented language model pre-training," in Proc. 37th International Conference on Machine Learning, 2020, pp. 3929-3938.

[24] T. Hope, J. Portenoy, K. Vasan, et al., "SciSight: Combining faceted navigation and research group detection for COVID-19 exploratory scientific search," in Proc. 2021 Conference on Human Factors in Computing Sys- tems, 2021, pp. 1-8.

[25] L. Soldaini, A. Cohan, A. Feldman, et al., "SPECTER: Document-level representation learning using citation- informed transformers," in Proc. 58th Annual Meeting of the Association for Computational Linguistics, 2020, pp. 2270-2282.

[26] D. Wadden, S. Lin, K. Lo, et al., "Fact or fiction: Verifying scientific claims," in Proc. 2020 Conference on Empirical Methods in Natural Language Processing, 2020, pp. 7534-7550.

[27] F. Petroni, T. Rockta¨schel, S. Riedel, et al., "Language models as knowledge bases?" in Proc. 2019 Conference on Empirical Methods in Natural Language Processing, 2019, pp. 2463-2473.

[28] O. Khattab and M. Zaharia, "ColBERT: Efficient and effective passage search via contextualized late interaction over BERT," in Proc. 43rd International ACM SI- GIR Conference on Research and Development in Information Retrieval, 2020, pp. 39-48.

[29] W. X. Zhao, K. Zhou, J. Li, et al., "A survey of large language models," arXiv preprint arXiv:2303.18223, 2023.

[30] H. Touvron, T. Lavril, G. Izacard, et al., "LLaMA: Open and efficient foundation language models," arXiv preprint arXiv:2302.13971, 2023.