iJRASET
International Journal For Research in Applied Science and Engineering Technology

INTERNATIONAL JOURNAL FOR RESEARCH
IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY

Volume: 14 Issue: VII Month of publication: July 2026

DOI: https://doi.org/10.22214/ijraset.2026.84169

www.ijraset.com

Call: 08813907089 | E-mail ID: ijraset@gmail.com

International Journal for Research in Applied Science & Engineering Technology (IJRASET)
ISSN: 2321-9653; IC Value: 45.98; SJ Impact Factor: 7.538
Volume 14 Issue VII July 2026- Available at www.ijraset.com

Intelligent Multi-Agent System for Research Automation

Siddhant Lohakar[1], Annlin Mathew[2], Atharva Kamde[3], Sakshi Bondre[4], Prof. Dhirajkumar Gupta[5] (Guide)
[1, 2, 3, 4]VI Semester, Computer Science and Engineering
[5]Assistant Professor
ST. Vincent Pallotti College Of Engineering And Technology
(An Autonomous Institute Affiliated to RTM University, Nagpur)
NAAC Accredited with 'A' Grade
Gavsi Manapur, Wardha Road, Nagpur - 441108

Abstract: This project proposes a domain-aware multi-agent research system that intelligently routes user queries to specialized AI agents across multiple domains such as biomedical, legal, market research, academic research, and technical programming. Traditional large language models (LLMs) often generate hallucinated or unverifiable responses due to lack of domain-specific grounding and validation [3]. To address this limitation, the system introduces an intelligent Query Router that classifies queries using hybrid techniques such as keyword analysis and embedding similarity, assigning confidence scores and selecting the most relevant domain agents [1]. Each domain agent interacts with domain-specific data sources and generates structured, citation-backed responses. A validation layer further verifies outputs to ensure reliability and reduce hallucinations.
The system is implemented using Node.js, LLM APIs, and databases such as MongoDB or PostgreSQL. It will be evaluated using domain-specific benchmarks, demonstrating improved accuracy, efficiency, and reliability compared to traditional single-agent systems [2].

I. INTRODUCTION

The emergence of Large Language Models (LLMs), exemplified by architectures such as GPT-4, has fundamentally transformed the landscape of automated question-answering systems by facilitating contextually nuanced and linguistically sophisticated responses across diverse knowledge domains [19]. These models utilize extensive pre-training on heterogeneous datasets to execute intricate cognitive reasoning tasks and synthesize detailed information with minimal prompting. Nevertheless, a salient scholarly concern remains the propensity of LLMs for "hallucination"—the generation of factually incorrect or unverifiable assertions delivered with high probabilistic confidence. This phenomenon is particularly problematic in high-stakes environments, such as medicine, jurisprudence, and technical research, where empirical accuracy is non-negotiable [3]. The persistent absence of dynamic, real-time knowledge verification further compromises the reliability of standalone LLM outputs.

In response to these diagnostic challenges, Retrieval-Augmented Generation (RAG) has been posited as a seminal framework that grounds generative processes in external, authoritative knowledge repositories. By retrieving and injecting relevant documentary context into the model's prompt, RAG paradigms effectively mitigate the risks of ungrounded synthesis. However, critical analysis of current RAG implementations reveals a predominant focus on single-domain architectures. In modern research environments characterized by fragmented data silos, a monolithic query approach is both computationally inefficient and analytically limited, as it lacks the granularity required to navigate the disparate semantic structures of specialized fields. Contemporary developments in multi-agent orchestration offer a more resilient and scalable alternative through the deployment of autonomous, domain-specific agents. These agents are governed by a centralized intelligent router that utilizes semantic similarity and confidence metrics to direct queries exclusively to relevant specialized units [1], [2]. This selective delegation not only optimizes resource allocation but also enhances the interpretive depth of cross-domain inquiries by synthesizing expert-level outputs from multiple distinct knowledge vectors. The proposed system builds upon these advancements by integrating intelligent query routing, domain-specific agents, and a robust validation layer into a unified architecture. The system not only retrieves and generates responses but also structures them into organized formats containing key insights, supporting evidence, and references. Additionally, the inclusion of a fact-checking mechanism ensures that generated outputs are verified against authoritative sources, thereby improving trustworthiness. By combining modular agent design with intelligent routing and validation, the proposed approach aims to significantly reduce hallucinations and enhance the accuracy, interpretability, and scalability of AI-driven research systems across multiple domains.

II. PROBLEM STATEMENT AND OBJECTIVES

Contemporary information retrieval infrastructures are largely constrained by a reliance on generalized AI architectures that fail to account for the idiosyncratic complexities of specialized domains. Such systems frequently exhibit a lack of domain-specific semantic depth, resulting in the propagation of hallucinations within critical sectors like healthcare and law. Furthermore, the absence of sophisticated query disambiguation leads to suboptimal routing and a subsequent degradation in the relevance and precision of synthesized intelligence.

Therefore, there is a need for an intelligent, scalable, and domain-aware research system that can accurately identify the nature of a query, route it to the most relevant domain-specific agents, and generate structured, validated outputs. Such a system should minimize hallucinations, improve response accuracy, and efficiently handle both single-domain and cross-domain queries while ensuring interpretability and reliability of results.

A. *Problem Statement*

Existing research frameworks predominantly utilize monolithic, general-purpose models, which results in persistent factual inaccuracies, improper query delegation, and a lack of specialized epistemic grounding [3]. Moreover, the exhaustive querying of all available data sources introduces significant latency and computational redundancy [1].

Therefore, there is a need for a domain-aware system that can intelligently route queries, utilize domain-specific knowledge, and provide validated outputs for complex queries.

B. *Objectives*

1) To design a domain-aware multi-agent research system.
2) To implement intelligent query routing using embeddings and keyword analysis.
3) To develop domain-specific agents for multiple knowledge areas.
4) To incorporate a validation layer to reduce hallucinations.
5) To generate structured outputs with citations.
6) To evaluate system performance using benchmarks.

C. *Scope of Work*

The project focuses on designing a multi-agent system with query routing and validation. It includes backend development, API integration, and evaluation. Large-scale deployment and proprietary data integration are considered future work.

III. LITERATURE REVIEW

Recent research highlights advancements in multi-agent systems, routing strategies, and hallucination reduction.

Table 3.1 Summary of Literature Review

Year	Author	Contribution	Gap
2025	Yue et al. [1]	MasRouter for agent routing	Not domain-specific
2025	Wu et al. [2]	RIRS multi-agent routing	No validation
2025	Chen et al. [4]	Medical multi-agent system	Limited domain
2025	Hao et al. [5]	LLM agents in bio domain	Not generalizable
2025	MultiAgentBench [8]	Benchmark system	No architecture
2025	Agentic RAG [9]	Retrieval + agents	No routing
2025	AgentRouter [6]	Graph-based routing	High complexity
2025	RCR Router [7]	Efficient routing	No validation
2024	Chain-of-Agents [10]	Multi-step reasoning	No dynamic routing

Year	Author	Contribution	Gap
2024	LLM Agent Survey [11]	Overview of agents	No implementation
2023	Toolformer [12]	Tool usage	Not multi-agent
2023	ReAct [13]	Reasoning + acting	No domain focus
2023	GPT-4 [14]	Advanced LLM	Hallucination persists
2023	LangChain [15]	Agent framework	No validation
2023	PaperQA [16]	Research QA	Single domain
2023	Hallucination Survey [17]	Study of hallucination	No solution
2022	MRKL [18]	Modular AI systems	No routing
2021	Domain QA [19]	Domain QA systems	Not scalable
RAG [20]	RAG	Retrieval generation	Single domain
2020	GPT-3 [21]	Few-shot learning	Low reliability

Summary: Existing systems lack a unified approach combining domain-aware routing, structured output, and validation. The proposed system addresses these limitations.

IV. PROPOSED METHODOLOGY

The proposed methodology is designed as a modular multi-agent architecture consisting of query routing, domain-specific processing, validation, and structured response generation. The workflow is divided into multiple stages to ensure scalability, reliability, and accuracy of outputs.

A. Query Processing And Routing

The system begins with user query input, which is preprocessed using natural language processing techniques such as tokenization and normalization. The Query Router converts the input query into an embedding representation using pre-trained embedding models. A hybrid routing mechanism is applied, combining keyword-based classification and embedding similarity. The router compares the query embedding with domain-specific embeddings maintained for each agent and assigns a confidence score. Based on a predefined threshold, the query is routed to one or more relevant agents. This approach avoids unnecessary computation and improves routing efficiency [1], [2].

B. Domain-Specific Agent Processing

Each domain agent is responsible for handling queries within a specific knowledge area. The agents operate independently and are designed with domain-specific retrieval and reasoning capabilities:

- Biomedical Agent retrieves data from PubMed and related APIs, ensuring evidence-based responses.
- Legal Agent processes legal queries using case law databases and verifies citations.
- Market Research Agent gathers financial and news data from external APIs.
- Academic Agent retrieves research papers from sources like arXiv and Semantic Scholar.
- Technical Agent accesses programming documentation and code repositories.

Each agent follows a three-step workflow:

- Retrieval of relevant documents
- Generation of response using LLM APIs
- Post-processing and formatting into structured output

C. Fact-Checking And Validation Layer

To reduce hallucinations, a validation layer is implemented. This layer verifies generated responses by cross-checking them with external sources. It evaluates:

- Citation authenticity
- Source reliability
- Recency of information

If inconsistencies are detected, the system either flags the response or triggers re-evaluation. This improves trustworthiness and aligns with recent research on hallucination reduction [3], [4].

D. Response Aggregation And Structuring

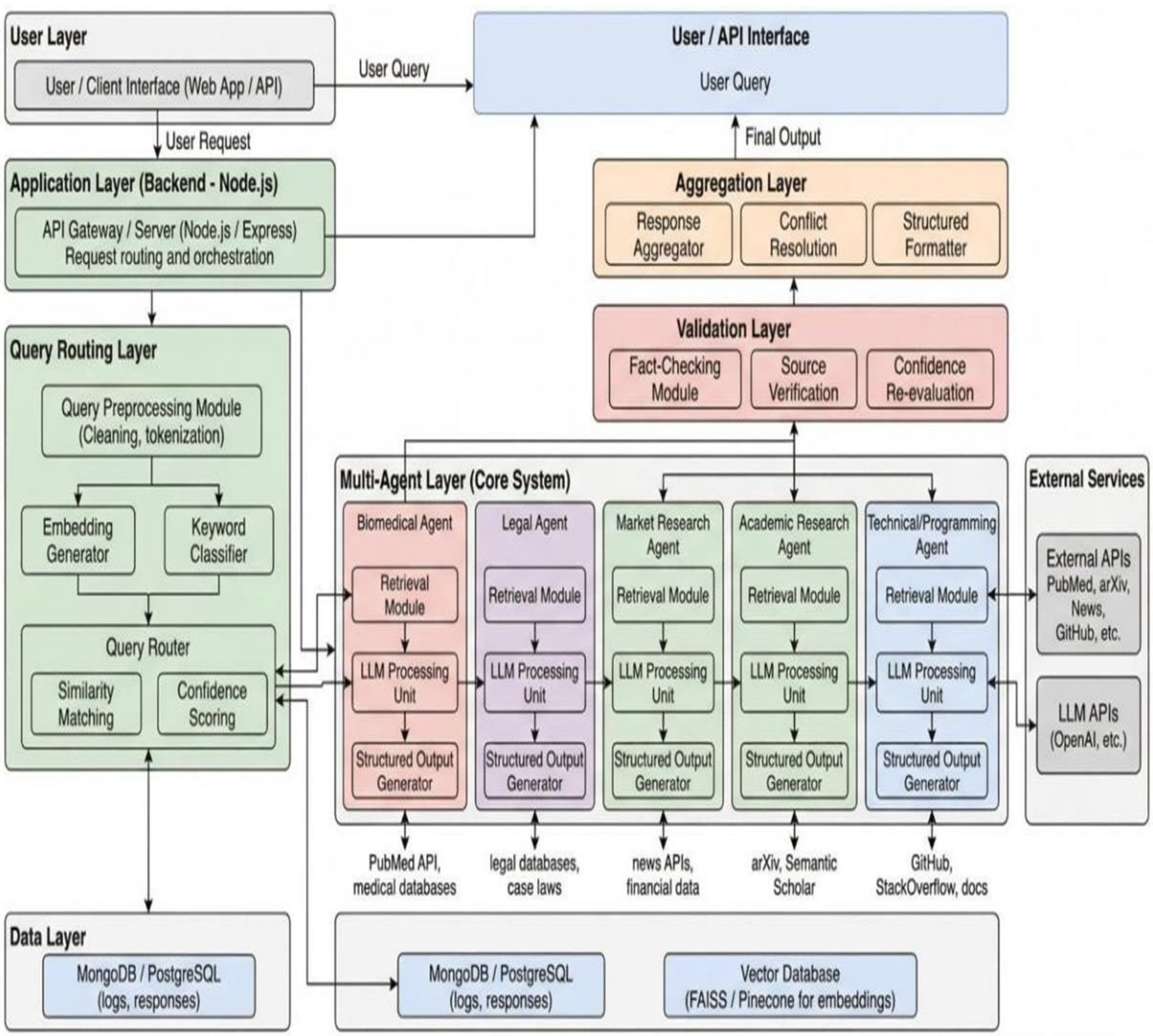
The responses from different agents are aggregated into a unified structured output. The output format includes:

- Summary
- Key findings
- Domain-specific insights
- References
- Confidence score

For cross-domain queries, outputs from multiple agents are merged coherently while maintaining clarity and traceability.

E. System Implementation

The system will be implemented using Node.js as the backend framework. APIs will be used for LLM interaction and data retrieval. MongoDB/PostgreSQL will store intermediate data and logs. The modular design ensures scalability and allows easy addition of new domain agents.



V. SOFTWARE AND HARDWARE REQUIREMENT

A. *Software Requirement*

Component	Specification / Use
Operating System	Linux (Ubuntu) or Windows
Programming Language	JavaScript (Node.js), React JS.
Backend Framework	Express.js
LLM APIs	Open AI, Gemini or Anthropic
Development Tools	VS code, Git
Database	MongoDB / PostgreSQL
Vector Database	FAISS/Pinecone
Libraries	LangChain, Axios, NLP libraries
External APIs	PubMed, arXiv, News APIs, Github

B. *Hardware Requirement*

Hardware	Need	Use in Project
System	16 GB RAM minimum	Development and execution
CPU	Multi-core processor	Backend processing.
Storage	100GB+	Data Storage
Network	High-speed internet	API communication

VI. EXPECTED OUTCOME

The expected outcome of the proposed system is the development of a fully functional domain-aware multi-agent research platform capable of handling complex queries across multiple domains with improved accuracy and reliability.

The system is expected to demonstrate efficient query routing, ensuring that queries are directed only to relevant domain agents. This will significantly reduce computational overhead and response latency compared to traditional systems.

Each domain agent is expected to produce structured, citation-backed responses, improving interpretability and usability. The inclusion of a validation layer is anticipated to reduce hallucinations and enhance the credibility of generated outputs.

The system will be evaluated using domain-specific benchmarks such as biomedical QA datasets, legal case analysis datasets, and programming evaluation tasks. Performance metrics will include accuracy, response time, citation correctness, and routing precision.

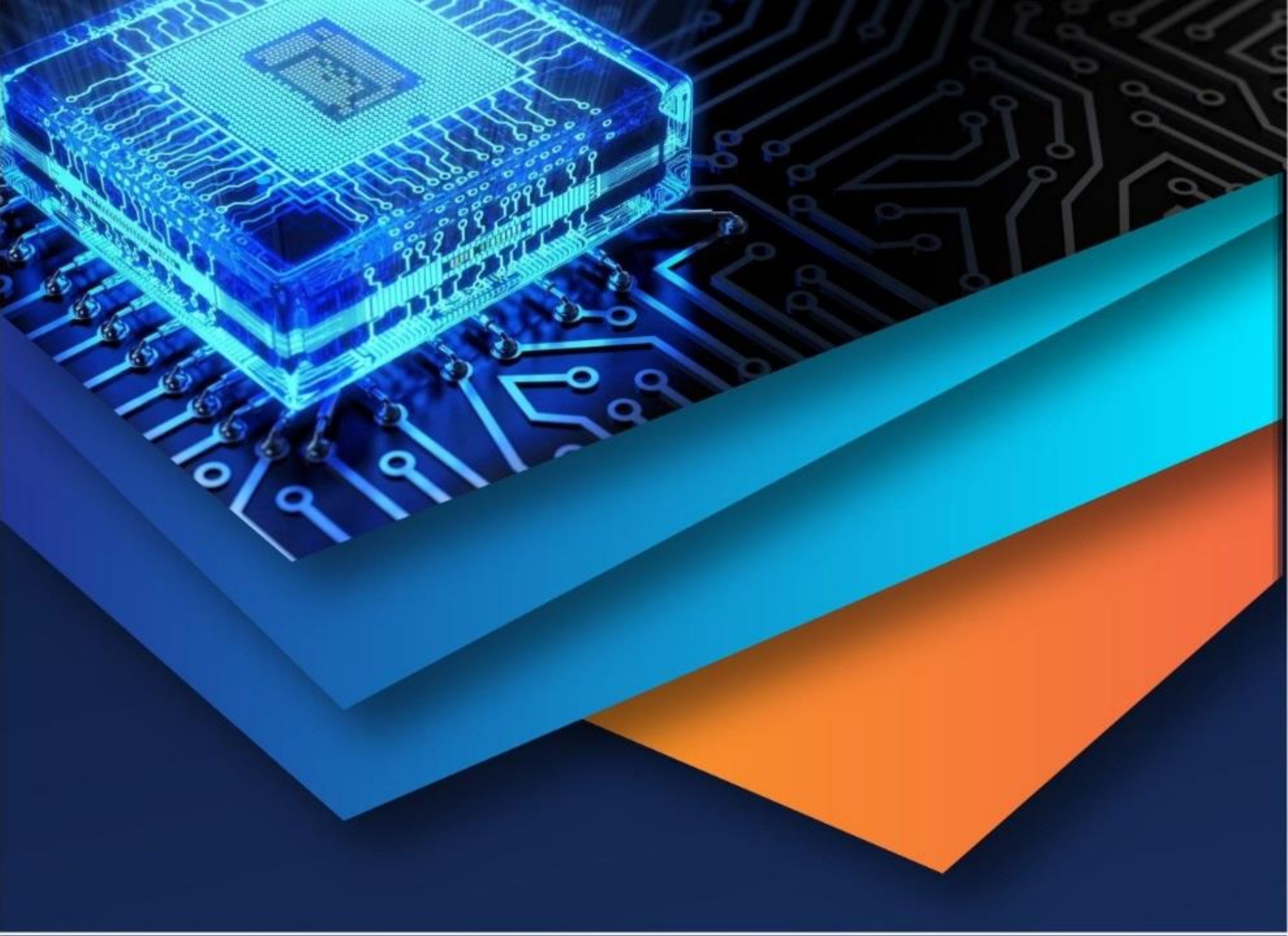
Additionally, the system is expected to handle cross-domain queries by integrating outputs from multiple agents, demonstrating collaborative reasoning capabilities.

Overall, the project aims to establish a scalable, modular, and reliable AI-driven research system that improves upon traditional single-model approaches and contributes to the advancement of multi-agent AI systems [5], [6].

REFERENCES

[1] Yue et al.; MasRouter; Proceedings of the Annual Meeting of the Association for Computational Linguistics (ACL), 2025.
[2] Wu et al.; RIRS Multi-Agent QA; arXiv preprint arXiv:2402.01680 (2025).
[3] Sadat et al.; DELUCIONQA; Findings of the Association for Computational Linguistics (ACL), 2023.
[4] Chen et al.; Multi-Agent Medical LLMs; NPJ Digital Medicine (Nature), 2025.
[5] Hao et al.; LLM Agents in Bioinformatics; Briefings in Bioinformatics, 26(6), bbaf601 (2025).
[6] AgentRouter; arXiv preprint arXiv:2510.05445 (2025).
[7] RCR Router; arXiv preprint arXiv:2508.04903 (2025).
[8] MultiAgentBench; arXiv preprint arXiv:2503.01935 (2025).
[9] Agentic Retrieval-Augmented Generation (Agentic RAG); arXiv preprint arXiv:2501.09136 (2025).
[10] Chain-of-Agents; arXiv preprint arXiv:2406.02818 (2024).
[11] LLM Agent Survey; arXiv preprint arXiv:2402.01680 (2024).

[12] Toolformer; arXiv preprint arXiv:2302.04761 (2023).
[13] ReAct: Synergizing Reasoning and Acting in Language Models; arXiv preprint arXiv:2210.03629 (2022).
[14] GPT-4 Technical Report; arXiv preprint arXiv:2303.08774 (2023).
[15] LangChain Documentation; LangChain Official Documentation, 2023.
[16] PaperQA; arXiv preprint arXiv:2312.07559 (2023).
[17] Hallucination Survey; arXiv preprint arXiv:2304.14779 (2023).
[18] MRKL Systems: A Modular Neuro-Symbolic Architecture; arXiv preprint arXiv:2205.00445 (2022).
[19] Domain Question Answering Systems; IEEE Xplore Digital Library, 2021.
[20] Retrieval-Augmented Generation for Knowledge-Intensive NLP Tasks; arXiv preprint arXiv:2005.11401 (2020).
[21] GPT-3: Language Models are Few-Shot Learners; arXiv preprint arXiv:2005.141 (2020).



10.22214/IJRASET

INDEX COPERNICUS 45.98

IMPACT FACTOR: 7.129

IMPACT FACTOR: 7.429

INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY

Call : 08813907089 (24*7 Support on Whatsapp)