Intelligent Campus Assistant A Hybrid Agentic Retrieval Augmented Generation Architecture for Dynamic Information Retrieval in Higher Education

Siddhant R. Gondane[1], Shaivik S. Shende[2], Anuj N. Pande[3], Anand A. Chaudhari[4], Aryan R. Rathod[5]
Dr. Sunil R. Gupta[6]
[1,2,3,4,6]*Artificial Intelligence and Data Science. Prof. Ram Meghe Institute of Technology and Research Amravati, India*
[5]*Professor, Computer Science and Engineering Prof. Ram Meghe Institute of Technology and Research Amravati, India*

***Abstract*—Implementation of Artificial Intelligence (AI) in terms of higher education administration is a paradigm shift with the stagnant information repositories being replaced by an interactive and conversational-based interface. Nonetheless, current conversational AI applications in the university setting have a bottleneck of freshness. Although traditional Retrieval-Augmented Generation (RAG) is useful in managing high inertia data, e.g., academic regulations and course syllabi, it often misses capturing low inertia real time updates like an impromptu schedule change, an event announcement or emergencies because of the latency incurred during the indexing of vectors and the stochastic nature of semantic retrieval. The concept proposed in this paper is that of the Intelligent Campus Assistant, a new conversational agent which uses a Hybrid Data Architecture and a Tool-Calling Agentic framework. In contrast to weak ReAct-based agents that tend to get stuck in reasoning loops and hallucinate, our system makes use of the deterministic tool-calling properties of modern Large Language Models (in the case Llama 3) to route queries dynamically between a static vector database (ChromaDB) and a dynamic, real time, and controllable by an administrator notice store. We suggest a scalable solution to the problem of the long tail of administrative student queries by decoupling the consumption of long-term knowledge and short-term updates and coordinating them with the help of a logic-aware agent. This paper has described the dual-source retrieval system and how stochastic reasoning is replaced by structured invocation of tools and the consequences of this architecture to institutional resilience, information trustworthiness and administrative efficacy.**

***Index Terms*—Retrieval-Augmented Generation (RAG), Agentic AI, Tool-Calling Framework, Large Language Models (LLMs), Vector Databases (ChromaDB), Real-Time RAG, Semantic Search Hallucination Mitigation, Llama 3.**

I. INTRODUCTION

1.1 The Administrative Crisis in the Modern University Ecosystem

The modern world of higher education is a highly complex, informational system where the speed of data creation and the volume of data may often exceed the ability of traditional distribution channels. This administrative burden is experienced by administrative departments especially ones charged with student affairs, admissions, and examination with an ever-increasing influx of repetitive and low-complexity queries, which have consumed a disproportionate portion of human capital resources which otherwise could be used to carry out high-touch student counseling and strategic planning. These questions are of a bimodal nature. On the one hand, institutions have in their charge a huge bulk of High-Inertia data: academic rules, codes of conduct, fee systems and syllabi. It is semantically rich information that is rarely changed, in most cases every semester. Conversely, every day a campus is run creates an onslaught of Low-Inertia data: room changes, alerts due to emergencies, and event schedules, among others. This information is very volatile and time saved.3

1.2 The Evolution of Conversational AI: From Decision Trees to Agency

Past institutions of higher learning have tried to reduce this load with the help of unresponsive Frequently Asked Questions (FAQ) pages or manually programmed chatbots with decision trees, which resorted to a hard and fast search based on keyword match and did not support any real semantic interpretation. Large Language Models (LLMs) and Transformer architecture [5] were a breakthrough, and the architectures obtained made it possible to generate human-like texts and recognize the nuances of natural language. In order to reduce the knowledge cutoff shortcomings of LLMs, Retrieval-Augmented Generation (RAG) [6] was introduced as the default solution to base responses on external, verifiable data. But the standard RAG architecture considers all the organizational data as a monolithic vector index, which brings the problem of freshness, when re-indexing every small update in the system is computationally expensive and causes latency.8 This requires a transition to Agentic AI in which the system becomes a self-organizing coordinator of tools and data sources.

II. LITERATURE REVIEW

2.1 The History of the Conversational AI in Education.

The search towards machines which are able to carry on natural language conversation has gone through three different epochs. Systems of the Symbolic Era (1960s-1980s) were rule-based and literally lacked any idea of intent. Probabilistic modelling was introduced during the Statistical Era (1990s-2010s), was more flexible, but has a problem with context retention. With the introduction of the Transformer architecture in 2017 [5], the Neural Era was initiated, allowing Conversational AI to complete complex, unstructured queries of students with great accuracy. Nonetheless, the propensity towards the generative models to hallucinate the presence of plausible and yet false facts, however, is one of the crucial issues when it comes to educational administration.[10]

2.2 Retrieval-Augmented Generation (RAG):

Concepts and Dynamic Limitations. RAG was officially suggested by Lewis et al. (2020) to close the parametric memory (model weights) and non-parametric memory (external databases) gap:[6] RAG is the industry standard on knowledge-intensive NLP tasks, and failure modes in dynamic environments are noted in recent literature

- Update Latency: Vector indexing (in particular, HNSW graph updates) is computationally expensive, which causes a bottleneck in freshness of real-time data.[7]
- Semantic Flattening: Vector search is excellent at semantic similarity, but not accurate keyword matching or temporal sequencing (i.e. between Exam Schedule 2024 and Exam Schedule 2025).[12]

2.3 The Common Denominator the Shift to Agentic AI:

ReAct to Tool-Calling. Researchers created Agentic AI in order to go beyond passive RAG. The ReAct (Reason + Act) model enabled LLMs to produce reasoning traces and actions in a feedback loop, but reasoning loops and failures in processing parsing tool outputs are also known to be associated with reasoning generated as unstructured text.8 Tool-Calling paradigm is a solution to this, by refining models to generate structured objects in JSON representing tool execution.[13] This decouples reasoning logic and syntax of execution. The recent benchmark results indicate that the Llama 3 models, fine-tuned to the tool use, are much more successful in ensuring schema compliance and routing accuracy compared to the general-purpose models.[8][3]. Theoretical Framework

The design of the Intelligent Campus Assistant is predicated on the theory that administrative knowledge is bimodal. Treating High-Inertia and Low-Inertia data with a uniform retrieval strategy is suboptimal.

III. THEORETICAL FRAMEWORK

The Intelligent Campus Assistant design is based on the premise; the administrative knowledge is bimodal in nature. One does not treat the High-Inertia and the Low-Inertia data according to the same retrieval strategy.

3.1 Vector Space Models and Semantic Retrieval

In the case of High-Inertia data, a Vector Space Model is used in the system. Documents are broken down and converted into high diameter vectors. To retrieve

conceptually similar documents by retrieval is based on Cosine

$$\text{similarity} = \cos(\theta) = \frac{A \cdot B}{\|A\|\|B\|}$$

Approximate Nearest Neighbor (ANN) search is provided with a complexity of.[1][1] and the vectors indexed with Hierarchical Navigable Small World (HNSW) graphs.

3.2 Theory of Agentic Tool Selection

In the case of Low-Inertia data, semantic search is inadequate in most cases. The system puts retrieval to be handled by an Agent. The LLM is a deterministic classifier in the Tool-Calling framework, and optimizes the likelihood of producing a valid JSON J schema of tool T $P(J_T \backslash Q)$ This direct mapping reduces cognitive load and eliminates the parsing fragility of ReAct.[13]

3.3 Hybrid Retrieval Fusion

In the case of Low-Inertia data, semantic search is inadequate in most cases. The system puts retrieval to be handled by an Agent. The LLM is a deterministic classifier in the Tool-Calling framework, and optimizes the likelihood of producing a valid JSON schema of tool

$$RRF\_Score(d) = \sum_{i=1}^{n} \frac{1}{k + r_i(d)}$$

where k is a constant (typically 60) and $r_i(d)$ is the rank of document d in the i-th retrieval method.

IV. SYSTEM ARCHITECTURE

It is a modular and four-layer system architecture that is developed to be clear and easy to maintain.

4.1 Data Layer: Decoupled Memory Architecture

- Static Knowledge Base (ChromaDB): It is a Long-term Memory. It maintains documents divided in 1000-character blocks having 200-character overlap in order to ensure semantic continuity.[17]
- Dynamic Notice Store (File System): Serves as "Short-Term Memory." It bypasses the latency of vector re-indexing by storing high-velocity updates as timestamped text files, which are read deterministically at query time.[3]

4.2 Orchestration Layer: Tool-Calling Agent

The tools create an orchestration layer that is further broken down into a tool-calling agent. We make use of Llama 3 model of LangChain and Groq. The agent is an agent serving the textual router

1. search_static_knowledge: Vector retrieval from ChromaDB for policy queries.
2. read_latest_notices: Lightweight file system scan for recent updates.The agent analyzes user intent to route queries, preventing static data from obscuring dynamic updates.

4.3 Interaction Layer: Admin Panel and UI

An obtained web interface enables authorized employees to post notices real-time. Since read_latest_notices tool reads which means read directly off the file system, updates are being propagated with practically zero-latency, resulting in "Real-Time RAG".[18]

V. IMPLEMENTATION DETAILS

5.1 Data Ingestion Pipeline

Chunking is done in the static pipeline by using Directory Loader and Recursive Character Text Splitter. All-MiniLM-L6-v2 vectors are produced and saved to disk so that they are not indexed more than once.

5.2 Agent Logic and Prompt engineering.

The agent is built with createtoolcalling agent. The system prompts a strategy of a Negative Constraint: Do not hallucinate. In case tools are not giving information, confess that you do not know. This is essential in ensuring academic integrity.[10]

5.3 Performance Optimization

A global server-side caching approach loads up at start-up the embedding model and database client. This makes the initiation of response close to instant inference as opposed to the time of about 5 seconds and makes the user experience responsive.[17]

VI. EVALUATION METHODOLOGY

We use the multi-dimensional model to the reliability of the assessment system [20]:

- Faithfulness: Determined by means of the RAGAS framework 18, which involves whether responses have sense-supported context.
- Tool Selection Accuracy (TSA): Measures the accuracy of the router at choosing the appropriate modality of data (Static vs. Dynamic).
- End-to-End Latency: Measured based on Time to First Token (TTFT) and total processing time.[15]

VII. DISCUSSION

7.1 The Freshness Advantage of Hybrid Architectures
The Hybrid Data Architecture is a useful solution to the cause of the latency gap in higher education. The system, by turning to the bimodality of institutional data, favours depth (vectors) and freshness (files). A campus-wide emergency can be communicated within the milliseconds making the chatbot a working channel of communication.[3]

7.2 Determinism in Agentic AI
ReAct to Tool-Calling is the required change that production AI needs. ReAct is susceptible to the use of free text API calls that involve hallucinating. Tool-Calling considers tool invocation structured prediction, which leads to greater accuracy and thoroughness to the university governance.[13]

VIII. ETHICAL CONSIDERATIONS AND GOVERNANCE

8.1 Hallucination Mitigation in High-Stakes Environments
Hallucinations can be mitigated in a large-stakes setting with the help of computational models (Witt, 2018). Hallucination is the main obstacle to the application of AI in university administration.[10] By basing the agent on the dual-source retrieval system, we reduce the usage of the parametric memory of the LLM. The Negative Constraint prompt strategy also makes sure that the assistant will rather say he/she does not know instead of lying about the policies.[12]

8.2 Privacy and Data Scrubbing
Although the present Proof-of-Concept does not consider transactional Personally Identifiable Information (PII), a functional system needs to have the most stringent data retention policies. Hybrid architecture is able to locally host dynamic stores so that sensitive notification of the campus can reside within the institution safe infrastructure.[17]

IX. FUTURE DIRECTIONS

The next step in work will be on the Agentic Chunking whereby the system automatically decides whether or not a dynamic notice has to be advanced into the static vector store. Secondly, we will install Multimodal RAG capable of ingesting campus maps and event posters [18], as well as Learning Management System (LMS) APIs capable of making personalized student concierges.

X. CONCLUSION

The Smart Campus is a blueprint that is given by the Intelligent Campus Assistant in a scalable manner. Adding a solid vector search of fixed policies with a nimble file-store of daily updates, all combined by a Tool-Calling Agent, provides the choice between accuracy in the past and relevance in the present. Such an architecture is a major enhancement on responsiveness and reliability used by monolithic RAG systems, and thus students get the right information at the right time.

REFERENCES

[1] Vaswani, A., et al. (2017). "Attention Is All You Need." Advances in Neural Information Processing Systems (NIPS), 30.
[2] Lewis, P., et al. (2020). "Retrieval-Augmented Generation for Knowledge-Intensive NLP Tasks." Advances in Neural Information Processing Systems (NeurIPS), 33.
[3] Yao, S., et al. (2025). "Large language models as autonomous agents: A survey." ACM Computing Surveys.
[4] Yao, S., et al. (2023). "ReAct: Synergizing Reasoning and Acting in Language Models." International Conference on Learning Representations (ICLR).
[5] Okonkwo, C. W., & Ade-Ibijola, A. (2021). "Chatbots applications in education: A systematic

review." Computers and Education: Artificial Intelligence, 2.

[6] Ersoy, P., & Ersahin, M. (2025). "A Comparative Evaluation of RAG Architectures for Cross-Domain LLM Applications." IEEE Access, 13.

[7] Huang, W. J., & Hew, K. F. (2025). "Chatbots and student motivation: a scoping review." International Journal of Educational Technology in Higher Education, 22(1).

[8] Hu, R., et al. (2025). "ICCA-RAG: Intelligent Customs Clearance Assistant Using RAG." IEEE Access, 13.

[9] Ji, Z., et al. (2023). "Survey of Hallucination in Large Language Models." ACM Computing Surveys, 56(2).

[10] Alansari, A., & Luqman, H. (2025). "A Comprehensive Survey of Hallucination in Large Language Models: Causes, Detection, and Mitigation." ACM Computing Surveys.

[11] Schick, T., et al. (2023). "Toolformer: Language Models Can Teach Themselves to Use Tools." Advances in Neural Information Processing Systems (NeurIPS), 36.

[12] Patil, S. G., et al. (2024). "Gorilla: Large Language Model Connected with Massive APIs." Advances in Neural Information Processing Systems (NeurIPS), 37.

[13] Huang, W. J., & Hew, K. F. (2024). "Facilitating online self-regulated learning and social presence using chatbots: Evidence-based design principles." IEEE Transactions on Learning Technologies, 18.

[14] Malkov, Y. A., & Yashunin, D. A. (2018). "Efficient and Robust Approximate Nearest Neighbor Search Using Hierarchical Navigable Small World Graphs." IEEE Transactions on Pattern Analysis and Machine Intelligence, 42(4).

[15] Basu, et al. (2025). "Evaluation and Benchmarking of LLM Agents: A Survey." arXiv preprint.

[16] Es, S., et al. (2024). "RAGAs: Automated Evaluation of Retrieval Augmented Generation." Proceedings of the 18th Conference of the European Chapter of the Association for Computational Linguistics (EACL).

[17] Saad-Falcon, J., et al. (2024). "ARES: An Automated Evaluation Framework for Retrieval-Augmented Generation Systems." Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics (NAACL).

[18] Qin, Y., et al. (2024). "ToolLLM: Facilitating Large Language Models to Master 16000+ Real-world APIs." International Conference on Learning Representations (ICLR).

[19] Mialon, G., et al. (2023). "Augmented Language Models: a Survey." Transactions on Machine Learning Research.

[20] Singh, A., et al. (2024). "Large Language Model-Driven Immersive Agent." Proceedings of the 2024 IEEE World AI IoT Congress (AIIoT).

[21] Shinn, N., et al. (2024). "Reflexion: Language Agents with Verbal Reinforcement Learning." Advances in Neural Information Processing Systems (NeurIPS), 36.

[22] Cormack, G. V., et al. (2009). "Reciprocal Rank Fusion Outperforms Condorcet and Individual Rank Learning Methods." Proceedings of the 32nd International ACM SIGIR Conference on Research and Development in Information Retrieval.

[23] Robertson, S. E., & Zaragoza, H. (2009). "The Probabilistic Relevance Framework: BM25 and Beyond." Foundations and Trends in Information Retrieval, 3(4).

[24] Hew, K. F., & Huang, W. (2021). "Using chatbots in flipped learning online sessions: Perceived usefulness and ease of use." Australasian Journal of Educational Technology, 37(2).

[25] Labadze, L., et al. (2023). "Role of AI Chatbots in education: Systematic Literature Review." International Journal of Educational Technology in Higher Education, 20(1).