IJARSCT
International Journal of Advanced Research in Science, Communication and Technology
International Open-Access, Double-Blind, Peer-Reviewed, Refereed, Multidisciplinary Online Journal
ISSN: 2581-9429
Volume 6, Issue 20, April 2026
Impact Factor: 8.2

Orchestrated Multi-Agent System with Deterministic RAG for CNC Process Optimization

Eesha Kale[1], Ms. Punam Chavan[2], Dr. Bharati Vasgi[3], Siddhi Pawar[4], Isha Deshmukh[5], Aditi Talnikar[6]
Associate Professor [3], Assistant Professor [2], Student [1,4,5,6]
Department of Information Technology [1,2,3,4,5,6]
Marathwada Mitra Mandal's College of Engineering, Pune, Maharashtra

Abstract: *For amateur CNC users it gets complicated and usually burdensome to acknowledge Computer Numerical Control (CNC) machine operations due to worn out manuals and shortfall of real-time operational guidance. This paper illustrates a deterministic multi-agent chatbot system done on purpose to assist users in learning and executing CNC operations through structured knowledge retrieval. The system incorporates a Retrieval Augmented Generation (RAG) based Large Language Model (LLMs) with vector embeddings extracted from Six Sigma methodologies (DMAIC) and CNC manuals. A multi-agent framework makes it able to handle tasks in modular manner such as query pre-processing, knowledge retrieval, and query response generation. To secure reliability, the system acquires a deterministic method that can lower hallucinations by grounding responses in adapted industrial data. The recent execution is limited to a simulated environment without direct CNC machine integration. Experimental result shows advancement in consistency and relevance in responses, presenting the potential knowledge of knowledge-grounded AI systems for industrial training and support.*

Keywords: CNC Operations, Multi-Agent Systems (MAS), Retrieval Augmented Generation (RAG), Large Language Models (LLMs), Deterministic AI, Six Sigma, OEE, Cycle Time

I. INTRODUCTION

Computer Numerical Control (CNC) machines are essential in advanced manufacturing productions, to ensure high precision, automation, and efficient production process. Although, to perform CNC machine operations it requires highly skilled operators which increases employment expenses in understanding machine instructions, process parameters, and operational workflows. For non-skilled operators, the significant challenges faced while learning due to the intricacy and modular nature of CNC operation manuals and manufacturing industry guidelines. Conventional learning material which includes Six Sigma documentation and CNC manuals, are often static, lengthy, which does not support interactive problem-solving or immediate guidance. As a result, encounter difficulties to interpret information appropriately, applying optimization approaches, and support evidence-based decisions during operational workflows. This produces divergence in the middle of available industrial knowledge and its functional efficiency, especially for beginners and trainees. Contemporary progress in the field of Artificial Intelligence, uncommonly Large Language Models (LLMs), catalyzes advancement in conversational AI for information expertise. Nevertheless, those kinds of systems habitually face hallucination issues and shortage of reliability when deployed within essential industrial sectors. Moreover, most of the existing approaches do not include structured industrial knowledge resources such as CNC manuals or process optimization approaches like Six Sigma in a controlled and deterministic manner. To resolve these challenges, this paper proposes a deterministic multi-agent chatbot system to perform CNC machine operations and learning. The system utilizes a Retrieval-Augmented Generation (RAG) approach to generate responses based on knowledge retrieved from CNC manuals and Six Sigma methodologies. A multi-agent chatbot framework is employed to perform component-based task execution to minimize human errors such as optimized query optimization, knowledge retrieval, and retrieved responses, refactoring system architecture for modularity and expansion to improve

IJARSCT
International Journal of Advanced Research in Science, Communication and Technology
International Open-Access, Double-Blind, Peer-Reviewed, Refereed, Multidisciplinary Online Journal
ISSN: 2581-9429 Volume 6, Issue 20, April 2026 Impact Factor: 8.2

system modularization and adaptability. In addition to this, deterministic approaches are adapted to lower hallucinations and enable consistent, reliable, accurate responses and thereby reduce human errors. The recent implementation pivots on a simulated environment, which illustrates the scalability of incorporating industrial knowledge with agent-oriented AI systems. The objective of recommended approach is to narrow the in the middle of complex technical documentation and smooth user interaction, providing a baseline for future development for real-time CNC automation and performance refinement using industrial manuals.

II. LITERATURE REVIEW

Recent inventions in the field of smart manufacturing to a great extent have an intelligent manufacturing and Industry 4.0 methodologies. Bandhana and Vokřínek [1] provide an wide-ranging investigation of intelligent manufacturing based on AI, focuses on collaboration of multi-agent systems (MAS) and manufacturing executable systems for logical judgement. This investigation spotlights the significance of customizable and concrete implementations, although highlighting constraints in actual operationalization and personnel flexibility. The disclosure of Agentic AI has permitted complementary autonomous and scalable techniques. Zhang et al. [2] conveys workflows dependent on agents that are competent of discrimination, reasoning, and exertion in shifting landscapes, guided by frameworks based on retrieval of knowledge for context-aware decision-making. Similar system performs the capability of multi-agent combination for critical industrial architectures and workflows. In the situations which require manufacturing guidance, Wulf and Meierhofer [3] pass through the Large Language Models (LLMs) for pipeline digitalization, involving answering to queries of users and recapitulation. Their contribution encompasses Retrieval-Augmented Generation (RAG) to spotlight domain-specific correctness, also investigating primary boundries such as hallucination and limitations in merging LLMs into automated production and manufacturing workflows. Rakes et al. [4] analyzed CNC machine performance from an operational standpoint using Overall Equipment Effectiveness (OEE), showing how data-driven maintenance techniques are capable of reducing downtime and boosting productivity. In a similar manner, AI-based methods for measuring cycle time [5] exhibit the growing function of intelligent systems in monitoring and optimization of manufacturing process. Despite these advancements, many current fact-findings ensure that either standalone LLM-based guiding systems, smart manufacturing automation, or CNC operational performance. Very poor study has been carried out on collaboration of RAG-based LLMs with multi-agent workflows to build deterministic, domain-specific chatbot mechanisms for CNC operations, especially for novice CNC operators. In addition to this, there is still an inadequate rumination to the problem of minimizing hallucinations while embedding structured industrial knowledge, such as CNC manuals and Six Sigma methodologies. This inequality demonstrates as the primary impact behind the above proposed survey, that take an attempt to build a deterministic multi-agent chatbot system i.e. conversational AI that upholds reliable and effective CNC machine interface by deploying RAG, CNC operational knowledge, and industrial optimization principles.

III. METHODOLOGY

This part confers the design and implementation of a deterministic multi-agent chatbot system for CNC machine assistance, combining Retrieval Augmented Generation (RAG), manufacturing data, six sigma DMAIC methodologies, and CNC operation manuals. The proposed system obeys a flexible architecture made up of data processing, knowledge presentation, multi-agent reasoning, and orchestration layers.

A. System Architecture

The suggested system is outlined as a hierarchical Multi-Agent System (MAS) involving four primary layers: (i) data and knowledge layer, (ii) knowledge retrieval Layer, (iii) cognitive reasoning layer, and (iv) orchestration layer. At the baseline, multiple context-sensitive knowledge sources are preserved, involving Six Sigma documentations, CNC operation manuals, Overall Equipment Effectiveness (OEE) datasets, and cycle time datasets. Every knowledge source is individually processed and indexed so that it boosts effective knowledge retrieval. Beyond this, the knowledge

IJARSCT
International Journal of Advanced Research in Science, Communication and Technology
International Open-Access, Double-Blind, Peer-Reviewed, Refereed, Multidisciplinary Online Journal
ISSN: 2581-9429
Volume 6, Issue 20, April 2026
Impact Factor: 8.2

retrieval layer made up of multiple specialized RAG modules, one and all responsible for a specific domain. These involve:

- CNC operation manuals RAG module for machine operation queries,
- Six Sigma RAG module for quality advancement strategies,
- OEE RAG module for equipment efficiency investigation,
- Cycle time RAG module for CNC production performance discriminations,
- A web scraping module works as a fallback system when context-aware domain-specific knowledge is not sufficient to answer queries.

The cognitive reasoning layer is made up of three agents depending on each other:

- Analyzer Agent - performs detailed reasoning, incorporates Six Sigma principles (DMAIC), CNC operation manuals, and analyses operational data such as cycle time and OEE,
- Predict and Plan Agent - Understands user queries intent and generates prediction and planning based on retrieved context,
- Mentor and Guide Agent - Improves responses to make sure clarity, usability, and suitability for non-expert CNC users.

At the topmost level, an orchestrator agent dynamically handles all the interaction between all above mentioned agents. It routes user queries, chooses relevant RAG modules, and ensures ordered execution of cognitive agents based on user query context.

B. Data Collection and Pre-processing

The proposed system utilizes diverse data sources, involving structured and unstructured formats. Six Sigma documentation and CNC operation manuals are collected in PDF format, whereas OEE and cycle time datasets are acquired as structured CSV data presenting real-world industrial production scenarios. All data goes through pre-processing steps incorporating cleaning, normalization, and segmentation. Unstructured text data gets split into smaller chunks to make it able for efficient embedding, whereas structured data is formatted for analytical querying. All the processed data gets stored in a PostgreSQL database.

C. Knowledge Representation and Retrieval

To permit efficient semantic retrieval the proposed system makes use of a Retrieval-Augmented Generation (RAG) process. Text chunks from all knowledge resources are converted into vector embeddings and preserved using the PGVector extention in PostgreSQL. One and all domain-specific data is indexed independently, permitting targeted knowledge retrieval based on user query context. Throughout query processing, relevant embeddings are retrieved using similarity search and transferred to the large language model as contextual input. To make sure deterministic behaviour and lower hallucinations, the proposed system restricts query responses strictly to retrieve knowledge. If there is no relevant context found, the system returns a fallback response which will indicate lack of information instead of generating hypothetical answers.

IJARSCT
International Journal of Advanced Research in Science, Communication and Technology
International Open-Access, Double-Blind, Peer-Reviewed, Refereed, Multidisciplinary Online Journal
ISSN: 2581-9429
Volume 6, Issue 20, April 2026
Impact Factor: 8.2

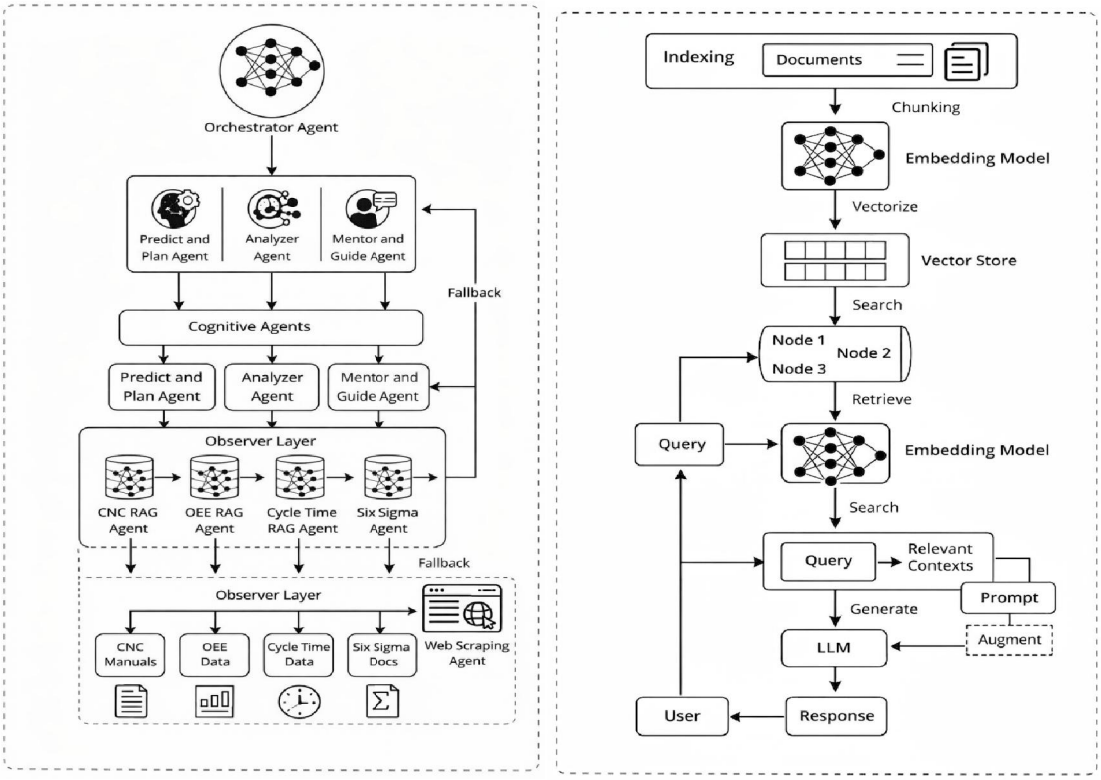

Fig.1. Document Indexing and Embedding Pipeline for RAG-Driven Knowledge Retrieval

Fig.2. System Architecture of a Multi-Agent Deterministic Framework for CNC Process Optimization

The processing pipeline of document is used for indexing, chunking, and embedding generation is decorated in Fig.1, which makes the baseline for efficient semantic retrieval in the above proposed system.

D. Multi-Agent Reasoning Framework

The reasoning methodology obeys a linear agent pipeline. After receiving a query from user, the orchestrator agent first calls out to the correct RAG modules. The retrieved context then delivered to the Predict and Plan Agent, which decides the logical flow of reasoning. The analyser agent eventually processes the context using domain-specific logic, incorporating Six Sigma DMAIC principles for fault diagnosis and process production optimization. This permits the proposed system to supply actionable insights relevant to performance progress. Finally, the Mentor and Guide Agent developers the response to make sure interpretability for non-expert CNC users, particularly for those who are not familiar with CNC machine operations.

F. LLM Integration

The system utilizes Gemini 2.5 Flash large language model for natural language acknowledgement and thereby response generation. The LLM works within a controlled RAG pipeline, which makes sure that outputs are grounded in retrieval knowledge instead of generative assumptions. The integration is executed using the AGNO agentic AI framework, which simplifies structured communication between agents and helps modular flexibility.

IJARSCT
International Journal of Advanced Research in Science, Communication and Technology
International Open-Access, Double-Blind, Peer-Reviewed, Refereed, Multidisciplinary Online Journal
ISSN: 2581-9429
Volume 6, Issue 20, April 2026
Impact Factor: 8.2

H. Web Interface and Deployment

A web-oriented interface is made using React to enable user interaction with the CNC chatbot system. The backend is executed using FastAPI in Python, providing APIs for query processing and orchestration of all agents. Recently, the system is deployed in a local environment for testing and validation. The structured design permits future deployment in industrial production settings with lesser changes.

G. Deterministic Design and Hallucination Control

A key offering of this work is the development of a deterministic agentic framework. Far from conventional LLM-driven systems, the presented technique enforces strict dependency on retrieved knowledge through:
domain-restricted RAG pipelines,
controlled agent communications,
exact fallback responses for unknown queries.

H. System Evaluation

The proposed system is right now evaluated through unit-level testing and scenario-oriented examination. CNC operations related queries, OEE analysis, cycle time reduction predictions, and Six Sigma recommendations are checked out for verifying accuracy and consistency. However, large-scale deployment is not yet conducted, preparatory results performs that the system effectively supplies correct, context-aware, and non-hallucinating responses to user queries, validating the feasibility of the proposed methodology.

I. Methodology Overview

The workflow proposed above settles a organized and flexible multi-agent architecture for CNC process enhancement by merging deterministic RAG-driven investigation within the ordered cognitive reasoning layer. Although the coordination of specialised agents coming under to orchestrator agent, the system validates correct, context-aware, and reliable decision guidance. The integration of fallback mechanisms and domain-specific knowledge resources added do further advancement in robustness and scope. Generally, this way permits continuous, automated, and valid observation, forming it well-suited for real-world industrial application having necessity of precision, scalability, and lower human intervention.

V. RESULTS

The deterministic multi-agent chatbot system presented in this paper was assessed with the help of several functional and scenario-based tests for monitoring performance across CNC operations, OEE analysis, cycle time evaluation, and Six Sigma based guidance manuals. The assessment ensures precise response generation, determinism, contextual awareness, reduced human error, minimized cost and system robustness.

A. Functional Validation

For diverse query categories, the proposed system has been examined including:

- CNC machine operation guidance,
- OEE performance assessment,
- Cycle time query processing enhancement,
- Six sigma based fault reduction suggestions.

The proposed multi agent system has the ability to retrieve appropriate domain-specific knowledge from guidance manuals using multi- agent RAG framework and generate contextual relevant responses. On combining structured datasets (CSV) and unstructured manuals (PDFs) allows the system to demonstrate both analytical and descriptive query execution effectively.

IJARSCT
International Journal of Advanced Research in Science, Communication and Technology
International Open-Access, Double-Blind, Peer-Reviewed, Refereed, Multidisciplinary Online Journal
ISSN: 2581-9429
Volume 6, Issue 20, April 2026
Impact Factor: 8.2

B. Deterministic Response Assessment

The primary assessment criterion was the ability of system to avoid hallucination in derived responses. Unlike rigid, script-based AI approaches, the proposed system strictly relies on knowledge retrieval. In cases where relevant data to query cannot be retrieved, the system consistently returns fallback responses using web scraping to ensure reliability. This deterministic behavior was examined through repetitive query testing, where identical queries consistently produced stable and reproducible outputs.

C. Multi-Agent Performance Analysis

The effectiveness of the multi-agent framework to generate context aware responses was examined on the basis of response generated by each agent layer:

- The Predict and Plan **Agent** successfully interprets intent of user queries and select appropriate knowledge sources for retrieval.
- The Analyzer Agent illustrates ability to apply Six Sigma DMAIC principles to identify inefficiencies in OEE and cycle time data effectively.
- The Mentor and Guide Agent enhance response clarity, making responses more precise and understandable for non-skilled manufacturing operators. Coordinate agent interactions by agent orchestrator, enable a seamless flow from retrieval to reasoning and response generation.

D. Knowledge Retrieval Accuracy

The Vector embeddings ensure efficient semantic search across multiple CNC machine operations and Six Sigma manuals. The system depicts high relevance in retrieved responses. The web scraping fallback mechanism is included to enhance system robustness by providing additional contextual support in case of internal knowledge sources are insufficient.

E. System Responsiveness

The multi-agent chatbot framework was tested in a local environment using Fast API backend services to examine its accuracy fir query processing. Query response generation times were examined to be within acceptable limits for real-time interaction, illustrating the feasibility of deploying such systems in industrial manufacturing sector.

F. LLM and RAG

On comparing with conventional chatbot systems and standalone LLM-based architecture, the proposed system provides:

- Enhanced reliability through deterministic design,
- Enhanced domain specific response generation via multi- RAG architecture,
- Better interpretability using multi-agent reasoning,
- Controlled knowledge grounding for non-hallucinating responses.

These observations represents that the system provides a real-time and scalable solution for intelligent CNC assistance.

V. DISCUSSION

The experimental results focus on effective collaboration multi-agent chatbot systems with retrieval- augmented generation for various manufacturing industry approaches. The proposed framework successfully reduces the gap between generative AI capabilities and domain-specific reliability necessity. Deterministic response generation emphasized in the design of the proposed system. By making the large language model limited to retrieval using structured knowledge retrieval and fallback mechanisms, the system generates non hallucinating responses with

IJARSCT
International Journal of Advanced Research in Science, Communication and Technology
International Open-Access, Double-Blind, Peer-Reviewed, Refereed, Multidisciplinary Online Journal
ISSN: 2581-9429 Volume 6, Issue 20, April 2026 Impact Factor: 8.2

more precision, which is a very difficult in conventional LLM-based systems. This makes the solution more suitable for high-risk manufacturing sector. Various RAG modules based upon specific domain enable the system to maintain modularity and scalability. Each knowledge source operates independent of each other, to enable future enhancements in automating manufacturing operations with less human errors such as predictive maintenance or IoT-based assessment without significant architectural changes. The Six Sigma principles incorporated within the Analyzer Agent creates an effective layer of manufacturing intelligence for enhancing the simple information retrieval. This ensures that system not only provide responses to queries but also gives actionable insights for process improvement. However, the recent system is dependent on controlled datasets and local deployment. Since the results are satisfactory and validated in large-scale industrial environments is necessary for enhancement of complete system performance assessment under dynamic conditions. Overall, the research illustrates that on collaborating MAS, RAG, and industrial knowledge methodologies are effective to enhance the reliability and usability of AI-driven manufacturing assistants.

VI. CHALLENGES AND LIMITATIONS

Though early responses were acceptable, still various challenges and limitations were found during the development and examination of the system.

The availability of real-time CNC machine operational data was limited. Currently system completely relies on sample knowledge base for OEE and cycle time analysis, which is not able to capture real-world variability and complexity completely.

The quality and coverage of knowledge resources determine the performance accuracy of the system. Traditionally available CNC machine operation manuals and Six Sigma documentation degraded retrieval accuracy and entire system performance.

Deterministic multi-agent chatbot reduces hallucination, which may restrict the system's capability to generate responses for queries beyond the information available knowledge base, resulting in frequent fallback responses.

In addition to the current system performs examination in a local environment, which can lead to difficulties in scaling the system to be used in large scale environments like dealing with large- scale concurrent query processing and integration with manufacturing systems remain unaddressed. Finally, the multi-agent architecture creates computational overhead, which may impact response latency in real-time industrial deployments. These difficulties focuses on the requirements for further research in real-time data integration precisely, scalable deployment, and adaptive learning mechanisms.

VII. CONCLUSION

This paper proposes a deterministic multi-agent chatbot system for CNC machine assistance with the integration of Retrieval- Augmented Generation, domain-based knowledge sources CNC machine operations manuals and Six Sigma methodologies. The proposed system tries to overcome several constraints of traditional AI-based chatbots, especially hallucination and lack of domain reliability. By using a hierarchical multi-agent system framework, the system efficiently combines retrieval, reasoning, and user assistance for providing precise and interpretable responses. On combining structured industrial knowledge resources with unstructured documentation ensures comprehensive support for CNC operations, OEE analysis, and cycle time optimization. The results depicts that the proposed approach provides a reliable and scalable system for intelligent manufacturing assistance, particularly for non-skilled manufacturing operators. The deterministic system ensures trustworthiness, reduces human errors and automated process executions to make system suitable for industrial environments where accuracy is important. Future work will highlight automated CNC integration, large-scale deployment, and improvement in adaptive learning capabilities to further improve system performance and applicability.

IJARSCT
International Journal of Advanced Research in Science, Communication and Technology
International Open-Access, Double-Blind, Peer-Reviewed, Refereed, Multidisciplinary Online Journal
ISSN: 2581-9429
Volume 6, Issue 20, April 2026
Impact Factor: 8.2

ACKNOWLEDGMENT

The authors express their sincere gratitude to both Parentheses Systems Private Limited, Pune, and Department of Information Technology, Marathwada Mitra Mandal's College of Engineering, Pune, for their valuable support and continuous guidance throughout the research which led to successful completion of this work. Additionally, gratitude is expressed to the contributors who provided CNC manuals and Six Sigma resources used in this study. Their contributions played a crucial role in building the domain- specific knowledge base for the system.

REFERENCES

[1]. Bandhana and J. Vokřínek, "AI-Driven Manufacturing: Surveying for Industry 4.0 and Beyond," Operations Research Forum, vol. 6, no. 4, pp. 145, Sep. 2025. doi: 10.1007/s43069-025-00554-6. [Online].

[2]. Y. Li, W. Zhang, Y. Yang, W.-C. Huang, Y. Wu, J. Luo, Y. Bei, H. P. Zou, X. Luo, Y. Zhao, C. Chan, Y. Chen, Z. Deng, Y. Li, H.-T. Zheng, R. Jiang, M. Zhang, Y. Song, and P. Yu, "Towards Agentic RAG with Deep Reasoning: A Survey of RAG-Reasoning Systems in LLMs," arXiv, Jul. 13, 2025. doi: 10.48550/arXiv.2507.09477. [Online].

[3]. Jochen Wulf, Jürg Meierhofer, "The Impact of Large Language Models on Task Automation in Manufacturing Sevices", Procedia CIRP – Proceedings of the 58th Conference on Manufacturing Systems (CMS 2025), University of Twente, The Netherlands.

[4]. David Rakes, Muhammad Arif, Agus Setiawan, Kerina Putri Nasution, and Yudi Prastyo, "Preventive Maintenance on CNC Machines Using the OEE Method to Reduce Downtime at PT. MTAT", July 2024, [Online].

[5]. Y. Li, W. Zhang, Y. Yang, W.-C. Huang, Y. Wu, J. Luo, Y. Bei, H. P. Zou, X. Luo, Y. Zhao, C. Chan, Y. Chen, Z. Deng, Y. Li, H.-T. Zheng, R. Jiang, M. Zhang, Y. Song, and P. Yu, "Towards Agentic RAG with Deep Reasoning: A Survey of RAG-Reasoning Systems in LLMs," arXiv, Jul. 13, 2025. doi: 10.48550/arXiv.2507.09477. [Online].

[6]. H. Sanchaniya, D. Sinha, A. Joshi, S. Kothimbire, B. Vasgi, and P. Chavan, "Enhancing CNC Machine Operator Accessibility through a Multimodal Chatbot," International Journal of Innovative Science and Research Technology (IJISRT), vol. 10, no. 4, pp. 3140–3146, Apr. 2025. doi: 10.38124/ijisrt/25apr1796. [Online].

[7]. S. Răileanu and T. Borangiu, "A Review of Multi-agent Systems Used in Industrial Applications," in Service Oriented, Holonic and Multi-Agent Manufacturing Systems for Industry of the Future, T. Borangiu, D. Trentesaux, and P. Leitão, Eds., SOHOMA 2022, Studies in Computational Intelligence, vol. 1083. Cham, Switzerland: Springer, 2023, pp. 1–23. doi: 10.1007/978-3-031-24291-5_1. [Online].

[8]. S. Raza, R. Sapkota, M. Karkee, and C. Emmanouilidis, "TRiSM for Agentic AI: A Review of Trust, Risk, and Security Management in LLM-based Agentic Multi-Agent Systems," arXiv, Jun. 4, 2025. [Online].

[9]. X. Li, S. Wang, S. Zeng, et al., "A survey on LLM-based multi-agent systems: workflow, infrastructure, and challenges," Vicinagearth, vol. 1, p. 9, 2024. doi: 10.1007/s44336-024-00009-2. [Online].

[10]. D. B. Acharya, K. Kuppan, and B. Divya, "Agentic AI: Autonomous Intelligence for Complex Goals—A Comprehensive Survey," IEEE Access, vol. 13, pp. 18912–18936, 2025. doi: 10.1109/ACCESS.2025.3532853.

[11]. R. S. Peres, X. Jia, J. Lee, K. Sun, A. W. Colombo, and J. Barata, "Industrial Artificial Intelligence in Industry 4.0—Systematic Review, Challenges and Outlook," IEEE Access, vol. 8, pp. 220121–220139, 2020. doi: 10.1109/ACCESS.2020.3042874.