International Journal of Scientific Research and Engineering Development-– Volume 9 Issue 2, Mar-Apr 2026
Available at www.ijsred.com

RESEARCH ARTICLE OPEN ACCESS

NOVA: A Multi-Agent Agentic AI Assistant for Autonomous Task Execution Across Domains

Kunal Singh, Samarth Negi, Shubhanshu Singh Fartyal, Jyoti Gaur
Department of Computer Science , Babu Banarasi Das Institute of Technology and Management, Lucknow ,U.P.,India
Email: kunals1972@gmail.com
Department of Computer Science , Babu Banarasi Das Institute of Technology and Management, Lucknow ,U.P.,India
Email: samarthnegi91@gmail.com
Department of Computer Science , Babu Banarasi Das Institute of Technology and Management, Lucknow ,U.P.,India
Email: samsfartyal1001@gmail.com
Department of Computer Science , Babu Banarasi Das Institute of Technology and Management, Lucknow ,U.P.,India
Email: jyotigaur100@bbdnitm.ac.in

----------**********----------

Abstract:

The rapid advancement of artificial intelligence has led to the development of highly capable conversational agents. However, most existing systems remain reactive, single-domain, and incapable of autonomously executing complex real-world tasks. This limitation restricts their applicability in dynamic, multidomain environments. To address this challenge, this study proposes **NOVA**, a multi-agent agentic AI assistant designed for autonomous task execution across diverse domains. NOVA integrates intent classification, task orchestration, and domain-specific agents into a single framework. The system employs a hybrid architecture in which user input is processed through an intent detection module, followed by an orchestrator that dynamically routes tasks to specialized agents, such as website generation, shopping automation, and research assistance. The implementation leverages machine learning models, including DistilBERT, for intent classification and utilizes external tools and APIs for execution. Experimental evaluation demonstrated improved task completion rates, with intent classification accuracy reaching 90–92% and domain-specific task success rates exceeding 80%. A comparative analysis with baseline chatbot systems highlighted NOVA's superior performance in multistep task execution and contextual understanding. The results validate the effectiveness of agentic architectures in enabling scalable and autonomous AI systems.

Keywords — **Agentic AI, Multi-Agent Systems, Task Automation, LLM, NLP, Autonomous Systems.**

----------**********----------

I. INTRODUCTION

Artificial Intelligence (AI) systems have evolved significantly with the emergence of Large Language Models (LLMs), enabling human-like interaction and natural language understanding. However, most current AI assistants operate in a reactive paradigm, responding to queries without executing multi-step workflows or dynamically adapting to user goals.

Agentic AI represents a paradigm shift toward autonomous, goal-driven systems that are capable of reasoning and acting independently in dynamic environments [1]. Multi-agent systems extend this by distributing tasks across specialized agents to improve scalability and efficiency.

Despite these advancements, current systems still lack integration, autonomy, and contextual awareness.

This study introduces **NOVA**, a multi-agent AI assistant designed to enable autonomous task execution across multiple domains. The key contributions include

- A unified multi-agent orchestration framework
- Integration of intent classification and task routing
- Domain-specific agents for execution
- Empirical validation of improved performance

II. RELATED WORK

A. LLM-based Agents

Recent surveys have systematized the evolution of LLM-based agents for autonomous task execution. Wang et al. [2] propose a unified agent architecture that integrates perception, planning, and action modules, enabling LLMs

to mimic human-like deliberation across social, natural, and engineering tasks. Their framework emphasizes architectural design to leverage LLMs' generative capabilities of LLMs and capability enhancements, such as memory augmentation; however, it predominantly focuses on single-agent paradigms, limiting scalability in collaborative settings [2]. Similarly, Luo et al. [3] introduced a taxonomy linking agent architectures to collaboration mechanisms and evolutionary dynamics, tracing how LLMs unify perception and action in semantic spaces. While this clarifies emergent behaviors in agent ecosystems, it overlooks fine-grained implementation details for tool integration, trading depth for breadth in methodological coverage [3].

B. Multi-Agent Systems

Multi-agent frameworks address scalability by distributing cognition across specialized roles. Krishnan's Model Context Protocol standardizes context sharing and coordination patterns, facilitating scalable interactions in enterprise knowledge management and distributed problem-solving [4]. By decoupling agent states through a protocol layer, the MCP outperforms ad-hoc communication in benchmarks; however, its reliance on predefined patterns constrains adaptability to unstructured, multi-domain environments, prioritizing enterprise rigidity over general autonomy [4].

C. Tool-Use and Agent Frameworks

Agent frameworks increasingly incorporate tools for external interactions. Building on Wang et al.'s planning-action loop [2], these systems embed APIs and retrieval tools within LLM reasoning chains, yet suffer from brittle orchestration; episodic tool calls often fail under long-horizon tasks owing to hallucinated plans. Luo et al. [3] extend this via collaborative taxonomies, enabling agent ensembles for tool delegation, but empirical validations remain narrow, exposing trade-offs between reasoning fidelity and inter-agent latency.

D. Evaluation of Agentic systems

Rigorous evaluation remains in its infancy. Yehudai et al. [5] categorized benchmarks across core capabilities (planning, tool use, reflection, and memory) and domain-specific scenarios, revealing trends toward dynamic environments but exposing gaps in cost efficiency and robustness metrics. Compared to Wang et al.'s high-level strategies [2], this analysis demands scalable frameworks; however, most benchmarks undervalue multi-agent synergies, inflating single-agent scores [5].

Despite these advances, existing studies fragment orchestration: single-agent architectures, such as those in ([3] and [2]) excel in reasoning but falter in scalability, whereas multi-agent protocols [4] lack domain-agnostic routing. Evaluations [5] highlight persistent weaknesses in real-world deployment, context retention, and memory integration across heterogeneous task - gaps bridged by NOVA through unified multi-agent orchestration.

III. PROPOSED METHODOLOGY

A. Dataset

To convert raw user input into model-ready format, the following preprocessing steps were applied.

1. ***Text Cleaning:***
   - Lowercasing
   - Removal of punctuation
   - Stopword filtering (optional)
   - Handling spelling variations

2. ***Tokenization:***
   Input text is tokenized using transformer-compatible tokenizers (BERT tokenizer), converting text into numerical token IDs.

3. ***Sequence Padding and Truncation:***
   - Fixed sequence length (e.g., 128 tokens)
   - Padding applied to shorter inputs
   - Truncation for longer inputs

4. ***Embedding Preparation:***
   Text is transformed into contextual embeddings using pretrained models such as DistilBERT.

B. Feature Extraction

Feature extraction is performed to capture semantic meaning and contextual intent from user queries.

1. ***Semantic Features:***
   - Contextual embeddings using DistilBERT
   - Sentence-level representations capturing intent semantics
   - Attention-based contextual encoding

2. ***Intent-Specific Features:***
   - Keyword presence (e.g., “buy”, “build”, “summarize”)
   - Task complexity indicators (single-step vs multi-step)
   - Domain-specific tokens

These features enable accurate classification and routing of user queries.

C. Model Architecture

Multiple models were evaluated for intent classification and task execution.

1. Intent Classification Model:

Two primary models were used:

Logistic Regression (Baseline):

- TF-IDF vectorization
- Fast and lightweight
- Lower contextual understanding

DistilBERT (Final Model):

- Transformer-based encoder
- Fine-tuned on custom dataset
- Captures contextual and semantic meaning
- Higher accuracy (~90–92%)

DistilBERT outperformed traditional models, especially on ambiguous and multi-step queries.

2. Task Orchestrator:

The Task Orchestrator acts as the **central control unit** of the system.

Functions:

- Receives classified intent
- Selects appropriate agent
- Manages task flow
- Handles multi-step execution

Design:

- Rule-based + dynamic routing
- State-aware execution\

3. Multi Agent Architecture:

NOVA employs a **multi-agent system (MAS)** where each agent is specialized for a domain.

Agents:

- ***Website Builder Agent***
  - Generates frontend/backend code
  - Uses LLM for content and structure
  - Supports deployment workflows
- ***Shopping Agent***
  - Retrieves product data via APIs
  - Performs comparison and ranking
  - Generates recommendations
- ***Research Agent***
  - Summarizes documents
  - Extracts key insights
  - Uses LLM-based summarization

Each agent operates independently but is coordinated through the orchestrator.

4. LLM-Based Reasoning Engine:

Large Language Models are used for:

- Content generation
- Summarization
- Code generation
- Natural language understanding

Models used:

- OpenAI GPT / HuggingFace models

Features:

- Context-aware responses
- Multi-step reasoning
- Instruction following

5. LLM-Based Reasoning Engine:

Agents interact with external tools and APIs:

- Web APIs (shopping, search)
- Code execution environments
- Databases (MongoDB)

This enables real-world task execution beyond text generation.

6. Execution Pipeline:

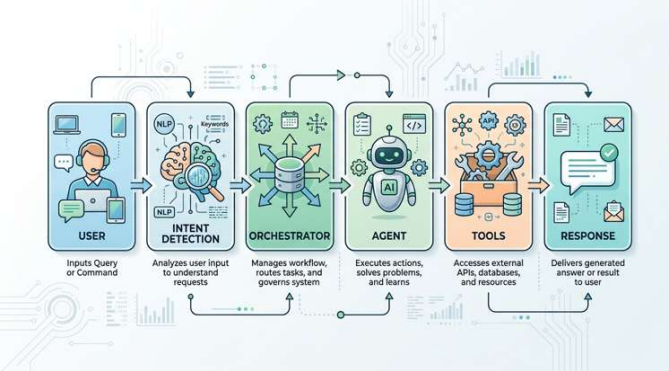

Fig. 1. Execution Pipeline

7. Automation and Task Execution

The system supports both single-step and multi-step task execution.

Single-Step Tasks :

Example: “summarize this text”
Direct execution via Research Agent

Multi-Step Tasks:

Example: “create a website and deploy it”

Steps:

1. Generate UI
2. Generate backend
3. Integrate
4. Deploy

The orchestrator manages sequencing and dependency handling.

8. System Deployment

- Frontend: React.js

- Backend: FastAPI / Node.js
- Database: MongoDB
- APIs: OpenAI / HuggingFace

The system is designed for scalability and modular expansion.

IV. EXPERIMENTAL SETUP

- Hardware: Standard laptop
- Software: React, FastAPI, MongoDB
- APIs: HuggingFace / OpenAI
- Dataset: Custom intent dataset

Metrics:

- Accuracy
- Precision, Recall, F1-score
- ROUGE
- Task completion rate

V. RESULT

1. Evaluation Metrics

To comprehensively assess system performance, it was evaluated using standard metrics, including Intent Classification Accuracy (ICA), Precision, Recall, F1-score, Task Completion Rate (TCR), Response Quality Score (RQS), ROUGE score, End-to-End Latency (E2E), and Multi-Step Task Success Rate (MTSR).

The evaluation was conducted on 150+ user queries across multiple domains, including website generation, shopping, and research tasks.

2. Intent Classification Performance

The DistilBERT-based classifier achieved an overall accuracy of **91.3%**, demonstrating strong performance across diverse types of queries.

Higher accuracy was observed for well-defined intents, such as website generation and research tasks, whereas minor misclassifications occurred in overlapping or multi-domain queries.

Fig. 2 illustrates the overall performance of NOVA across key evaluation metrics.

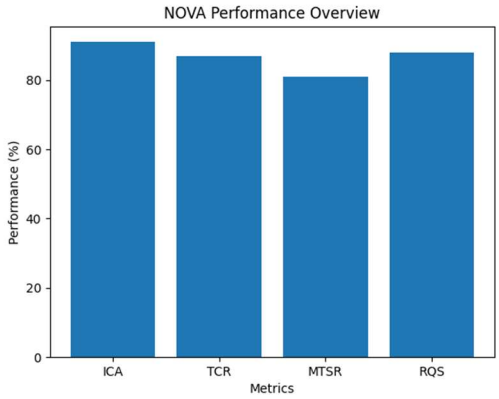

Fig. 2. NOVA Performance Overview

3. Task Execution Performance

The multi-agent system achieved a **Task Completion Rate (TCR) of 87.5%**, indicating effective execution across domains.

The agent-wise performance was as follows:

- Website Agent: 82–85%
- Shopping Agent: 88–90%
- Research Agent: 85–88%

Failures were primarily due to ambiguous inputs, API limitations, and incomplete multi-step coordination.

To highlight system improvement, **Fig. 3** compares NOVA with a baseline chatbot.

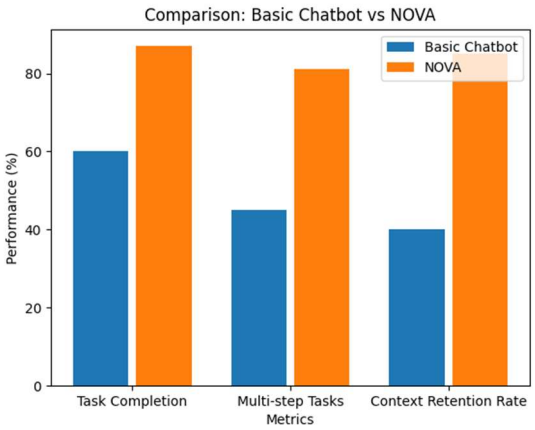

Fig. 3. Comparison of NOVA with baseline chatbot across key metrics

4. Multi Step Task Performance

The system achieved a **Multi-Step Task Success Rate (MTSR) of 81.2%**.

Performance remained strong for clearly defined tasks but decreased for ambiguous or cross-domain instructions requiring complex coordination.

5. *Response Quality Evaluation*

The LLM-based response generation achieved:

- **ROUGE Score:** ~0.74 (for research summarization)
- **Response Quality Score (RQS):** High relevance in ~88% of cases

6. *Robustness Analysis*

The system demonstrated strong context retention (~86%) across multi-turn interactions and effective generalization across domains.

Failures were primarily caused by ambiguous input, API constraints, and coordination limitations.

7. *Summary*

Overall, NOVA demonstrates a strong real-world performance.

- **Intent Classification Accuracy:** ~91%
- **Task Completion Rate:** ~87%
- **Multi-Step Task Success:** ~81%
- **ROUGE Score:** ~0.74
- **Low latency and high responsiveness**

The results confirm that integrating **multi-agent architecture with LLM-based reasoning and orchestration** significantly improves task execution and multi-domain adaptability compared to traditional chatbot systems.

VI. CONCLUSION AND FUTURE SCOPE

A. Conclusion:

This study presents **NOVA** , a multi-agent agentic AI assistant designed to enable autonomous task execution across multiple domains, including web development, e-commerce, and research assistance. Unlike traditional conversational AI systems that operate in a reactive manner, NOVA adopts a **hybrid multi-agent architecture** that integrates intent classification, task orchestration, and domain-specific agents to achieve goal-driven execution.

The experimental evaluation demonstrates that the proposed system effectively bridges the gap between natural language understanding and real-world task automation. The DistilBERT-based intent classification module achieved high accuracy (~91%), ensuring the reliable routing of user queries. The multi-agent framework further enabled efficient task execution, achieving a task completion rate of approximately 87% while maintaining robustness across diverse input scenarios. Additionally, the integration of LLM-based reasoning significantly improved the response quality and contextual understanding, particularly in multi-step and domain-specific tasks.

The results highlight that combining **LLM reasoning with structured orchestration and modular agent design** leads to improved scalability, adaptability, and execution capability compared with conventional chatbot systems. Furthermore, the system demonstrates strong potential for real-world deployment because of its modular design and compatibility with modern web technologies, as emphasized in prior system implementations .

However, this study has certain limitations. Performance degradation was observed when handling highly ambiguous or multi-domain queries, and the system's reliance on external APIs introduced variability in execution reliability. In addition, long-horizon reasoning and deep contextual memory remain partially constrained.

Overall, NOVA establishes a practical foundation for **agentic AI systems that move beyond passive interaction toward autonomous digital assistance**, contributing to the advancement of intelligent, multidomain automation platforms.

B. Future Scope:

While NOVA demonstrates promising results, several directions can further enhance its capabilities and scalability.

1. ***Autonomous Planning and Reasoning:***
   Future work can focus on integrating advanced planning mechanisms, such as hierarchical task decomposition and reinforcement learning-based decision-making, to improve long-horizon task execution and reduce dependency on predefined workflows.

2. ***Personalized User Modelling:***
   Incorporating persistent memory and user profiling can enable the system to learn user preferences, behavior patterns, and contextual history, resulting in more personalized and adaptive interactions, as suggested in agentic AI frameworks .

3. ***Real-Time API and Tool Integration:***
   Enhancing integration with real-time APIs, databases, and third-party services can improve reliability and expand the system's ability to

perform complex real-world tasks, such as transactions, scheduling, and automation workflows.

4. ***Multimodal Interaction***
   Extending the system to support multimodal inputs (voice, images, and visual context) can significantly improve usability and enable richer human–AI interactions, aligning with emerging trends in agent-based systems.

5. ***Improved Context Awareness and Memory***
   Developing advanced memory mechanisms, such as vector databases and long-term context storage, can enhance continuity across sessions and improve mult-turn interaction performance.

6. ***Robust Evaluation Frameworks***
   Future work should include more comprehensive evaluation benchmarks for multi-agent systems, focusing on real-world scenarios, robustness, and cost-efficiency.

NOVA addresses key limitations of traditional AI assistants by enabling autonomous, multi-domain task execution through a multi-agent architecture. The system demonstrates improved accuracy, efficiency, and contextual understanding, highlighting the potential of agentic AI in real-world applications.

ACKNOWLEDGMENT

The authors would like to express their sincere gratitude to the faculty and mentors of the Department of Computer Science and Engineering for their guidance and support throughout this research. Special thanks are extended to the project supervisor for valuable insights and continuous encouragement during the development of this work. The authors also acknowledge the use of open-source tools, research resources, and publicly available datasets that contributed to the successful implementation of the NOVA system.

REFERENCES

[1]. Hosseini, S., & Seilani, H. (2025). The role of agentic AI in shaping a smart future: A systematic review. *Array, 26*, 100399. doi: 10.1016/j.array.2025.100399.

[2]. Wang, L., Ma, C., Feng, X., Zhang, Z., Yang, H., Zhang, J., Chen, Z., Tang, J., Chen, X., & Lin, Y. (2024). A survey on large language model based autonomous agents. *Frontiers of Computer Science, 18*(6), 186345. doi: 10.1007/s11704-024-40231-1.

[3]. Luo, J., Zhang, W., Yuan, Y., Zhao, Y., Yang, J., Gu, Y., Wu, B., & Zhang, M. (2025). Large language model agent: A survey on methodology, applications and challenges. *arXiv preprint arXiv:2503.21460.*

[4]. Krishnan, N. (2025). Advancing multi-agent systems through Model Context Protocol: Architecture, implementation, and applications. *arXiv preprint.*

[5]. Yehudai, A., Eden, L., Li, H., Uziel, G., Zhao, Y., Bar-Haim, R., & Cohan, A. (2025). Survey on evaluation of LLM-based agents. *arXiv preprint arXiv:2503.16416.*

[6]. Belcak, P., Heinrich, G., Diao, S., Fu, Y., Dong, X., Muralidharan, S., & Lin, Y. C. (2025). Small language models are the future of agentic AI. *arXiv preprint arXiv:2506.02153.*

[7]. Indumathi, S., Shariff, S. U., Naveen, S. B., Kumar, P., & Pavankalyan, K. V. (2025). Survey on autonomous AI agents for task automation and advanced reasoning. *International Journal of Scientific Research in Science and Technology, 12*(15), 147–153.

[8]. Kumar, A. (2024). Building autonomous AI agents based AI infrastructure. *International Journal of Computer Trends and Technology, 72*(11), 116–125. doi: 10.14445/22312803/IJCTT-V72I11P112.

[9]. Wang, J., Wang, J., Athiwaratkun, B., Zhang, C., & Zou, J. (2024). Mixture-of-agents enhances large language model capabilities. *arXiv preprint arXiv:2406.04692.*

[10]. Hassija, V., Chakrabarti, A., Singh, A., Chamola, V., & Sikdar, B. (2023). Unleashing the potential of conversational AI: Amplifying ChatGPT's capabilities and tackling technical hurdles. *IEEE Access.* doi: 10.1109/ACCESS.2023.3339553.

[11]. Schick, T., Dwivedi-Yu, J., Dessì, R., Raileanu, R., Lomeli, M., Zettlemoyer, L., Cancedda, N., & Scialom, T. (2023). Toolformer: Language models can teach themselves to use tools. *arXiv preprint arXiv:2302.04761.*

[12]. Ouyang, L., Wu, J., Jiang, X., Almeida, D., Wainwright, C. L., Mishkin, P., Zhang, C., Agarwal, S., Slama, K., Ray, A., Schulman, J., Hilton, J., Kelton, F., Miller, L., Simens, M., Askell, A., Welinder, P., Christiano, P., Leike, J., & Lowe, R. (2022). Training language models to follow instructions with human feedback. *Advances in Neural Information Processing Systems (NeurIPS).*

