Volume: 13 Issue: XI Month of publication: November 2025

DOI: https://doi.org/10.22214/ijraset.2025.75758

International Journal for Research in Applied Science & Engineering Technology (IJRASET)
ISSN: 2321-9653; IC Value: 45.98; SJ Impact Factor: 7.538
Volume 13 Issue XI Nov 2025- Available at www.ijraset.com

Survey on Multi-Agent and Multimodal Architectures for Intelligent Task Automation (AVIA Project)

Munaf Irfan Shaikh[1], Ayaan Badshah Khan[2], Hitesh Vinod Dadlani[3], Prof. Jayshri Kandekar[4]
Dept. Artificial Intelligence and Data Science, MET Institute of Engineering Nashik, Maharashtra, India

Abstract: The rapid evolution of Large Language Models (LLMs) and multi-agent systems (MAS) has paved the way for advanced Virtual Personal Assistants (VPAs) capable of performing complex, real-world tasks beyond simple, single- query responses. Traditional AI assistants are often limited in scope, lacking the deep integration, persistent memory, and adaptability required for cross-platform workflows, forcing users to rely on multiple tools. This survey examines the architectural and methodological shift toward Intelligent Task Automation, focusing on systems that leverage multimodal and multi-agent frameworks, exemplified by the AVIA (Autonomous Virtual Intelligent Assistant) project. We analyze core components including specialized agents orchestrated by a central planner (like n8n), the integration of LLMs for sophisticated Natural Language Understanding (NLU), and the use of multimodal capabilities (voice/image input, text/audio output). We explore key technical concepts, including the Transformer architecture and Retrieval-Augmented Generation (RAG) for conversational memory. The findings highlight the significant potential of multi- agent and multimodal systems to provide a unified, efficient, and context-aware solution for digital task automation, improving productivity and moving toward a more versatile Agentic AI future.

Index Terms: Survey, Large Language Models (LLMs), Multi- Agent Systems (MAS), Virtual Personal Assistants (VPAs), Multimodal Interaction, Intelligent Task Automation, Retrieval- Augmented Generation (RAG)

I. INTRODUCTION

The rapid growth of Large Language Models (LLMs) and intelligent automation has marked a significant shift in how humans interact with digital systems. As everyday workflows become increasingly complex and distributed across multiple platforms, the demand for autonomous, context-aware assistants continues to rise. Traditional virtual assistants remain limited, offering only single-query responses without persistent memory, deep tool integration, or adaptive reasoning capabilities. These constraints force users to manually coordinate tasks across communication apps, cloud services, and productivity tools — creating fragmentation and inefficiency. Recent advancements in multimodal LLMs and multi-agent architectures present a promising pathway toward highly capable, personalized, and autonomous virtual personal assistants (VPAs).

Modern assistant systems leverage agentic AI, where multiple specialized agents collaborate to perform complex tasks that go far beyond conversational interaction. Instead of a single monolithic model, multi-agent systems (MAS) employ domain-specific agents — such as planning agents, search agents, task executors, and summarizers — allowing for structured reasoning and division of labor. Coupled with LLM-driven Natural Language Understanding (NLU), these systems can interpret abstract user goals, decompose them into actionable steps, and execute tasks autonomously with high accuracy. Orchestration platforms like n8n act as central planners, enabling seamless integration with APIs, automation workflows, and third-party services, transforming VPAs from passive responders into fully operational digital workers.

Modern VPAs now incorporate multimodal capabilities, using inputs such as voice, text, images, and documents to enhance interaction richness. Multimodal LLMs enable the assistant to analyze visual data, understand spoken commands, and generate audio responses, creating a more natural and accessible user experience. Memory systems powered by Retrieval-Augmented Generation (RAG) further extend assistant intelligence by storing and recalling long-term contextual information. This enables persistent personalization, continuity across sessions, and improved decision-making, addressing a major limitation of traditional AI assistants. As a result, modern assistants evolve from simple query responders into deeply integrated, intelligent systems capable of supporting diverse real-world tasks.

This survey provides a comprehensive examination of the architecture, methodologies, and technological foundations behind AVIA (Autonomous Virtual Intelligent Assistant), a multi-agent, multimodal VPA designed for real-world task automation. The paper synthesizes contributions across LLM research, agentic frameworks, orchestration pipelines, and multimodal processing. It highlights key challenges such as real-time automation reliability, multi-agent coordination, memory management, and cross-platform integration. Finally, it outlines future directions toward more robust, scalable, and user-centric VPA ecosystems capable of autonomously handling the complexity of modern digital workflows.

II. LITERATURE SURVEY

The rapid advancement of generative AI, particularly Large Language Models (LLMs) and multi-agent architectures, has significantly influenced the evolution of Virtual Personal Assistants (VPAs). Recent research increasingly focuses on enhancing autonomy, contextual understanding, and cross- platform automation through multimodal interaction, memory- augmented systems, and orchestrated agent frameworks. This section provides a structured review of existing works, high- lighting key developments in LLM-driven assistants, multi- agent coordination, task automation workflows, and memory- augmented conversational systems.

A. *Systematic Reviews and Surveys*

Early surveys on intelligent assistants primarily examined rule-based chatbots and dialogue systems, identifying founda- tional limitations in scalability, contextual reasoning, and do- main adaptability. Traditional VPAs lacked persistent memory and relied heavily on static intents, resulting in constrained capabilities for multi-step reasoning or automated task execu- tion. More recent reviews categorize modern assistant systems based on their integration of LLMs, multimodal inputs, and autonomous decision-making. Studies highlight the shift from single-turn conversational models to agentic AI frameworks capable of planning, decomposing tasks, and interacting with external tools. Comprehensive analyses also emphasize the importance of modularity, cross-platform integration, and per- sonalization in next-generation assistants. User studies indicate increasing demand for assistants that deliver accurate informa- tion retrieval, reliable automation, and full workflow execution rather than simple responses.

B. *LLM-Based Systems and Classical Approaches*

Early virtual assistant technologies were dominated by symbolic AI, rule-based NLU pipelines, and keyword-driven systems. While effective for predefined tasks, such approaches suffered from limited generalization and poor handling of complex queries. The introduction of deep learning–based NLP models, such as RNNs and early Transformer variants, improved intent classification and dialogue generation, but still lacked global reasoning capabilities. With the emergence of large-scale Transformers such as BERT, GPT, and T5, assistants gained improved contextual understanding and the ability to generate coherent natural language responses. Mod- ern VPAs further leverage instruction-tuned LLMs capable of performing multi-step reasoning, tool use, and chain-of- thought–guided decision-making. These developments repre- sent a major leap over classical chatbot frameworks, enabling higher autonomy, more natural interactions, and the execution of increasingly complex commands.

C. *Multi-Agent Systems and Organizational Frameworks*

Multi-Agent Systems (MAS) have become an integral part of next-generation VPAs, offering improved modularity, parallelization, and specialization. Research demonstrates that distributing tasks across specialized agents — such as planning agents, retrieval agents, execution agents, and monitoring agents — enhances both reliability and task efficiency. Studies propose hierarchical agent structures where a central orches- trator or planner delegates subtasks to domain-specific agents, enabling decomposition of complex workflows into manage- able steps. Agent coordination has been widely explored in environments like AutoGPT, BabyAGI, MetaGPT, and LangChain Agents, each emphasizing different architectures for collaboration, memory retrieval, and tool usage. Emerging research also highlights the importance of human-in-the-loop feedback, adaptive agent prompting, and error-correction loops for increasing robustness in autonomous systems.

D. *Memory Systems and Retrieval-Augmented Approaches*

A critical limitation in many traditional VPAs is the lack of long-term memory and context persistence. Retrieval- Augmented Generation (RAG) has emerged as a dominant solution, combining embedding-based search with generative reasoning to maintain continuity across sessions. Studies show that RAG-enhanced assistants outperform purely generative models in factual accuracy, personalization, and long-term task management.

Memory systems such as vector databases (FAISS, Chroma), episodic and semantic memory modules, and hybrid memory architectures enable assistants to store user preferences, past interactions, and task outcomes. Several works also explore multimodal memory retrieval, where stored images, documents, or voice logs can be integrated into future conversation context. This advancement significantly increases the usefulness and reliability of autonomous agents operating across multiple platforms and switching tasks over extended periods.

E. Workflow Automation and Orchestration Tools

To support real-world task execution, modern assistant architectures increasingly integrate workflow automation en- gines like n8n, Zapier, and Make. Research identifies these tools as essential components for bridging LLM intelligence with external applications such as email, cloud storage, mes- saging platforms, and social media APIs. Studies show that automation frameworks improve reliability by enforcing struc- tured task flows, error-handling routines, and step-by-step agent coordination. Moreover, combining MAS with orches- tration platforms provides a scalable foundation for complex, multi-app tasks — such as scheduling events, generating reports, retrieving data, and managing digital content au- tonomously. This hybrid approach marks a major advancement over standalone LLM systems, allowing VPAs to evolve into intelligent digital workers capable of independent operation.

III. CHALLENGES

Despite major advancements in Large Language Models (LLMs), multi-agent architectures, and multimodal interaction, current VPA systems still face several challenges that limit their reliability, robustness, and real-world adoption. These challenges span across interaction quality, technical perfor- mance, memory consistency, and integration with external platforms.

A. Multimodal Interaction and Real-World Understanding

Although multimodal LLMs can process voice, text, and images, achieving accurate real-world understanding remains difficult. Assistant performance heavily depends on the quality of inputs—noisy audio, low-resolution images, or incomplete visual contexts often lead to misinterpretation. VPAs also struggle with grounding user queries to the correct context, particularly when switching between applications or handling ambiguous visual inputs. Ensuring consistent responsiveness across diverse environments (e.g., background noise, varied lighting, multiple speakers) remains a major challenge. Ad- ditionally, maintaining a natural conversational flow across modalities often requires highly optimized pipelines and model alignment strategies.

B. Technical Limitations and Performance Bottlenecks

Autonomous multi-agent systems require significant com- putational resources for planning, retrieval, execution, and monitoring tasks. Running multiple agents in parallel increases latency, especially when interacting with external APIs or performing complex reasoning. Workflow orchestrators like n8n introduce additional delays during multi-step automation. Furthermore, LLMs incur high inference costs, making contin- uous real-time operation expensive. Memory-driven systems (such as RAG pipelines) also suffer from embedding mis- matches, vector retrieval errors, and degraded performance when memory size grows. Ensuring reliable, low-latency exe- cution across platforms—while keeping system resource usage manageable—remains an ongoing technical challenge.

C. Integration, Security, and Reliability Concerns

Real-world digital automation requires deep integration with third-party applications such as email, cloud storage, messag- ing platforms, social media, and file systems. However, API rate limits, platform restrictions, inconsistent response formats, and authentication failures can disrupt task execution. VPAs must also handle failures gracefully, including malformed API responses, tool misfires, and unexpected workflow errors. From a security standpoint, ensuring safe delegation of tasks is critical—especially when the assistant accesses personal data, schedules events, or interacts autonomously with external services. Privacy risks increase when storing long-term user memories, requiring robust encryption and secure retrieval methods.

D. Accessibility, User Experience, and Adoption Barriers

Although AI assistants are becoming more capable, widespread adoption is hindered by usability challenges. Many users find multimodal systems overwhelming, especially when switching between voice, text, and visual interfaces. Long- term personalization is difficult to maintain due to inconsistent memory retrieval, forgetting past preferences, or providing overly generic responses.

Costs associated with high-quality LLMs, cloud inference, and API integrations also limit accessibility for users in low-income regions. Additionally, depen- dence on stable internet connectivity affects system reliability in rural or low-bandwidth areas, making fully autonomous operation difficult to guarantee.

IV. FUTURE DIRECTIVES

A. Multilingual and Cross-Cultural Generalization

Future development of multimodal VPAs must prioritize broader linguistic and cultural adaptability to support users across diverse regions. While current systems primarily op- erate in English, expanding capabilities to include multilin- gual comprehension and generation—especially for regional and low-resource languages—will significantly enhance ac- cessibility. This involves improving language coverage within LLMs, integrating localized datasets, and adapting models to account for cultural nuances in conversation and workflow expectations. Additionally, future prototypes should be eval- uated across varied digital ecosystems, communication styles, and user behaviors to ensure that autonomous agents remain robust, inclusive, and globally relevant.

B. Efficient and Scalable Architectures

As multi-agent and multimodal systems become increas- ingly complex, scalability and computational efficiency will be crucial for widespread deployment. Future research should explore lightweight LLM variants, agent optimization strate- gies, and caching mechanisms to reduce inference costs during continuous operation. Techniques such as model distillation, quantization, and hardware acceleration (e.g., GPU offloading, edge TPU support) may enable faster and more affordable agent execution. Moreover, hybrid cloud–edge architectures could be utilized, where heavy reasoning is performed in the cloud while routine automation runs locally. Such approaches will help ensure that VPAs like AVIA remain responsive, cost- effective, and suitable for long-term real-world use.

C. Context-Aware and Explainable Systems

Future VPAs must advance toward deeper contextual intel- ligence, enabling them to understand user intent, history, and environment with greater precision. Enhancing RAG-based memory to provide more reliable long-term personalization will be essential for maintaining continuity across sessions. Further integration of multimodal context—such as screen content, documents, images, and audio—can allow assistants to make more accurate decisions. Additionally, Explainable AI (XAI) techniques should be incorporated to help users understand why an agent made a specific decision or executed a particular action, increasing trust and transparency. Finally, research should explore proactive agents capable of predicting user needs, initiating tasks autonomously, and adapting behav- ior based on evolving user preferences.

V. CONCLUSION

The rapid evolution of Large Language Models (LLMs), multimodal processing, and multi-agent architectures has transformed the landscape of Virtual Personal Assistants (VPAs). This survey examined key advancements from early rule-based dialogue systems to modern agentic AI frameworks capable of reasoning, planning, and automating complex dig- ital workflows. While LLM-powered assistants significantly outperform traditional approaches in language understanding and contextual reasoning, they still face several limitations in scalability, reliability, and real-time decision-making across diverse platforms.

A major gap identified across current research is the limited ability of many VPAs to serve as fully autonomous, end-to-end task executors. Most systems excel at isolated functions—such as question answering, scheduling, or document analysis—but struggle to integrate these abilities into cohesive, long-duration workflows. Challenges such as latency in multi-agent coordi- nation, inconsistent memory retrieval, tool integration failures, and dependency on stable internet connectivity continue to hinder robust real-world deployment. Additionally, multimodal processing, while powerful, remains sensitive to noisy environ- ments, ambiguous images, and shifting user context.

Key areas such as long-term personalization, proactive task initiation, secure data handling, and cross-platform automation require further enhancement. Future progress will depend on developing more efficient LLM variants, optimizing multi- agent pipelines, and strengthening memory systems using advanced Retrieval-Augmented Generation (RAG) techniques. Improved orchestration frameworks, hybrid cloud–edge exe- cution models, and standardized interfaces will also play a critical role in enabling seamless automation. Ultimately, the future of VPAs like AVIA lies in combining technological innovation with user-centered design to create assistants that are not only intelligent and autonomous but also trustworthy, accessible, and adaptable to everyday digital environments.

REFERENCES

[1] W. S. Wong, H. Hamid-Aghvami, and S. Wolak, “Context-Aware Per- sonal Assistant Agent Multi-Agent System,” in Proc. Int. Conf., Oct. 2008.

[2] G. Cebula, A. M. Ghiran, I. Gergely, and G. S. Cojocaru, “IPA: An Intelligent Personal Assistant Agent for Task Performance Support,” in Proc. Int. Conf., Sep. 2009.

[3] S. Li, S. Wang, Z. Zeng, Y. Wu, and Y. Yang, “A Survey on LLM- Based Multi-Agent Systems: Workflow, Infrastructure, and Challenges,” Vicinagenth, vol. 1, no. 3, 2024.

[4] B. Li et al., “Beyond Self-Talk: A Communication-Centric Survey of LLM-Based Multi-Agent Systems,” arXiv preprint, 2025.

[5] A. K. Patil, “Agentic AI: A Comprehensive Survey of Technologies, Applications, and Societal Implications,” IEEE Access, 2025.

[6] C. Sharma, “Retrieval-Augmented Generation: A Comprehensive Survey of Architectures, Enhancements, and Robustness Frontiers,” 2025.

[7] M. M. Hasan et al., “Model Context Protocol (MCP) at First Glance: Studying the Security and Maintainability of MCP Servers,” 2025.

[8] Y. Du, “The Impact of Artificial Intelligence on People’s Daily Life,” The Frontiers of Society, Science and Technology, vol. 6, no. 6, pp. 12–18, 2024.

[9] P. M. G. Arias, “Disen˜o, Desarrollo e Implementacio´n de una Asistente Virtual para la Resolucio´n de Dudas sobre los Procesos Acade´micos de la Universidad Polite´cnica Salesiana,” Univ. Polite´cnica Salesiana, Ecuador, Tech. Rep., 2022.

[10] A. P. Mendoza et al., “NAIA: A Multi-Technology Virtual Assistant for Boosting Academic Environments—A Case Study,” 2025.

[11] S. D. Mishra, A. Dhiman, and D. Dhyani, “AI Assistant for Daily Use,” Int. J. Sci. Dev. Res. (IJSDR), vol. 9, no. 10, 2024.

[12] A. B. M. V. Shinde et al., “AI-Based Virtual Assistant Using Python: A Systematic Review,” Int. J. Res. Appl. Sci. Eng. Technol. (IJRASET), vol. 11, no. 3, pp. 814–818, 2023.

[13] A. B. Singh et al., “Automating Desktop Tasks with a Voice-Controlled AI Assistant Using Python,” Int. J. Research Publication and Reviews, vol. 5, no. 5, pp. 12615–12620, 2024.

[14] A. S. Reddy M., Vyshnavi, C. R. Kumar, and Saumya, “Virtual Assistant Using Artificial Intelligence and Python,” J. Emerging Technologies and Innovative Research (JETIR), vol. 7, no. 3, pp. 1116–1119, 2020.