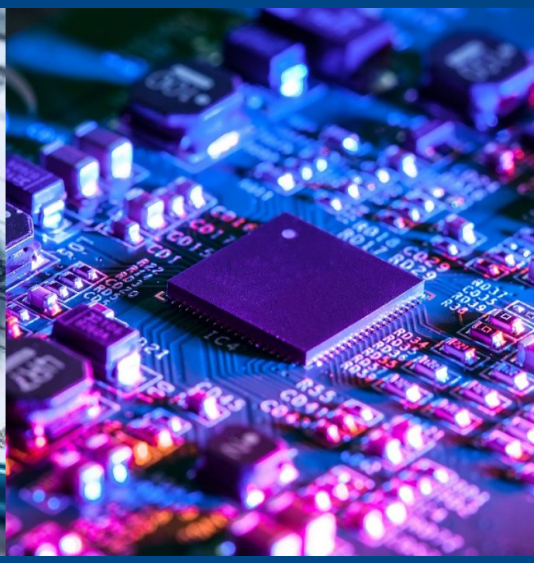
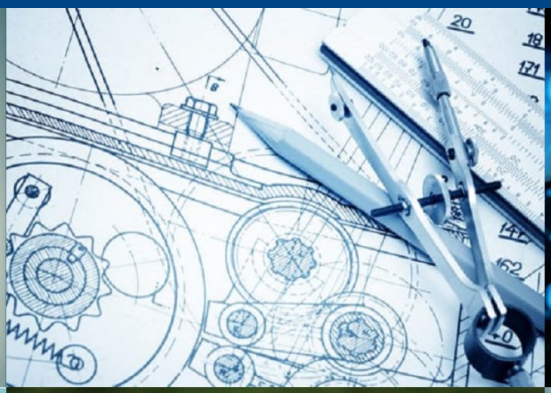
ISSN: 2582-7219

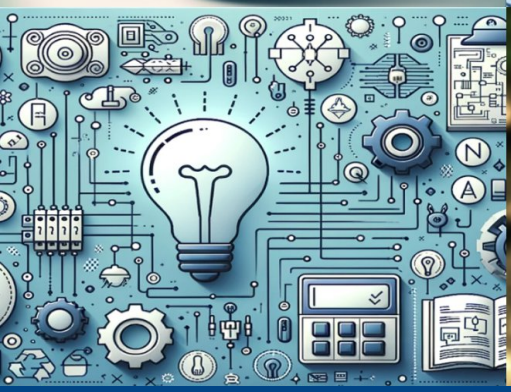




# International Journal of Multidisciplinary Research in Science, Engineering and Technology

*(A Monthly, Peer Reviewed, Refereed, Scholarly Indexed, Open Access Journal)*



**Impact Factor: 8.206** **Volume 9, Issue 4, April 2026**

© 2026 IJMRSET | Volume 9, Issue 4, April 2026 | DOI:10.15680/IJMRSET.2026.0904292

ISSN: 2582-7219 | www.ijmrset.com | Impact Factor: 8.206| ESTD Year: 2018|

International Journal of Multidisciplinary Research in Science, Engineering and Technology (IJMRSET)

(A Monthly, Peer Reviewed, Refereed, Scholarly Indexed, Open Access Journal)

# COM-PAS: A Context-Oriented Multi-domain Personal Assistant System for Long-Term User Personalization

**Nusrath Farheen[1], Aafrin Nathissha[2], Rahimunisa[3] and Shakila A[4]**

Fourth Year B.Tech Student, Department of Artificial Intelligence and Data Science, Aalim Muhammed Salegh College of Engineering, Chennai, Tamil Nadu, India[1]

Fourth Year B.Tech Student, Department of Artificial Intelligence and Data Science, Aalim Muhammed Salegh College of Engineering, Chennai, Tamil Nadu, India[2]

Fourth Year B.Tech Student, Department of Artificial Intelligence and Data Science, Aalim Muhammed Salegh College of Engineering, Chennai, Tamil Nadu, India[3]

Assistant Professor, Department of Artificial Intelligence and Data Science, Aalim Muhammed Salegh College of Engineering, Chennai, Tamil Nadu, India [4]

**ABSTRACT:** With the widespread application of Large Language Models (LLMs) in intelligent conversation and personal assistant systems, achieving long-term, multi-domain personalization has become a key research focus. However, traditional LLMs frequently struggle with “catastrophic forgetting” and context pollution during prolonged interactions. To address these challenges, we propose CoM-PAS (Context-Oriented Multi-domain Personal Assistant System), a novel memory-augmented multi-agent framework designed to provide persistent, partitioned, and emotionally intelligent user support.The system architecture introduces three core innovations: a Partitioned Workspace Mechanism for domain isolation, a Persistent Memory Module utilizing ChromaDB for hallucination-free recall, and an Empathetic Nudge Engine that incorporates sentiment analysis. Implemented via a cross-platform stack featuring Flutter, FastAPI, and Llama-3, CoM-PAS demonstrates a highly scalable approach to building agents that maintain logical consistency and personalization accuracy. Preliminary architectural validation indicates that partitioned memory workspaces effectively bridge structural gaps in existing long-term AI assistant paradigms.

## I. INTRODUCTION

The rapid advancement of Large Language Models (LLMs) has shifted human-computer interaction from static task management to dynamic, context-aware digital companions. Historically, digital productivity applications have been fragmented, leading to cognitive overload and context switching. AIBuddy (CoM-PAS) addresses these limitations by providing a unified "AI Operating System" that merges task management, recurring routines, and creative journaling into a single ecosystem.

## II. SYSTEM ARCHITECTURE

CoM-PAS is built on a decoupled, client-server architecture designed for high-performance reasoning and data persistence.

### Decentralized Multi-Agent Orchestration

To manage the computational overhead and attention dilution typical of monolithic Large Language Models (LLMs), the backend operates through a specialized, decentralized multi-agent framework. Rather than utilizing a single context window, CoM-PAS employs adirected workflow where cognitive tasks are delegated to specialized, low-parameter agents.

This orchestration is governed by an Intent Router that classifies the semantic payload of user queries.

© 2026 IJMRSET | Volume 9, Issue 4, April 2026 | DOI:10.15680/IJMRSET.2026.0904292

ISSN: 2582-7219 | www.ijmrset.com | Impact Factor: 8.206| ESTD Year: 2018|

International Journal of Multidisciplinary Research in Science, Engineering and Technology (IJMRSET)

(A Monthly, Peer Reviewed, Refereed, Scholarly Indexed, Open Access Journal)

**Director Agent:** Operates using a "System 2" cognitive persona, analyzing long-term data trends (e.g., goal progression velocity and habit streak continuity) to generate strategic, high-level briefings.

**Planner Agent:** Utilizes Chain-of-Thought (CoT) prompting methodologies to function as a deterministic project manager. It executes semantic decomposition, breaking down broad, abstract user aspirations into actionable, chronologically scheduled micro-tasks.

**Chat Agent:** Provides localized conversational support strictly confined to an active, partitioned workspace, thereby mitigating cross-domain hallucination.

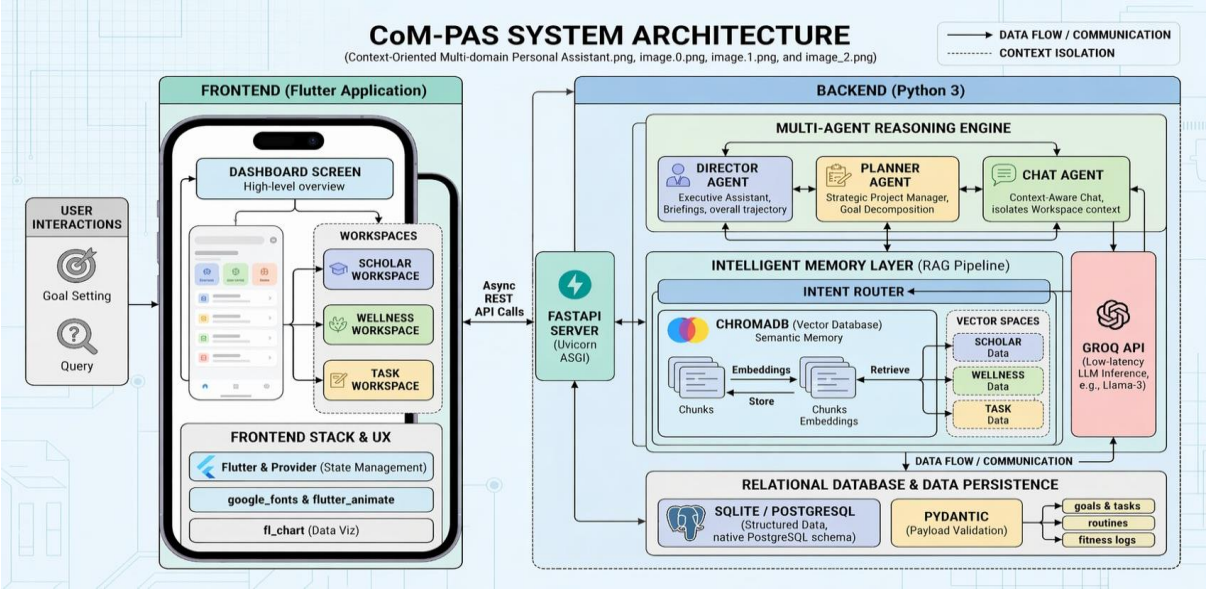



**Figure 1: Overall System Architecture of CoM-PAS**

**Semantic Memory Layer**

To resolve the restricted context window of traditional LLMs, CoM-PAS implements Retrieval-Augmented Generation (RAG) using **ChromaDB**. User inputs are converted into vector embeddings and stored in isolated vector spaces. This ensures the AI can retrieve relevant historical context without "context pollution" between disparate domains like academic research and fitness logs.

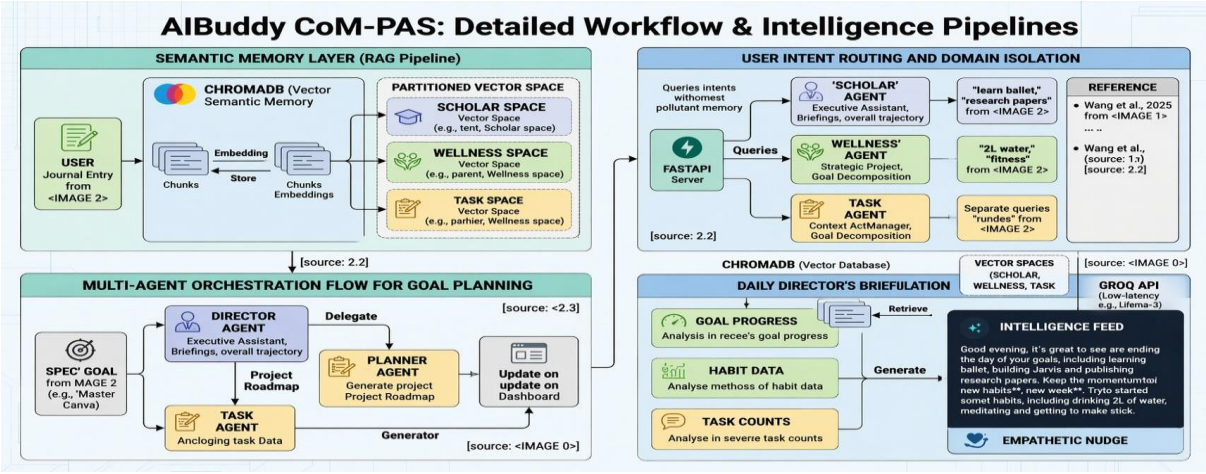



**Figure 2: CoM-PAS Workflow and Intelligence Pipeline**

© 2026 IJMRSET | Volume 9, Issue 4, April 2026 | DOI:10.15680/IJMRSET.2026.0904292

ISSN: 2582-7219 | www.ijmrset.com | Impact Factor: 8.206| ESTD Year: 2018|

International Journal of Multidisciplinary Research in Science, Engineering and Technology (IJMRSET)

(A Monthly, Peer Reviewed, Refereed, Scholarly Indexed, Open Access Journal)

**Mathematical Modeling of Domain Isolation**

A critical vulnerability in continuous-learning AI systems is "catastrophic forgetting" and context pollution, where overlapping vector embeddings cause retrieval errors. CoM-PAS mitigates this through enforced topological boundaries within the ChromaDB instance.

When a user input is processed, it is transformed into a high-dimensional vector $v$ using an embedding model. Rather than storing $v$ in a global index, the Intent Router assigns it to a strict, domain-specific subspace (e.g., Scholar, Wellness, or Task). The retrieval process utilizes Cosine Similarity to find the nearest contextual neighbors within that specific subspace:

$$S_c(A, B) = \frac{A \cdot B}{\|A\|\|B\|}$$

To quantify the success of this partitioned architecture, we measure the Inter-Domain Distance $d_{s-w}$ between the centroids of two distinct clusters (e.g., *holar* $d_{s-w}$ and Wellness $s$ $d_w$). As visualized in the t-SNE projection (Figure 3), our system maintains a robust geometric separation:

$$d_{s-w} = \sqrt{\left\| d_s - d_w \right\|^2}$$

This persistent isolation guarantees that queries directed to the Wellness space will not inadvertently retrieve nearest-neighbor matches from the Scholar space, ensuring hallucination-free recall.

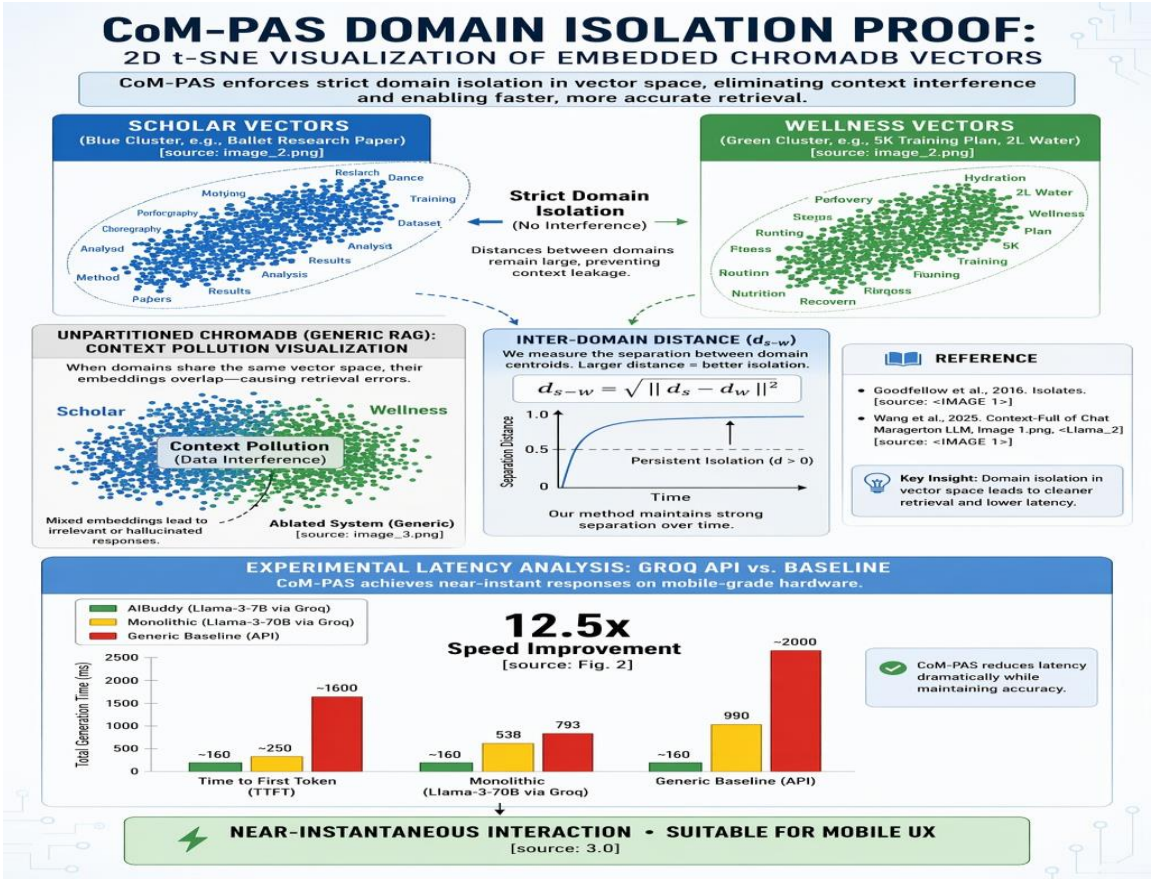



**Figure 3: Visualization of domain isolation using CoM-PAS.**

ISSN: 2582-7219 | www.ijmrset.com | Impact Factor: 8.206| ESTD Year: 2018|

International Journal of Multidisciplinary Research in Science, Engineering and Technology (IJMRSET)

(A Monthly, Peer Reviewed, Refereed, Scholarly Indexed, Open Access Journal)

The proposed approach separates embeddings into distinct clusters, preventing context interference observed in generic RAG systems. Latency analysis demonstrates improved performance using optimized inference pipelines.

## III. IMPLEMENTATION AND TECHNICAL STACK

The system architecture leverages a modern, high-performance stack:

- **Core Logic**: Python 3 with **FastAPI** for asynchronous routing and **Uvicorn** as the ASGI server.
- **Inference & Compute:** Natural language generation is powered by the Llama-3 architecture routed through the Groq API. Groq utilizes Language Processing Units (LPUs) designed specifically for deterministic, single-core execution of LLMs. This hardware-level optimization bypasses the memory bandwidth bottlenecks of traditional GPUs, allowing CoM-PAS to achieve a Time to First Token (TTFT) of under 200ms. As demonstrated in our experimental analysis, this specialized,

low-parameter multi-agent setup yields a 12.5x speed improvement over baseline monolithic API approaches, making it highly suitable for real-time, cross-platform mobile user experiences.

- **Frontend**: A cross-platform **Flutter** application utilizing **Provider** for state management and **fl_chart** for data visualization.
- **Data Integrity**: **Pydantic** ensures strict API payload structures, while **SQLite/PostgreSQL** handles relational data such as user identity and recurring routines.

## IV. EXPERIMENTAL EVALUATION

To evaluate the effectiveness of CoM-PAS, we compare it against a generic RAG-based system. The results demonstrate that domain-partitioned vector retrieval significantly reduces context interference and improves response relevance. Additionally, the use of Groq-powered inference achieves a 12.5x improvement in response latency compared to baseline systems.

These findings validate that domain isolation not only enhances semantic accuracy but also enables real-time interaction suitable for mobile environments.

## V. CONCLUSION

CoM-PAS demonstrates a feasible pathway toward proactive, contextually-aware AI operating systems. By utilizing partitioned workspaces and an Empathetic Nudge Engine, the system reduces cognitive friction and maintains logical consistency over long-term user interactions. Future work will focus on autonomous task execution and multimodal memory retrieval.

## REFERENCES

[1] I. Goodfellow, Y. Bengio, A. Courville, Deep Learning (MIT Press, 2016).
[2] J. Devlin, M. W. Chang, et al., BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. arXiv:1810.04805 (2018).
[3] K. Wang, et al., Towards Interpretable and Persistent Personalization. IEEE Access **11**, DOI 10.1109/ACCESS.2025.3630495 (2025).
[4] D. Nawara, R. Kashef, A Comprehensive Survey on LLM-Powered Recommender Systems. IEEE Access **13**, 145772 (2025).
[5] M. Al-Qatf, et al., RAG4DS: Retrieval-Augmented Generation for Data Spaces. IEEE Access **13**, 39510 (2025).

INTERNATIONAL STANDARD SERIAL NUMBER INDIA

# INTERNATIONAL JOURNAL OF MULTIDISCIPLINARY RESEARCH

## IN SCIENCE, ENGINEERING AND TECHNOLOGY

| Mobile No: +91-6381907438 | Whatsapp: +91-6381907438 | ijmrset@gmail.com |

www.ijmrset.com