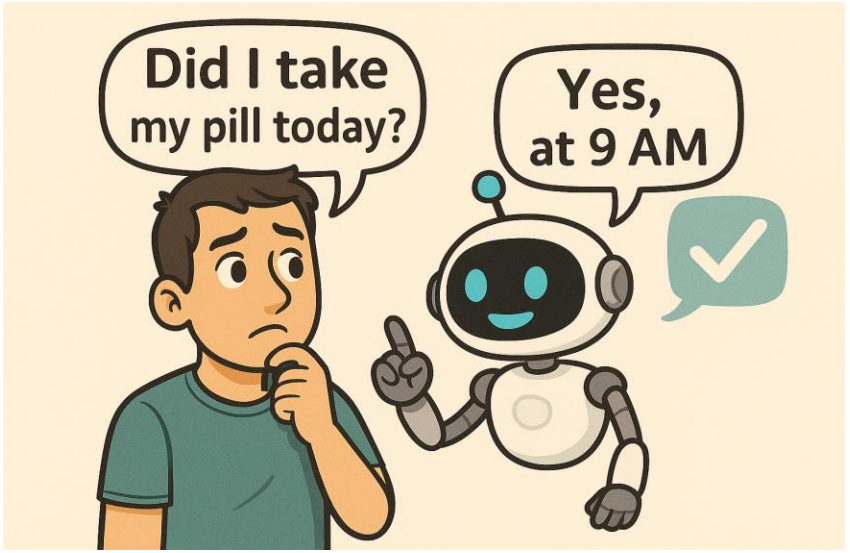
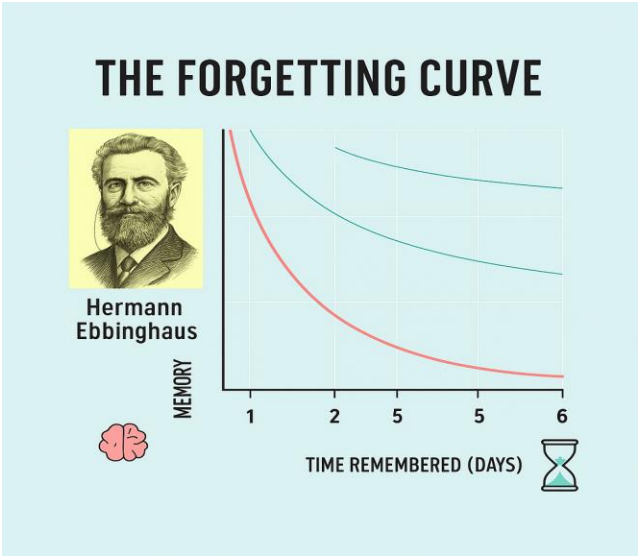


Human memory is fragile!

Adhwaa Alchaab & Pavan Kumar

Rutgers University, ECE and CS Departments.



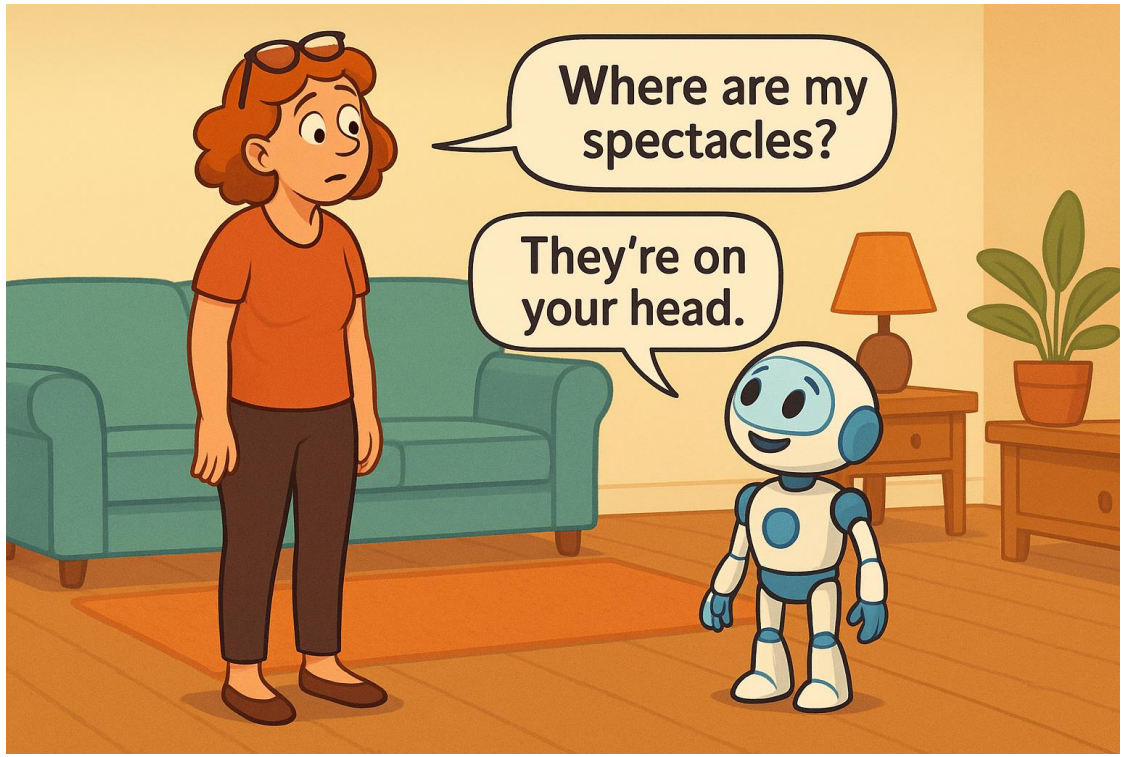
MEMENTO: Memory-Enhanced Modular Entity for Noticing, Tracking, and Organization

- **People forget ~50% of new information within an hour and 70% within a day unless reinforced.**

Ebbinghaus, H. (1885). Memory: A Contribution to Experimental Psychology.

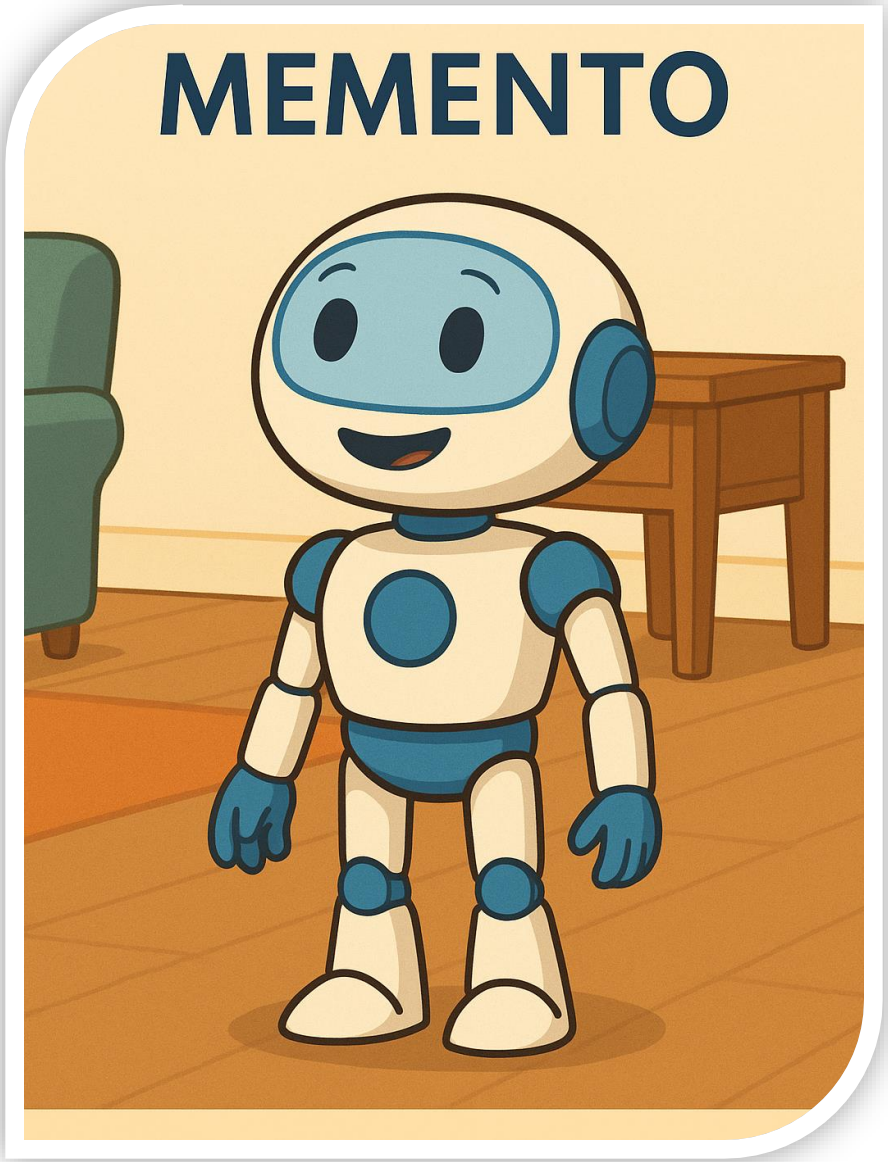
- **62% forgot to take a medication.**

Gadkari & McHorney, BMC Health Services Research, 2012



Problem statement

- **Prospective memory is remembering to perform a task in the future (e.g., taking medicine, attending a meeting). It's highly prone to failure, especially in stressful or multitasking environments.**
- Einstein, G. O., & McDaniel, M. A. (1990). Normal aging and prospective memory. Journal of Experimental Psychology
- Existing reminder apps are passive, require manual input, and lack real-time context or physical presence
- **Need**: an adaptive, context-aware assistant that proactively supports memory in everyday life



Introduction

MEMENTO: Memory-Enhanced Modular Entity for Noticing, Tracking, and Organization

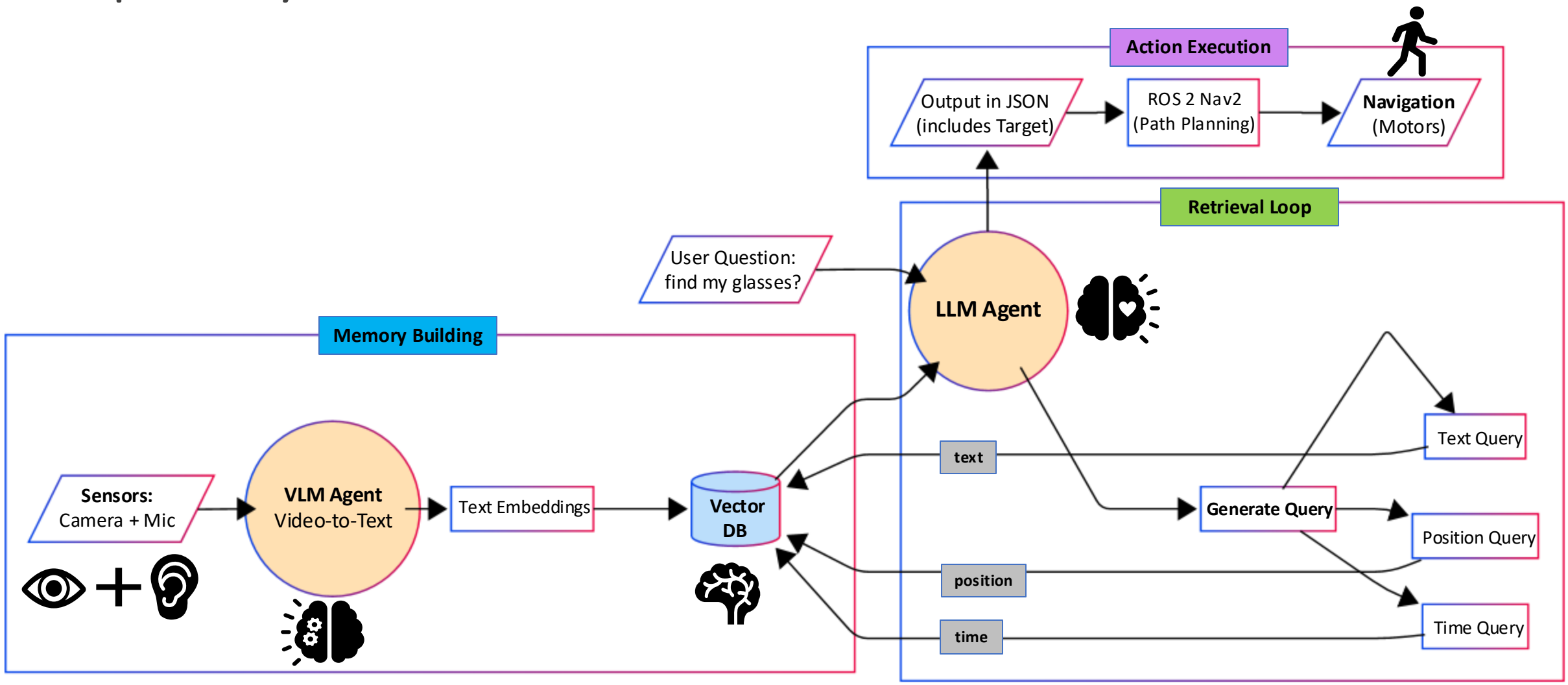
Designed as a socially cognizant assistive robot for real-time memory support

Integrates speech, behavior, and environmental data into a structured, quarriable memory

Supports natural dialogue and context-aware retrieval of past events and actions

Built with privacy, transparency, and user agency at its core

Proposed System: MEMENTO Robot



On-Chip AI Models for Memory, Retrieval & Navigation

The On-Device Models that are being deployed and tested by Nvidia from the research paper "ReMEmbR" are the Following:

Memory Building:

- **Video Captioning:** Quantized VILA-3B
- **Text Embedding:** mxbai-embed-large-v1

Retrieval Loop (LLM Agent):

- Llama3-8B | Qwen-8B | Llama 3.2-3B

Action Execution:

- Whisper for speech recognition
- ROS 2 Nav2 Stack with AMCL using 3D LiDAR

On-Device Vector Database:

- Quantized approximate nearest-neighbor (ANN) vector DB for fast similarity search over caption embeddings + timestamp & pose vectors

Deployment setup

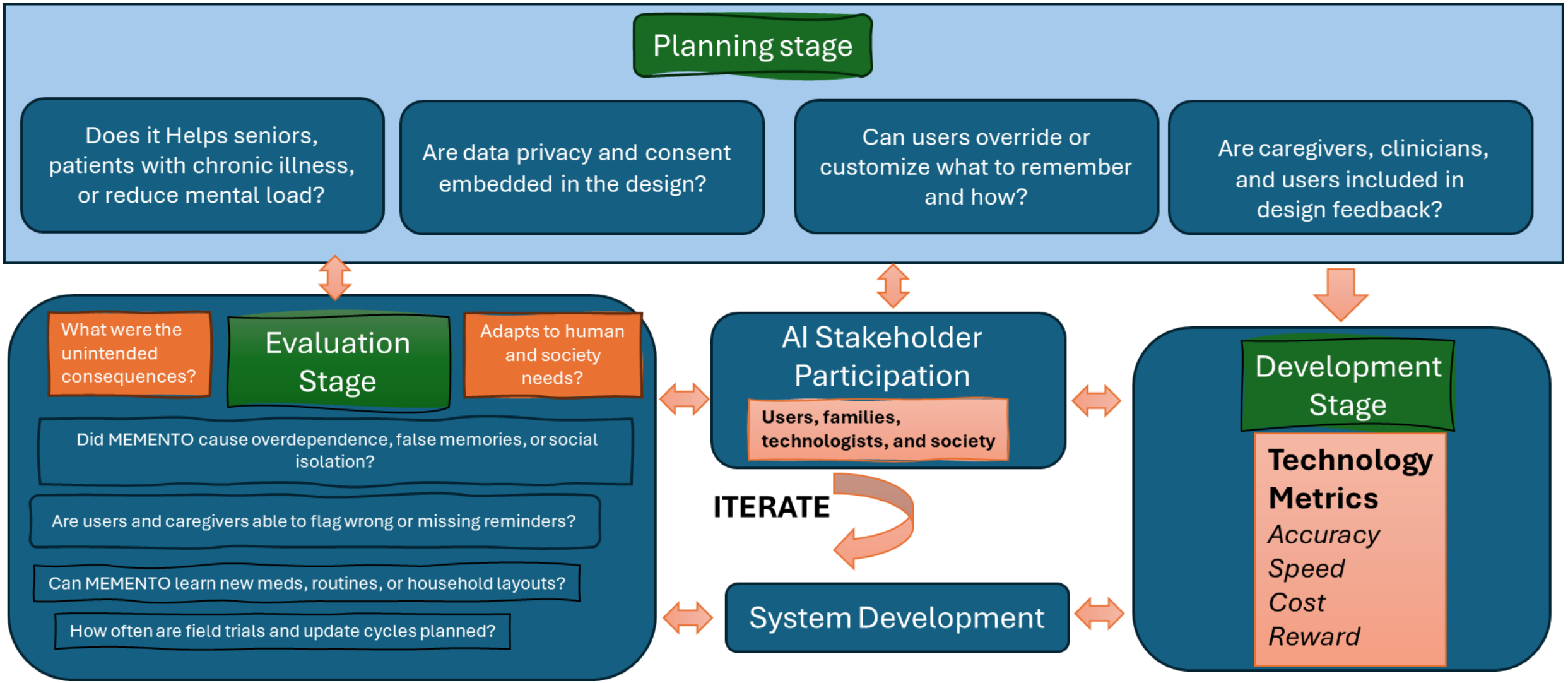
We deployed MEMENTO on a Linux workstation running Ubuntu 22.04.

The system has an RTX 4090 GPU, a 32-core CPU, and 68GB of RAM.
Our software stack includes:

- we benchmarked **3 lightweight models** (LLaMA 3, LLaMA 3.2, and Qwen3) for memory recall accuracy and reasoning ability
- **MilvusDB** for memory indexing (Docker image)
- A fully Python-based environment
- We design a lightweight **Gradio GUI** for user interaction

Each memory entry is structured using a MemoryItem class that includes a caption, timestamp, and spatial data. These entries are indexed in Milvus for retrieval.

Socially Cognizant Design Framework for MEMENTO: A Memory-Assistive Robot



Socially Cognizant Design

Privacy & Control

-Users mark private zones and delete any memory at will

-All data stays encrypted and stored locally on the robot

Clear Consent

-Before recording, MEMENTO asks for permission

-Users can opt in or out for specific tasks or locations

Socially Cognizant Design

Co-Design with Stakeholders	Workshops with elders, caregivers, and clinicians to gather needs	Rapid prototypes tested and refined based on real user feedback
Defined roles:	developers build, operators deploy, guardians oversee Ongoing	Easy buttons to correct or remove any entry
Evaluation:	Regular safety checks and bias scans on the system. Field studies measure success rates and user satisfaction.	Long-term trials to track cognitive and emotional impact

Tech-Only vs. Socially Cognizant MEMENTO

Aspect	Tech-Only	Socially Cognizant
Focus	Speed, accuracy	Ethics, inclusion, trust
Risk	Overdependence, privacy loss	Mitigated via user controls and transparency
Data Control	Centralized, limited control	Override, delete, audit memory entries
Inclusivity	One-size-fits-all	Multilingual, accessible, culturally aware
Trust	Assumed intelligence, false reliance	Transparent behavior and explainability

Conclusion

- MEMENTO reimagines reminders as an active, socially aware memory partner
- Integrates spatio-temporal embeddings, multimodal sensing, autonomous navigation, and LLMs
- Bridges symbolic and sub-symbolic representations for dynamic recall and anticipatory reasoning
- Ethical, user-centric design ensures privacy, consent, and inclusive accessibility

Future Work

Develop adaptive retrieval algorithms that account for individual affective states, habitual patterns, and memory retention profiles to deliver contextually tailored prompts.

Migrate core inference pipelines to embedded hardware, optimizing for low latency, power efficiency.

Continual and Reinforcement Learning through online learning strategies that leverage user feedback to refine interaction policies and memory retrieval over time.

DEMO

Box Link to the Short Demo(Video):

Link: https://rutgers.box.com/s/gmxb0laog7e0gjsvnw4ol342tl65l063

References

1. Einstein, G. O., & McDaniel, M. A. (1990). Normal aging and prospective memory. Journal of Experimental Psychology
2. Grattafiori, Aaron, et al. "The llama 3 herd of models." *arXiv preprint arXiv:2407.21783* (2024).
3. A. Asai, Z. Wu, Y. Wang, A. Sil, and H. Hajishirzi, "Self-RAG: Learning to Retrieve, Generate, and Critique through Self-Reflection," arXiv preprint arXiv:2310.11511, 2023.
4. Anwar, Abrar, et al. "Remembr: Building and reasoning over long-horizon spatio-temporal memory for robot navigation." *arXiv preprint arXiv:2409.13682* (2024).
5. Dana, Kristin J., et al. "Socially cognizant robotics for a technology enhanced society." *arXiv preprint arXiv:2310.18303* (2023).