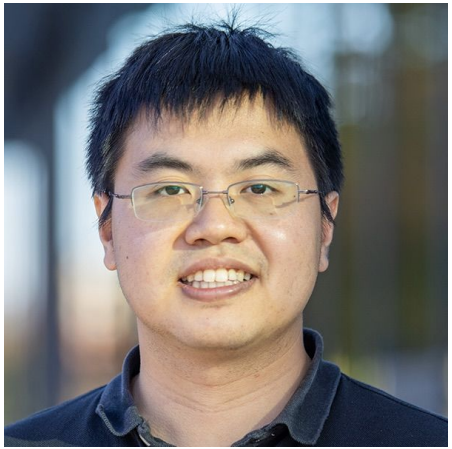
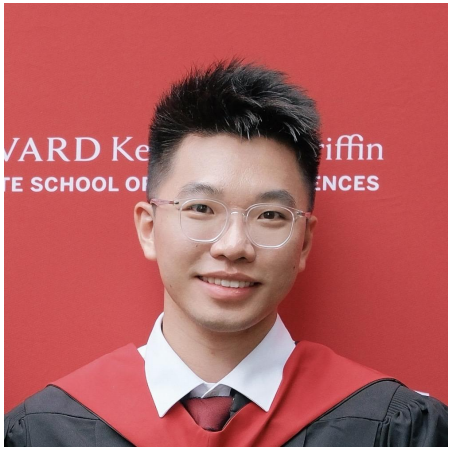
Inside Out

Decentralized Multi-agent Memory

Courtesy of Harvard Embodied Minds Lab



Yilun Du



Zhenting Qi

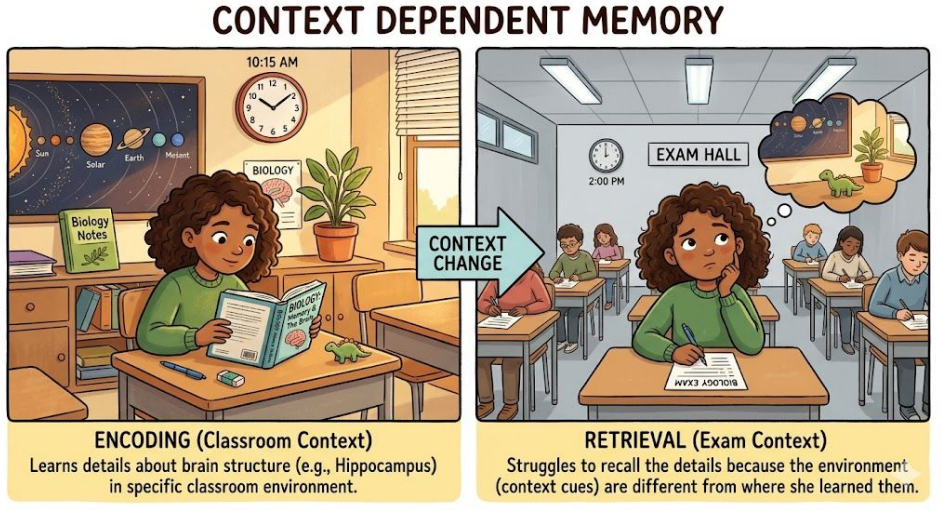
Jack Fan

I. What do we take from human memory?

What do we take from human memory?

Memory is...

- **context-dependent**
- multifaceted
- distilled/selective



What do we take from human memory?

Memory is...

- context-dependent
- **multifaceted**
- distilled/selective

What do we take from human memory?

Memory is...

- context-dependent
- multifaceted
- **distilled/selective**

II. Existing limitations

Existing limitations

- **Lack of portability**: Memory is overfit to the system it operates in.
- **Monolithic:** in a world of perfect precision and recall, it’s hard to find the needle in the haystack; also, it’s hard to synthesize heterogeneous cross-domain associations

III. Our design

(inspired by Inside Out, all credits to Disney-Pixar)

A Multiagent Memory System

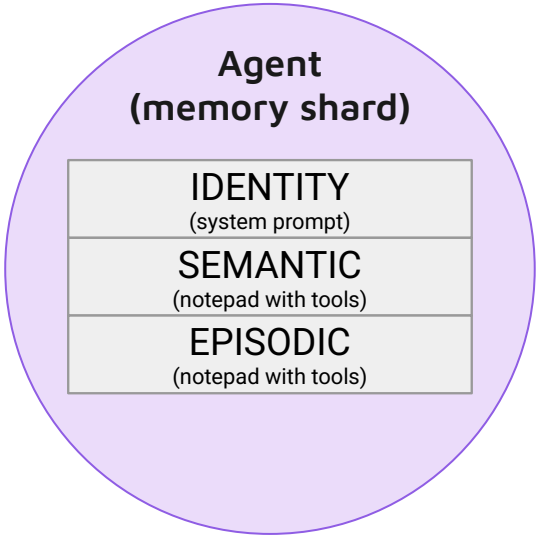
- **Decentralization**: hierarchical organization and heterogeneous construction of memories
- **Modularity:** different agents have unilateral control over different aspects (“shards”) of memory
- **Dynamism:** adaptability to different domains and simple memory management (create, update, delete)

Example use case: Education

- **Multiple learning settings:** where/when each student is learning; preparing exams, working on research projects, etc.
- **Capturing learning preferences:** “I learn by doing with hands-on examples”, “I need clear summaries of lectures”
- **Cross-session:** surfacing new connections across previously siloed projects (shoutout exocortex)

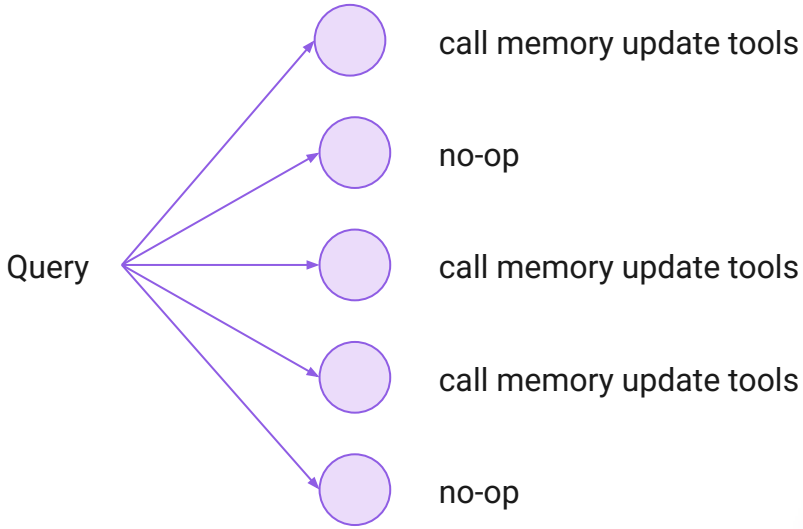
System Architecture

Atomic unit: Agent with a “Memory Shard”



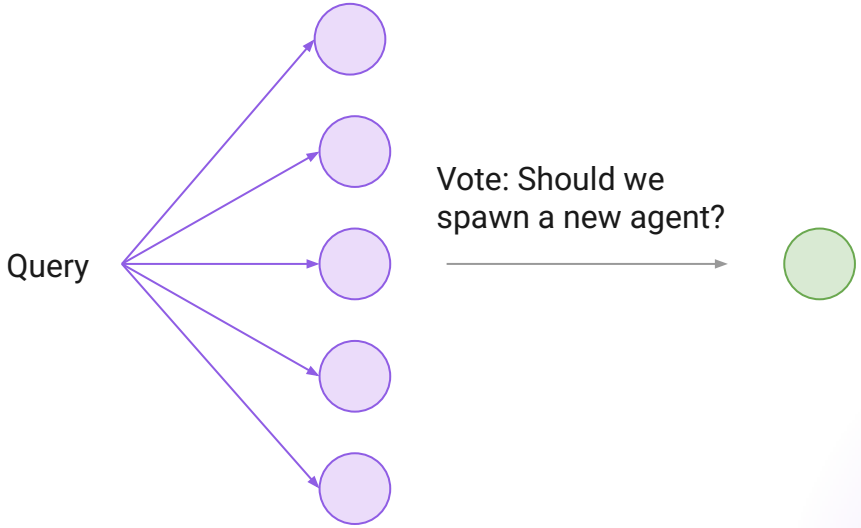
System Architecture

Phase 1: User input + agents (maybe) update memory



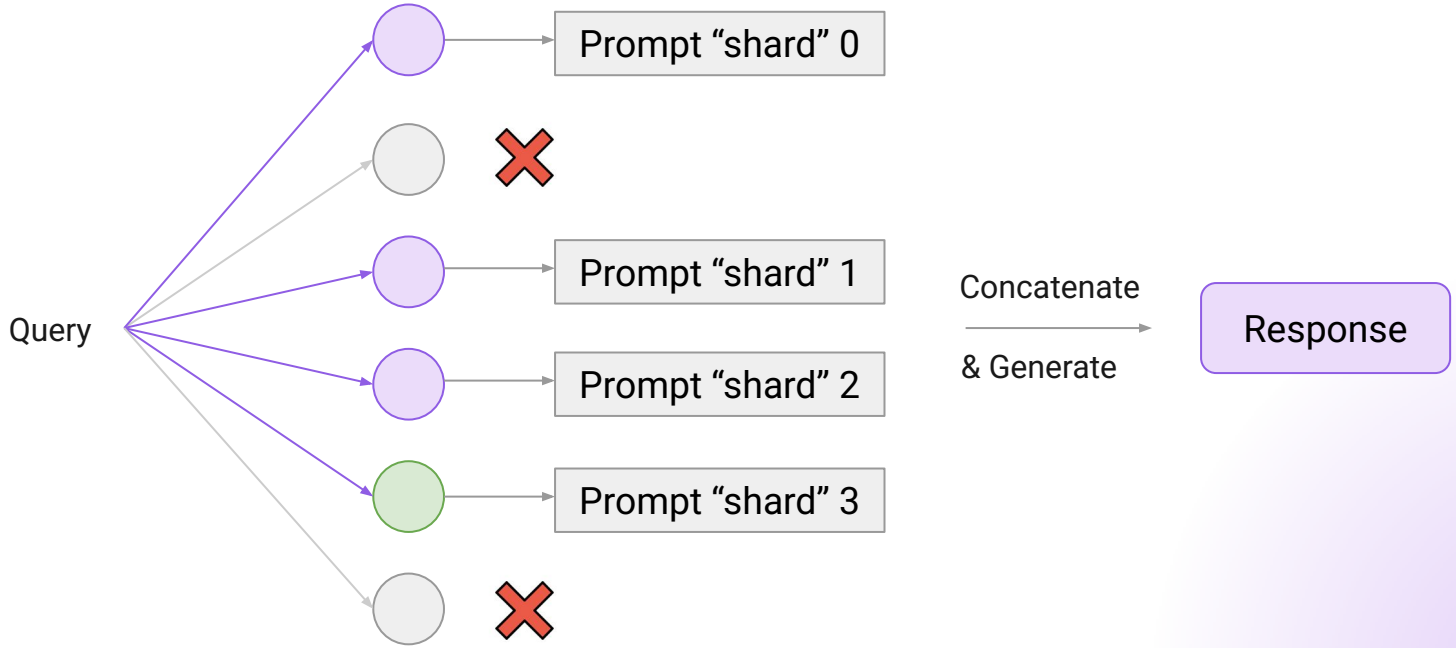
System Architecture

Phase 2: Voting for Spawning New Agent



System Architecture

Phase 3: Assembling (Activated) Shards



DEMO!

Reflections

- Governance considerations
- Portability and interoperability
- Scalability pathways
- Pathways to real-world piloting

IV. Future work

Future work: experimental directions

- **“The Society of Mind”:** forming hierarchical structure between shards, dynamic deconstruction of memories
- **Self-assembly:** different memory “shards” can evolve, specialise, conglomerate and work together; think stem cell-like growth (or microbots in *Big Hero 6*)
- **Continual learning:** weight-space (RL/LoRA/etc.) and token-space (context management)

Future work: guiding principles

- **Generalizability:** extensibility + composability → effective memory system for any domain + across domains
- **Efficiency:** modularity + composability = smaller models, portable and edge-viable systems

Thank you!
Questions?