Core of Multi-Agent Systems

Xiaming Chen，道夕

08/01/2025

Agenda

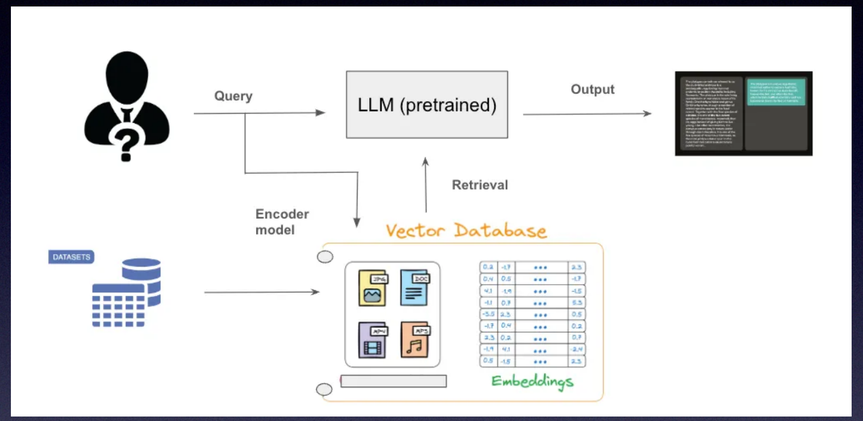
- Background
- Intro. to Agentic AI
- Multi-agent system (MAS)
- Advanced Topics

Background

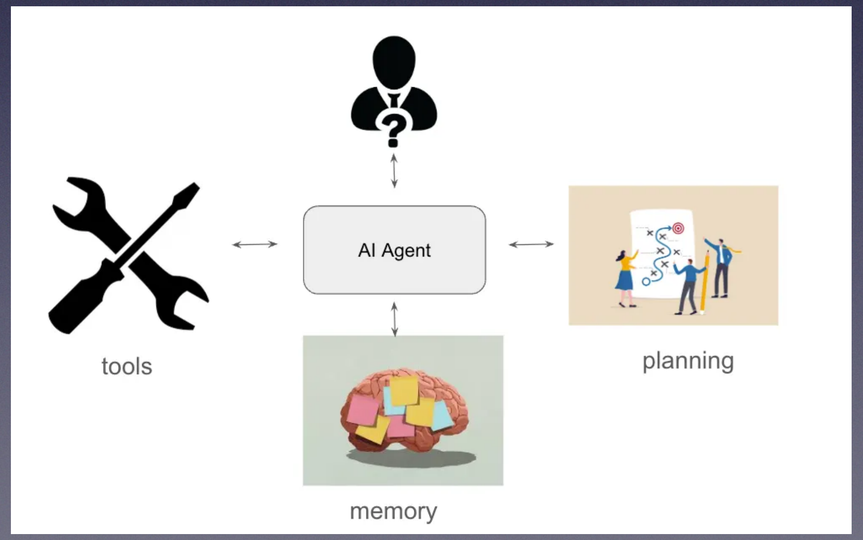
- Preliminaries: Know what is a large language model
- Personal research area: cognitive computation
- This talk focuses on:
  - Core techs of multi-agent systems
  - How to build a MAS

From RAG to AI Agent

- Both built on LLM
- Core of RAG
  - Vector store & Embed model
- Core of AI Agent: **ReAct**
  - Planning (orchestration)
  - Tools (e.g., MCP)
  - Memory (short & long term)



RAG



Agent

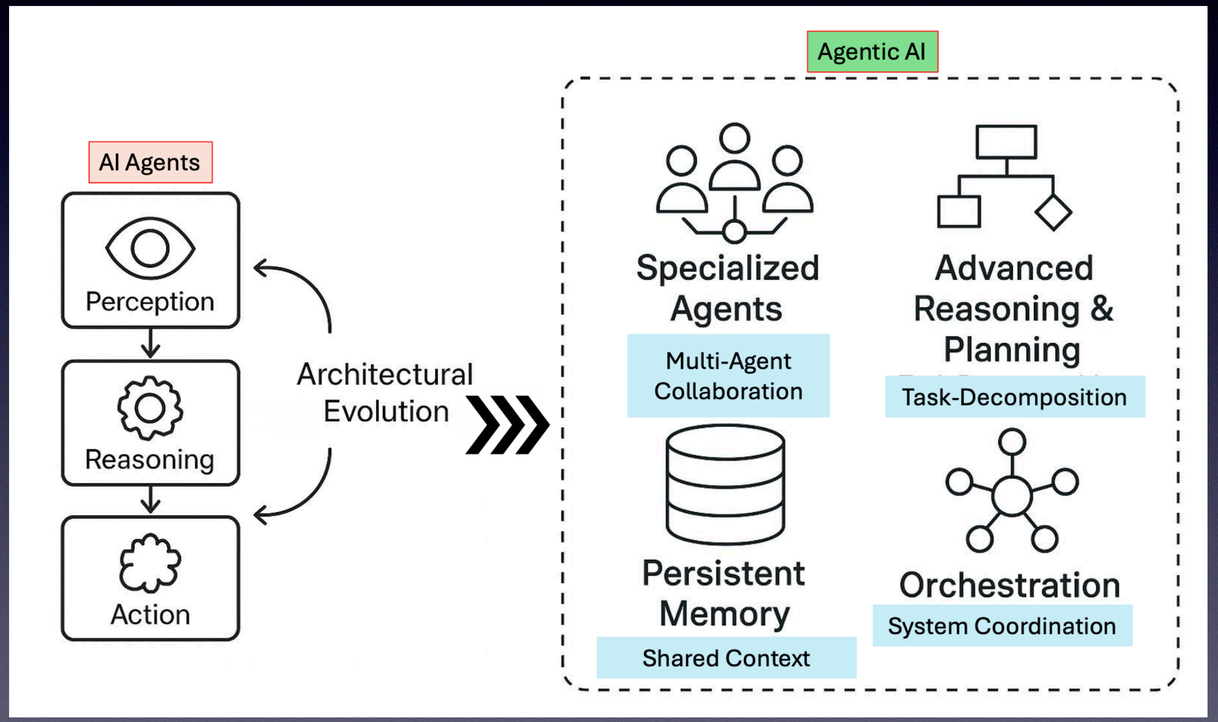
ReAct: Synergizing Reasoning and Acting in Language Model, ICLR 2023

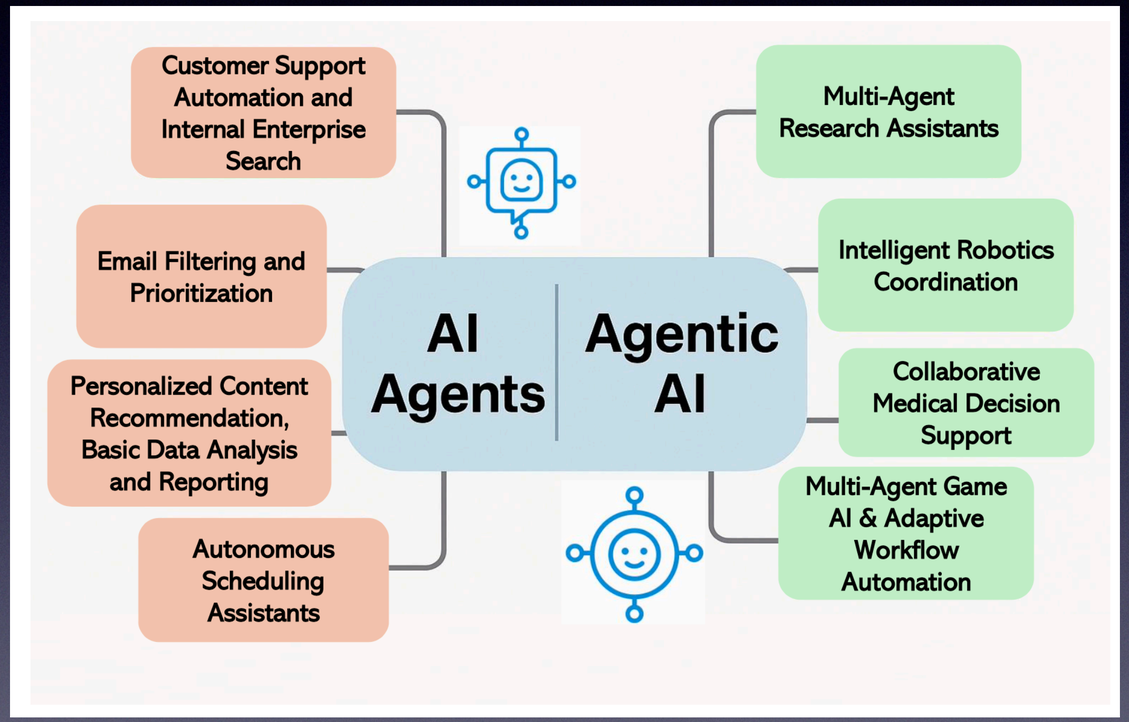
An informal def.

- Agent: a system that uses an LLM to decide (reasoning) the control flow (action) of an application.
  - In brief, a piece of code powered by LLM
- Multi-agent system: a computerized system comprising multiple intelligent agents that interact within a shared environment (context).
  - Not even new

Advances and Challenges in Foundation Agents: From Brain-Inspired Intelligence to Evolutionary, Collaborative, and Safe System, arXiv, 2025

AI Agent vs. Agentic AI





Agentic AI -> Multi-agent system

AI Agents vs. Agentic AI: A Conceptual Taxonomy, Applications and Challenges, arXiv, 2025

2025 AI Agents Infrastructure Stack

Powered by Agency AI + AgentOps

PLATFORM

PaaS/BaaS
fly.io · griptape · Render · netlify · NEON · supabase · Vercel · Railway · AXIOM

Logging
Grafana · RAYGUN

Observability, Tracing, and Evaluation
AgentOps · Metoro · LangSmith · Langfuse · braintrust · Patronus AI · COVAL · opik

Agent Frameworks
AgentStack · LangGraph · crewai · AG · Boundary · AG2 · CAMEL-AI · ControlFlow · LlamaIndex · Prolson · smolagents · OpenAI Agents SDK · Microsoft

ORCHESTRATION

Persistence
inngest · hatchet · Trigger.dev · Temporal

Agent Routing
LangGraph · crewai · Letta

Model Routing
Martian · arcee.ai

DATA

Memory
cognee · mem0 · zep

Storage
NEON · supabase · Pinecone · chroma · Weaviate · mongoDB · Fireproof · MotherDuck · neo4j · qdrant · tinybird

ETL
LlamaIndex · reducto · DATAVOLO · verodat

TOOLS

AGENTS AS A SERVICE

Search
sonar · exa · Serper · glean · meilisearch · Search1API · Tavily

Data Extraction
Parallel · Firecrawl · TINY FISH · Browse AI · oxylabs · NIMBLE · bright data

UI Automation
Browser Use · Browserbase · bytebot · LaVague · AGI,inc · notte · OS-ATLAS · Open Interpreter · HyperWrite · Anthropic Computer Use · OpenAI Operator · Google Project Mariner

Payments
Open Commerce · payman · Skyfire · protegee · Stripe Agent SDK

Secure Tool Usage
composio · ARCADE · LAYER · Paragon · mcprun · Unified · QL · wildcard · Toolhouse · glean

Auth
Auth0 · clerk · ANON · okta · OpenFGA · authzed

Browser Infrastructure
Browserbase · Anchor browser · Browserless · APIFY · CLOUDFLARE · platform.sh · Browser Use

Sandboxes
E2B · MODAL · CodeSandbox · Pig · SCRAPYBARA · CLOUDFLARE · RIZA · >>>ForeverVM · Daytona · WebContainers

AGENTS

Next Gen Copilots
perplexity · gradial · Cleric · glean · Canopy

Agent Teammates
Astral · bolt.new · Common Room · DEVIN · Dropzone AI

Agent Swarms
aaru · SOCIETIES

500+ Agents

marketplace.agen.cy

Powered by Agency AI + AgentOps

Source
substack.com/@jonturow & marketplace.agen.cy

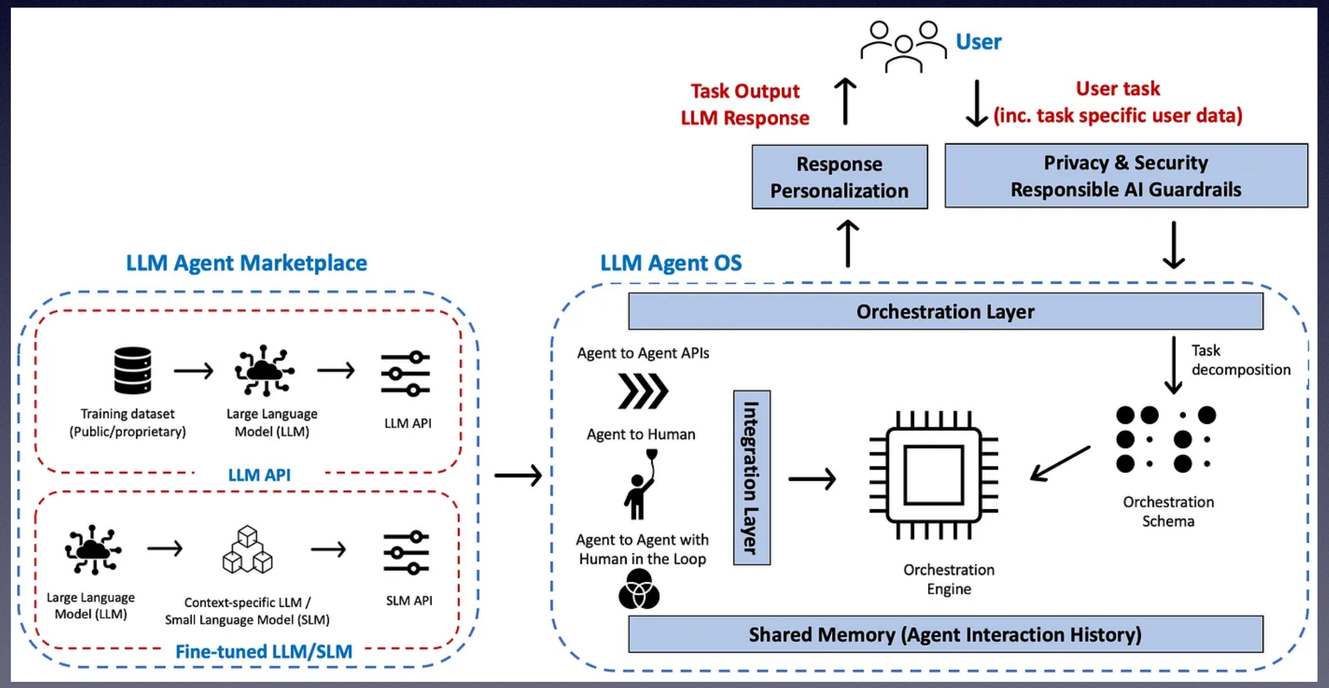
SOTA of MAS

- Vibe Coding: Claude Code, Cursor, Gemini CLI, Windsurf, Jules (Google) etc.
- Designing: Lovable, Sigma, Stitch, Figma etc.
- General domain: Genspark, MiniMax etc.
- Search: Perplexity, Google AI Mode, Fellou, Dia, Sigma, Genspark etc.
- Research: STORM (Stanford), SciMaster, Elicit etc.

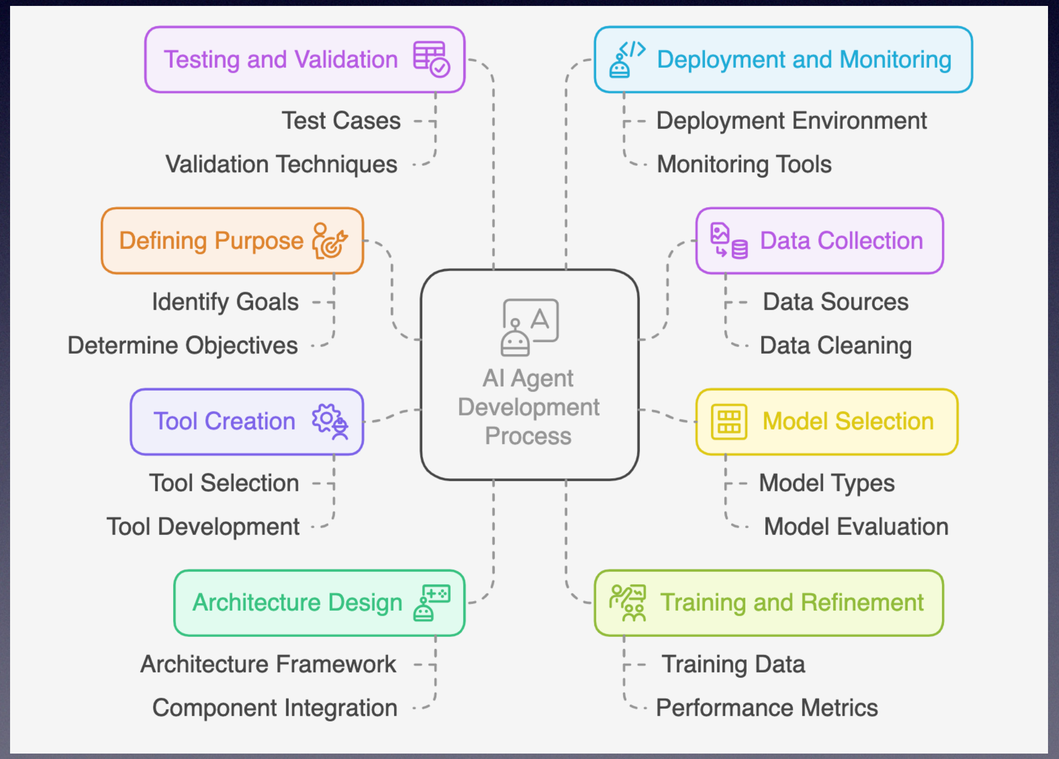
A comprehensive category of AI tools: https://www.toolify.ai/category

Engineering Essentials

Design a MAS…



but start from an agent

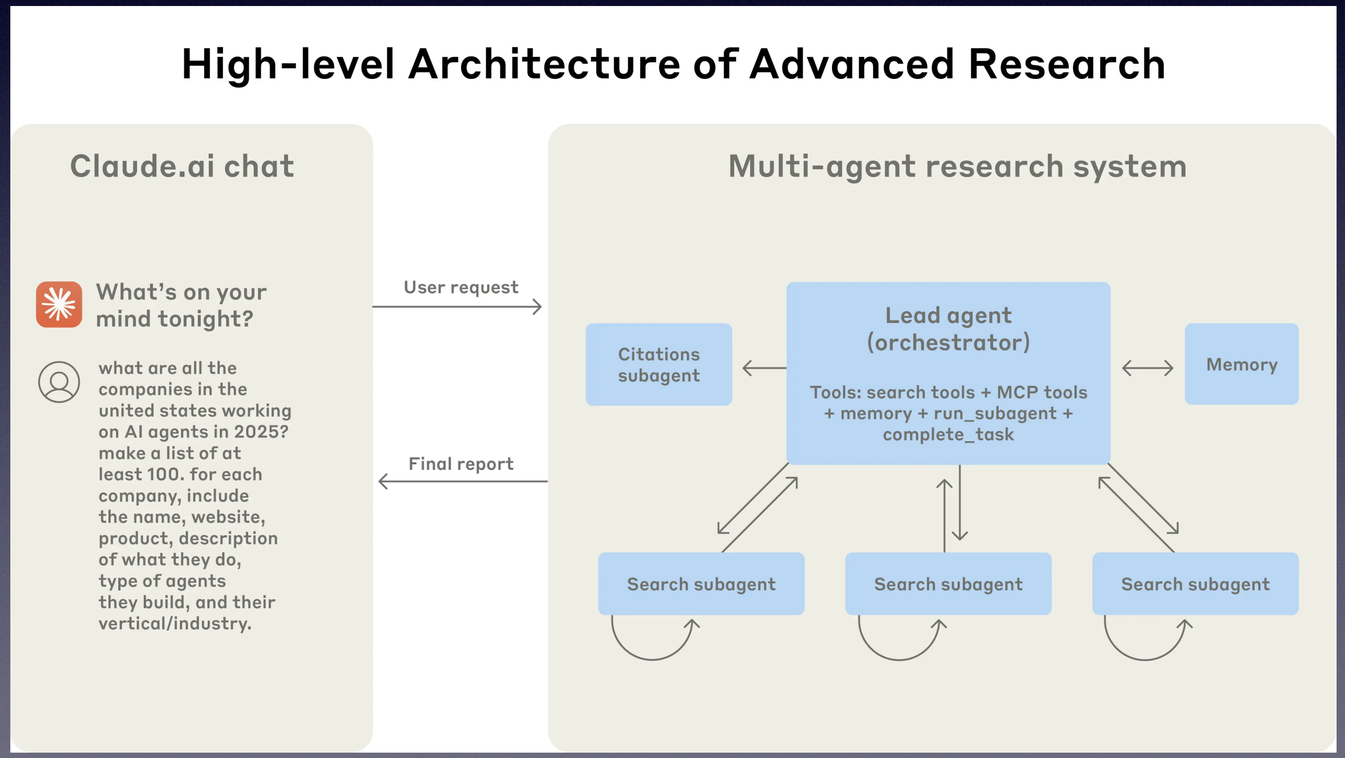


https://medium.datadriveninvestor.com/ai-agent-platform-reference-architecture-0be5b19d0eba

https://relevanceai.com/blog/how-to-build-an-ai-agent-a-comprehensive-guide-for-2025

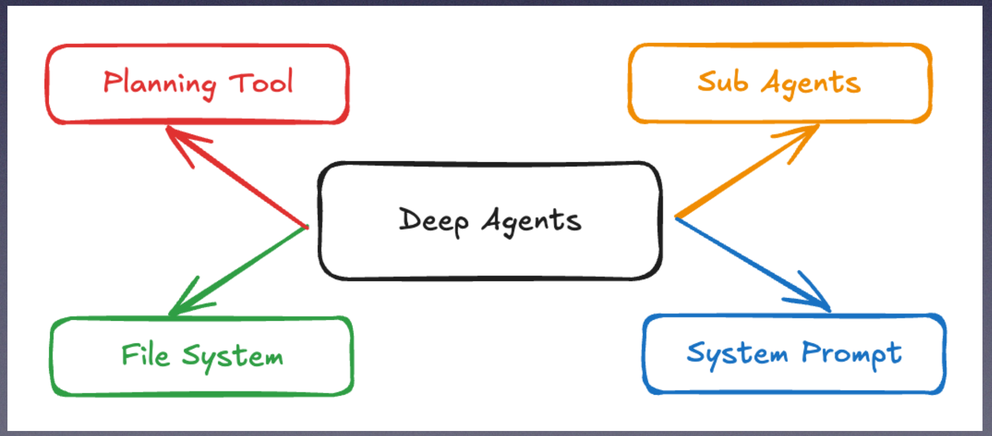
Learning from 1st tier

From Anthropic team



https://www.anthropic.com/engineering/built-multi-agent-research-system

- Main components
  - A planning tool + subagents + file system + delicate prompts



https://blog.langchain.com/deep-agents/

Demo: DeepResearch Agent

- https://github.com/caesar0301/mas-talk-2508/tree/master/code

Ideas Execution Matters

- Validate product-market fit as early as possible.
- Start with a powerful model (e.g. gemini), then to a multi-model, multi-agent system.
- Harness AI coding assistance.

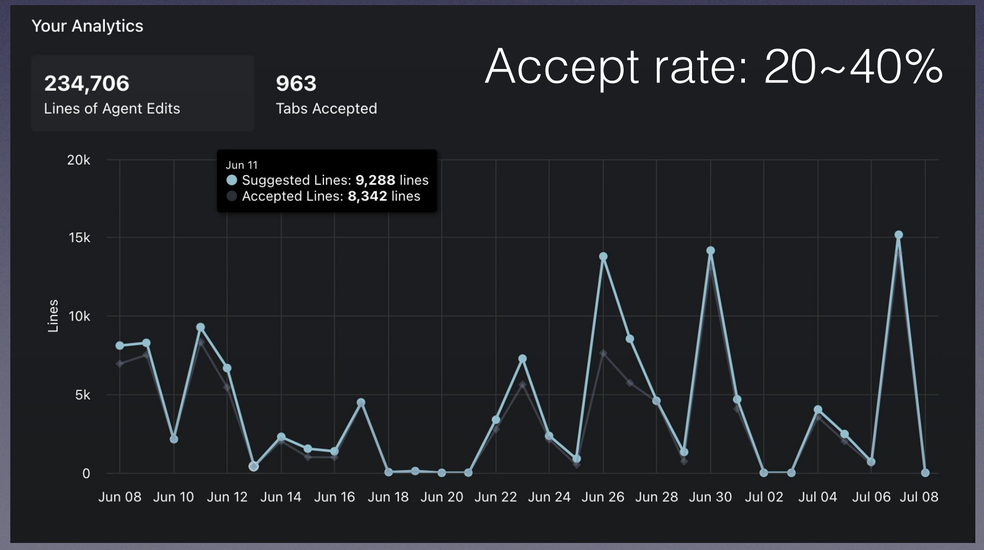
Talk is cheap, code is fast — only ideas that land create value.

Multi-Model is the Future

Let me start with a reality check that every AI practitioner knows: **no single model right now excels at everything.**

Through our partnerships with leading model teams, we have extensive experience with each major AI system's strengths and limitations. OpenAI excels at deep research and creative writing tasks. Anthropic's Claude demonstrates superior agentic reasoning, tool use and complex coding capabilities. Gemini consistently outperforms others in multimodal understanding—analyzing images, videos, or complex visual data. Grok Heavy delivers impressive capabilities for large-scale, complex reasoning tasks. And even the Kimi+Groq combination provides unbeatable speed and cost-effectiveness with solid quality. and so on...

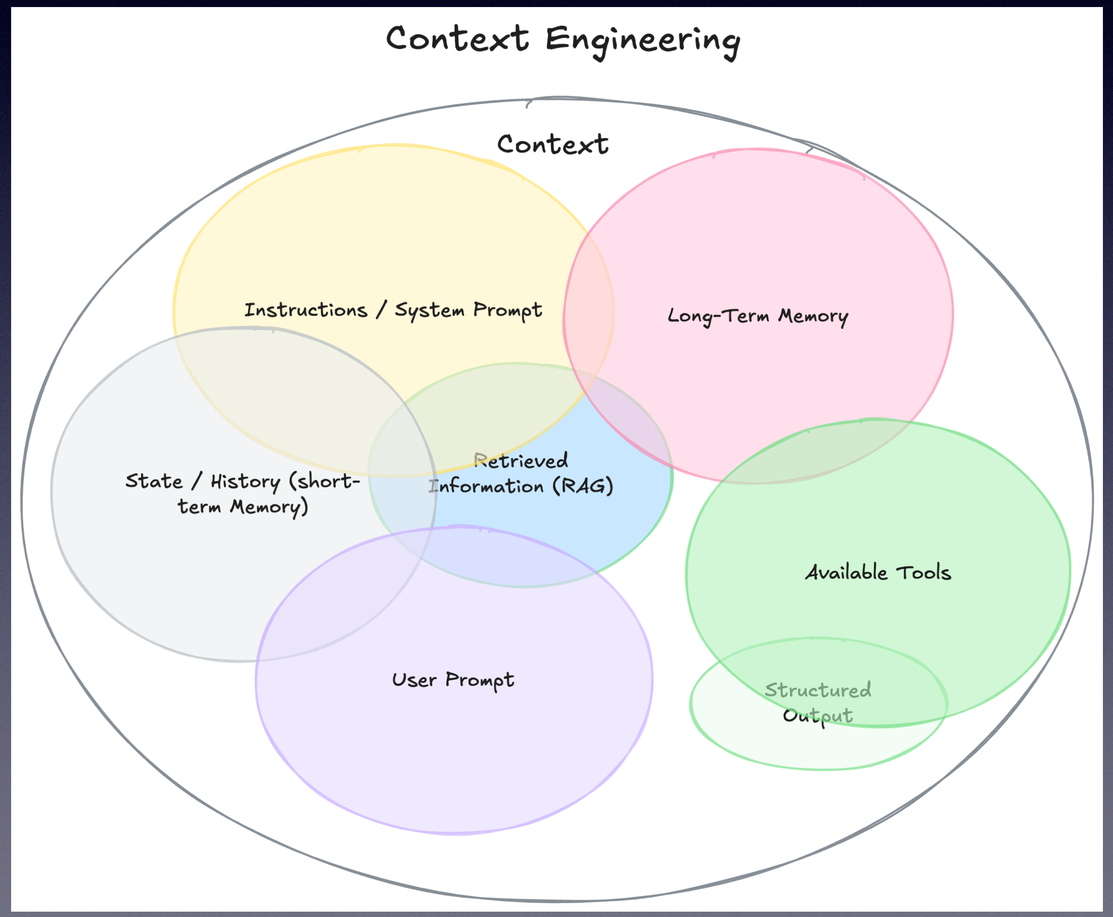
— From Eric Jing, Genspark CEO



https://x.com/drjaminchen/status/1942480133438218548

Context Engineering

— the final frontier of agentic AI



- Lacks robust tools for managing evolving context
- No standardized guidelines or best practices yet
- Critical for memory, orchestration, and long-horizon planning

Keep Agents on Track

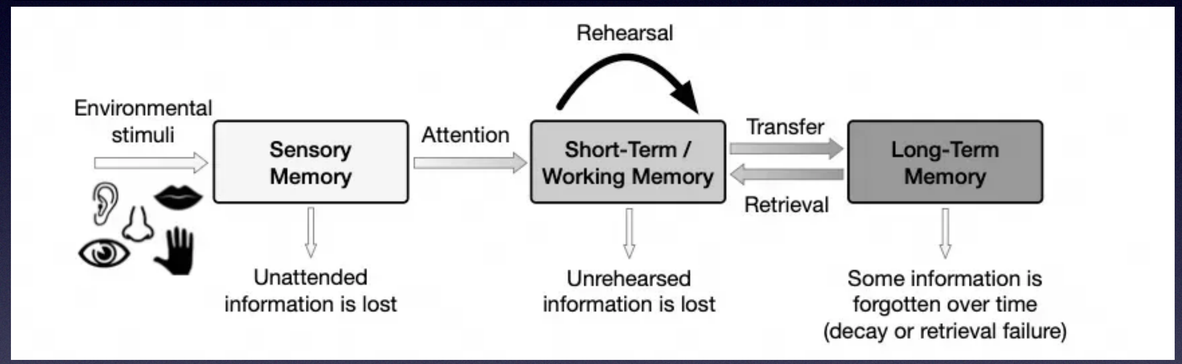
— from a practical perspective

- **Embrace modular design**: easy to test, easy for LLMs to reason.
- **Prompting still matters**: even in the era of “context engineering”.
- **Use a clear coordinator–worker arch:** structure enables scale.
- **Start wide, then narrow down**: validate each agent and tool rigorously.
- **Context engineering begins and ends with memory**

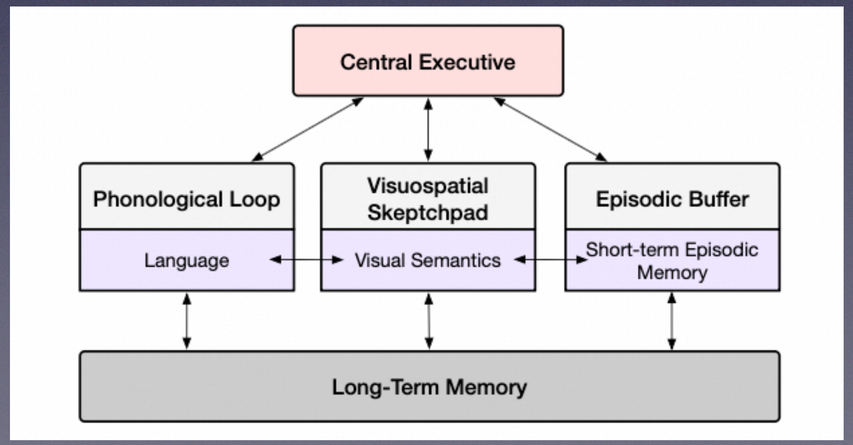
Agentic Memory

— rooted in human cognitive architecture (HCA)

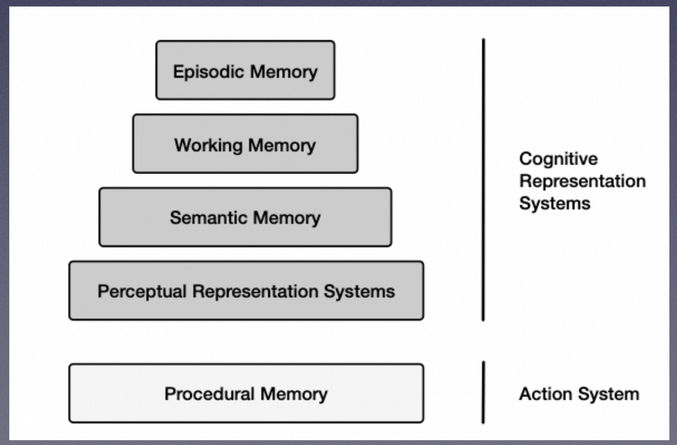
Multi-Store Model, Richard C Atkinson, 1968



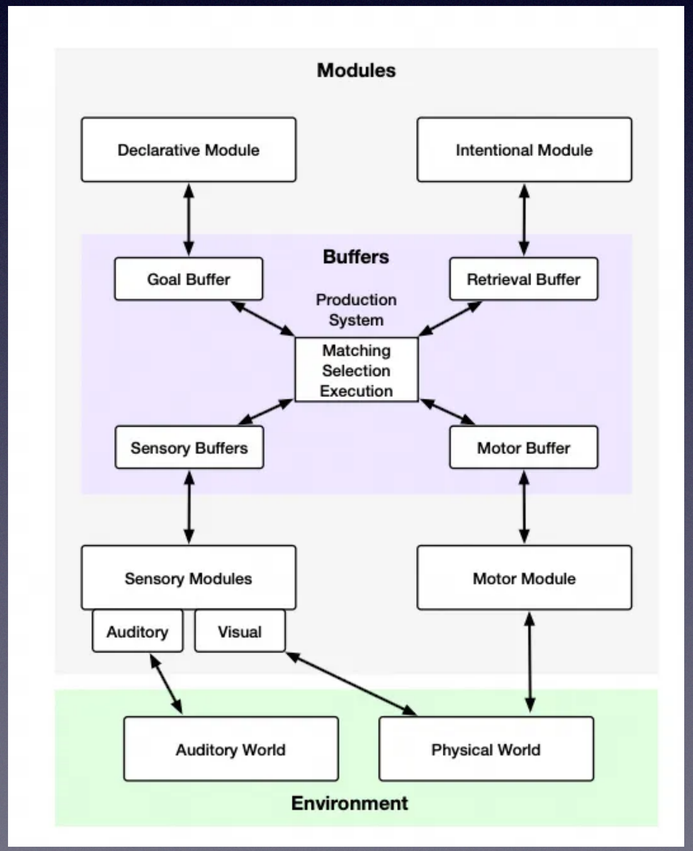
Working Memory Model, Alan Baddeley, 1974



SPI-Model, Endel Tulving, 1985



ACT-R, John R. Anderson, 2009

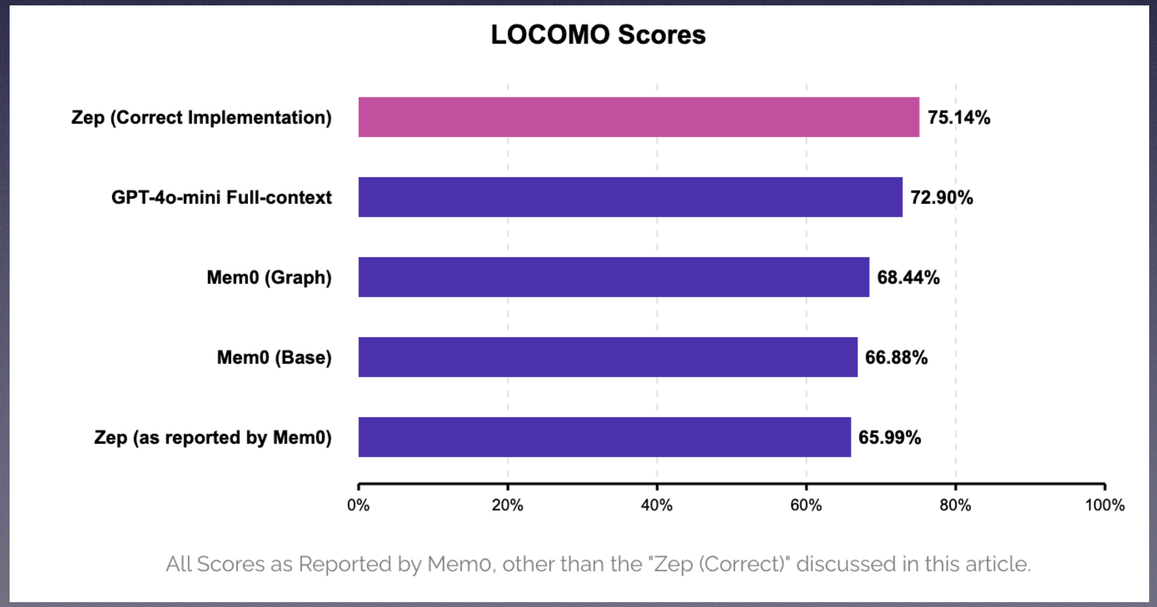


Agentic memory is booming

- We have a bunch of open source projects on agent memory in 2025
  - Zep (graphiti), Cognee, mem0, MIRIX, MemOS, Second Me, M+, MemoRAG, MEM1 etc.
- **Under the hood in common**: human memory concepts + RAG + KG

GAIA Leaderboard

Agent name	Model family	organisation	Averag
Su Zero Ultra		Suzhou AI Lab	80.4
h2oGPTe Agent v1.6.33	claude-3-7-sonnet-20250219, gemini-2.5-pro-preview-06-05 (extended thinkin	h2o.ai	79.73
Agent2030-v2.3	o3, GPT 4.1, Gemini 2.5 Pro		79.4
h2oGPTe Agent v1.6.32	claude-3-7-sonnet-20250219, gemini-2.5-pro-preview-06-05	h2o.ai	79.07
Agent_v0.1.0	gpt-4.1		79.07
AWorld (Run Instantly)	GPT-4o, DeepSeek V3, Claude-Sonnet-4, Gemini-2.5-Pro	inclusionAI	77.08
SU AI Zero	Anthropic, Google, openaI	Suzhou AI Lab	76.41
Agent2030-v2.2	o4-mini, GPT 4.1, Gemini 2.5 Pro		76.08



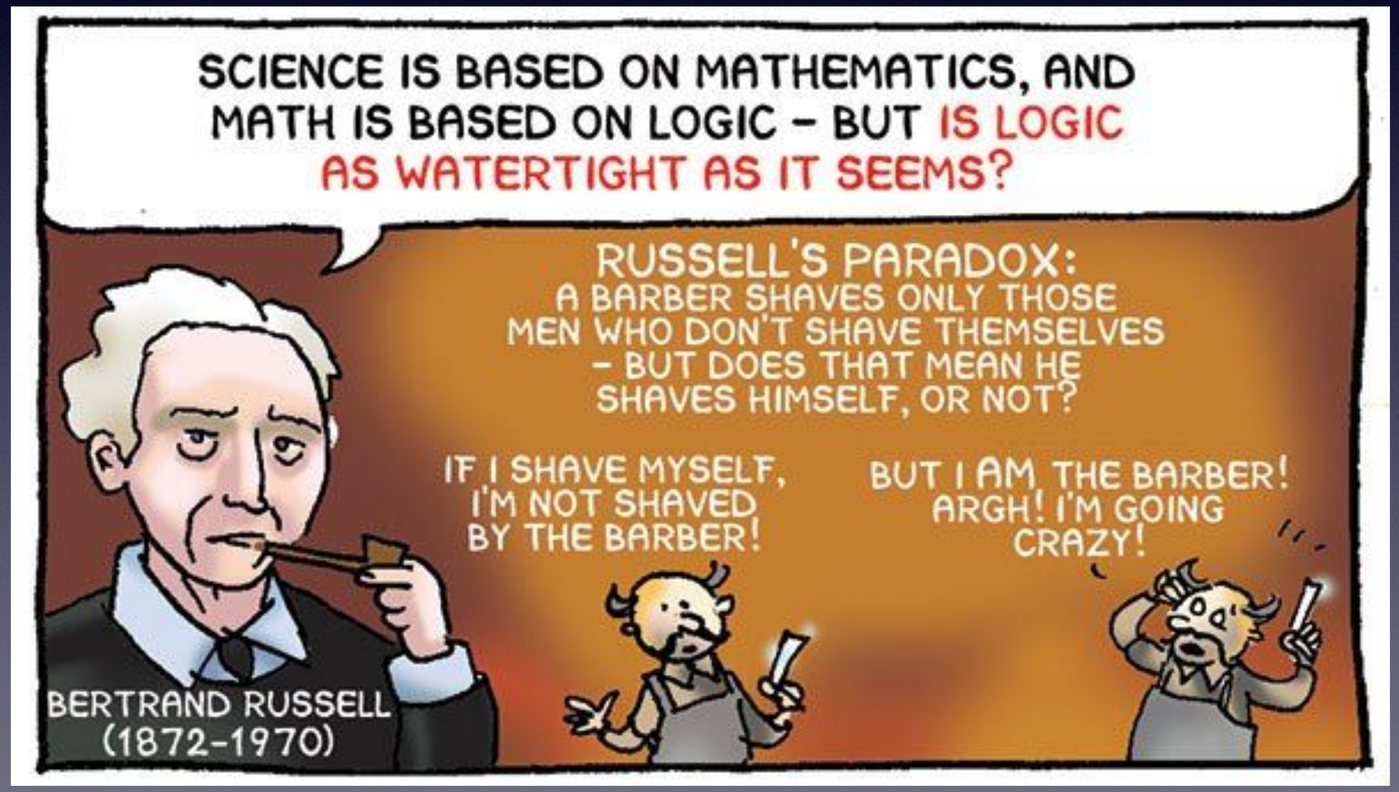
A Survey on the Memory Mechanism of LLM based Agents, ACM TIS, 2025

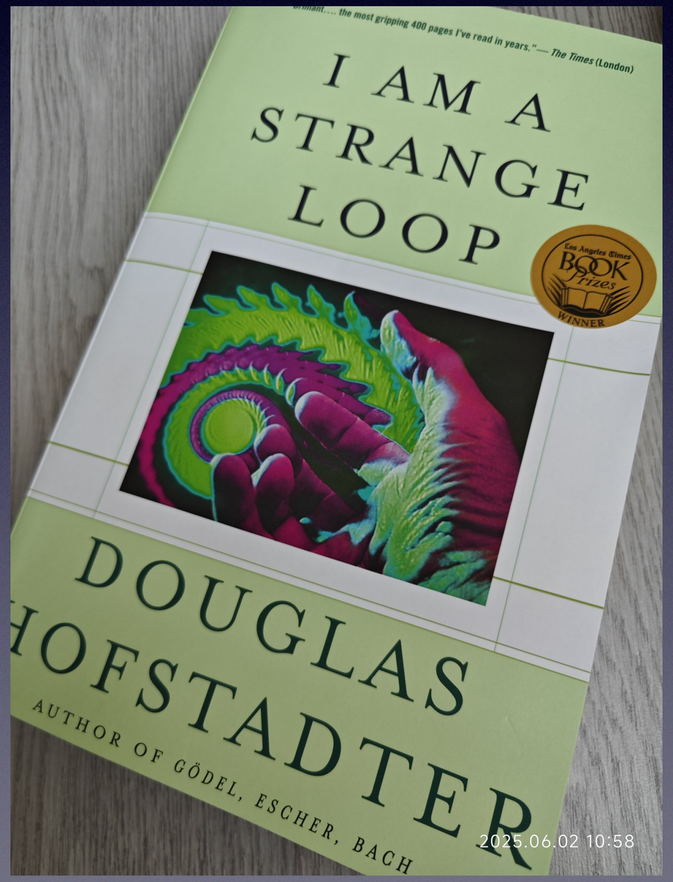
Advanced Topics

- Foundation innovation of AI
- From ReAct to Proactive agents (discuss)
- Intelligence vs. Memory (discuss)

Foundation Innovation of AI

One of theoretical foundations of agents: Gödel's incompleteness theorems





Foundation Innovation of AI

Example: LLM Dense Model vs. MoE

- A formal system adheres to Gödel's incompleteness theorems.
- A Turing machine is a formal system.
- Neural networks are not Turing-complete.
- However, transformers and recurrent neural networks (RNNs) have been proven to be Turing-complete under specific theoretical conditions.
- So, LLM cannot achieve consistency and completeness simultaneously.

https://www.xiaming.site/2025/02/08/neurosymbolic-part2

Must Read Papers on Agents

On Agent Memory

- Mem0: Building Production-Ready AI Agents with Scalable Long-Term Memory
- A Survey on the Memory Mechanism of LLM-Based Agents
- HippoRAG: Neurobiologically Inspired Long-Term Memory for Large Language Models
- MemoRAG: Moving Towards Next-Gen RAG via Memory-Inspired Knowledge Discovery
- Memory^3: Language Modeling with Explicit Memory

Must Read Papers on Agents

On Graph Retrieval

- From Local to Global: A Graph RAG Approach to Query-Focused Summarization
- HybridRAG: Integrating Knowledge Graphs and Vector Retrieval-Augmented Generation for Efficient Information Extraction
- G-Retriever: Retrieval-Augmented Generation for Textual Graph Understanding and Question Answering
- LightRAG: Simple and Fast Retrieval-Augmented Generation
- GRAG: Graph Retrieval-Augmented Generation

Must Read Papers on Agents

On Cognitive Architectures

- Cognitive Architectures for Language Agents
- Human Problem Solving – Newell & Simon (1972)
- Consciousness Is Computational: The LIDA Model of Global Workspace Theory
- Survey on Memory-Augmented Neural Networks: Cognitive Insights to AI Applications
- A Theory of Consciousness from a Theoretical Computer Science Perspective: Insights from the Conscious Turing Machine
- Consciousness in Artificial Intelligence: Insights from the Science of Consciousness