Costruire sistemi multi-agente efficaci con Semantic Kernel: tecnologie e sfide

Anna Fabris

Machine Learning Engineer
Impresoft 4ward

Matteo Pagani

Cloud Solution Architect
Modern Work Global CSU
Microsoft

global AI community

WHAT IS A MULTIAGENT SYSTEM?

A multiagent system (MAS) is a group of autonomous AI agents working together to solve complex tasks.

AGENTS

Individual parts of the system.

Each agent has its own abilities, knowledge, and goals. Agents can be language models, tools...

INTERACTIONS

Agents interact with each other with various methods, such as talking to each other, working together, or competing.

CHALLENGES OF MULTIAGENT SYSTEMS

CONTROLLABILITY

Difficulty having control on the behavior of the system

COORDINATION COMPLEXITY

Preventing conflicts and ensuring coherent collective behavior

UNPREDICTABLE BEHAVIOR

Detecting issues is complicated due to the independency of agents

ERROR PROPAGATION

Error from a single LLM agent can propagate through the network

COST

Calling too many LLM inferences can become very expensive

TASK TERMINATION

Avoiding infinite loops in long task can be challenging.

WHY NOT JUST ONE LLM?

FLEXIBILITY

Multiagent systems can adjust to varying environments by adding, removing or adapting agents.

SCALABILITY

A single LLM can only "remember" a limited amount of input (limited context window)

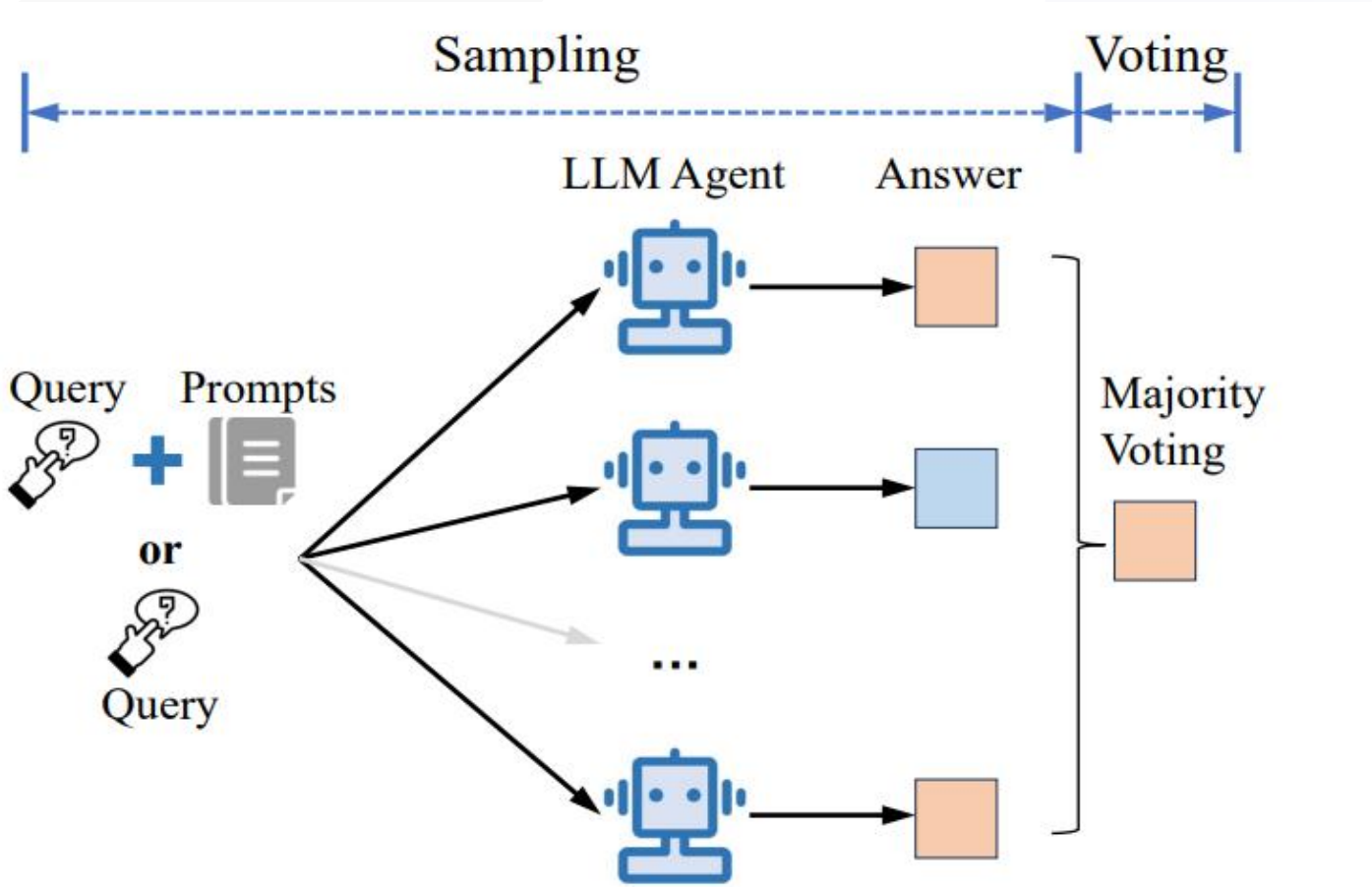
SPECIALIZATION

Different agents can be trained or prompted to excel at specific tasks

PERFORMANCE

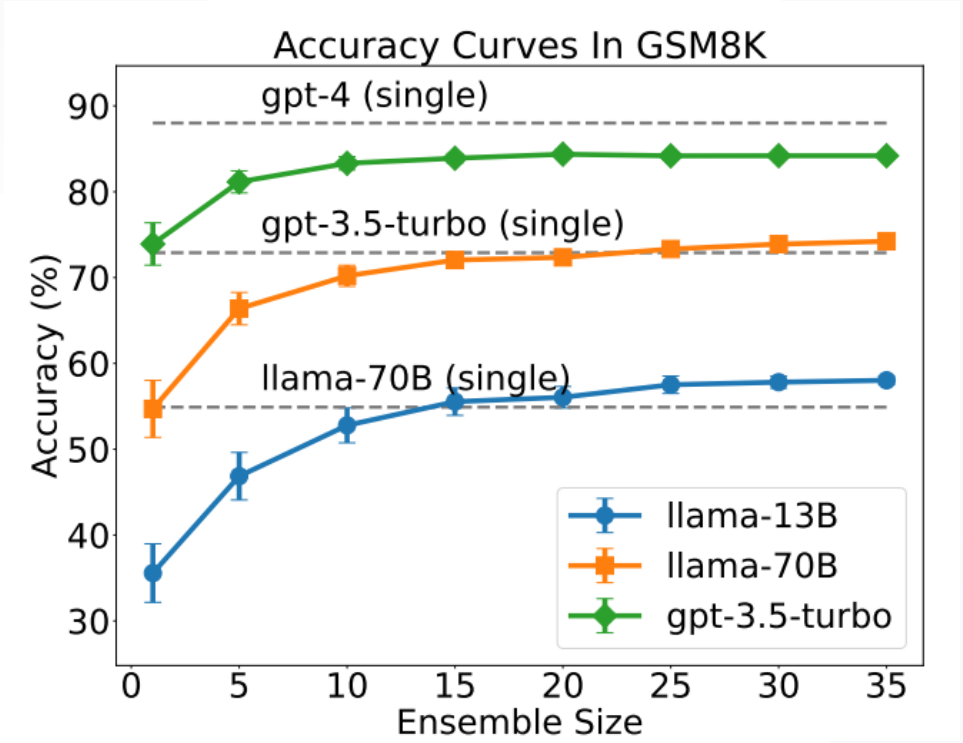
Multiagent frameworks tend to outperform singular agents

ARCHITECTURE: AGENT FOREST



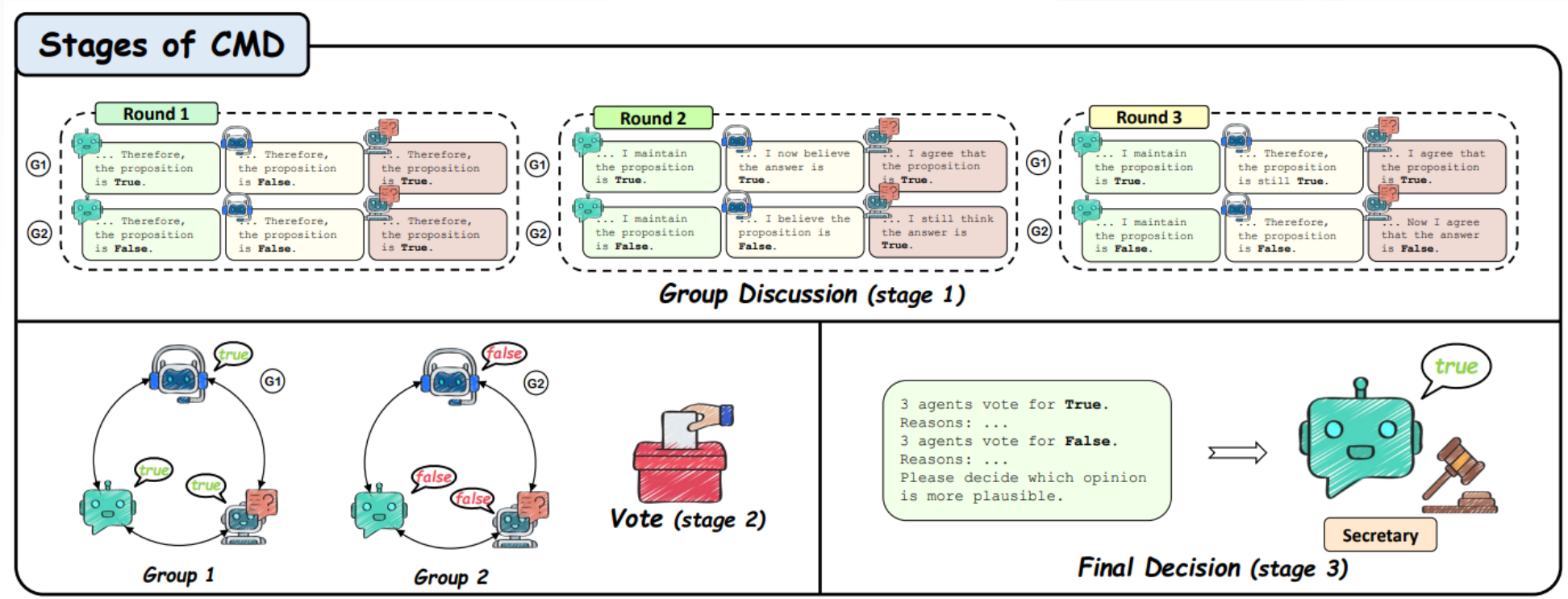
Li, Junyou, et al. "More agents is all you need." *arXiv preprint arXiv:2402.05120* (2024)

ARCHITECTURE: AGENT FOREST



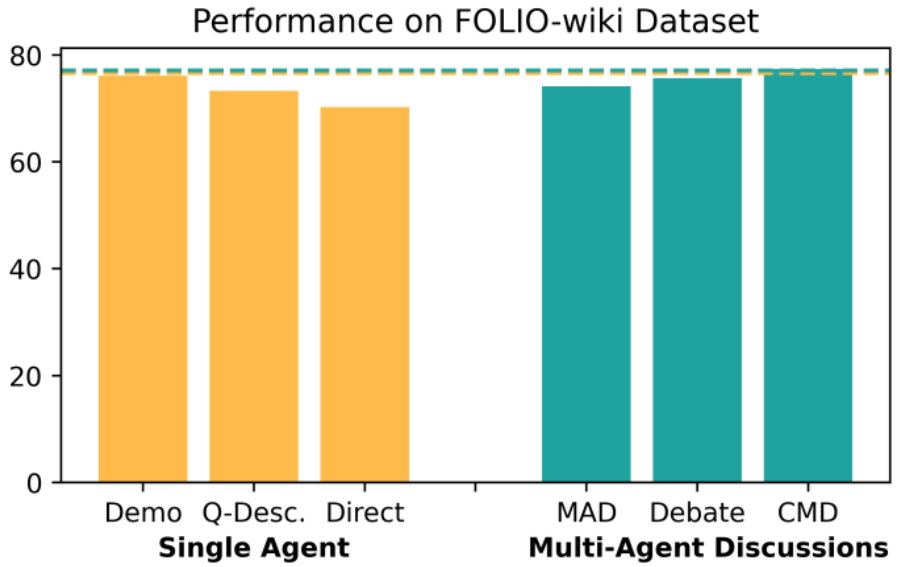
Li, Junyou, et al. "More agents is all you need." *arXiv preprint arXiv:2402.05120* (2024)

ARCHITECTURE: CONQUER-AND-MERGE DISCUSSION



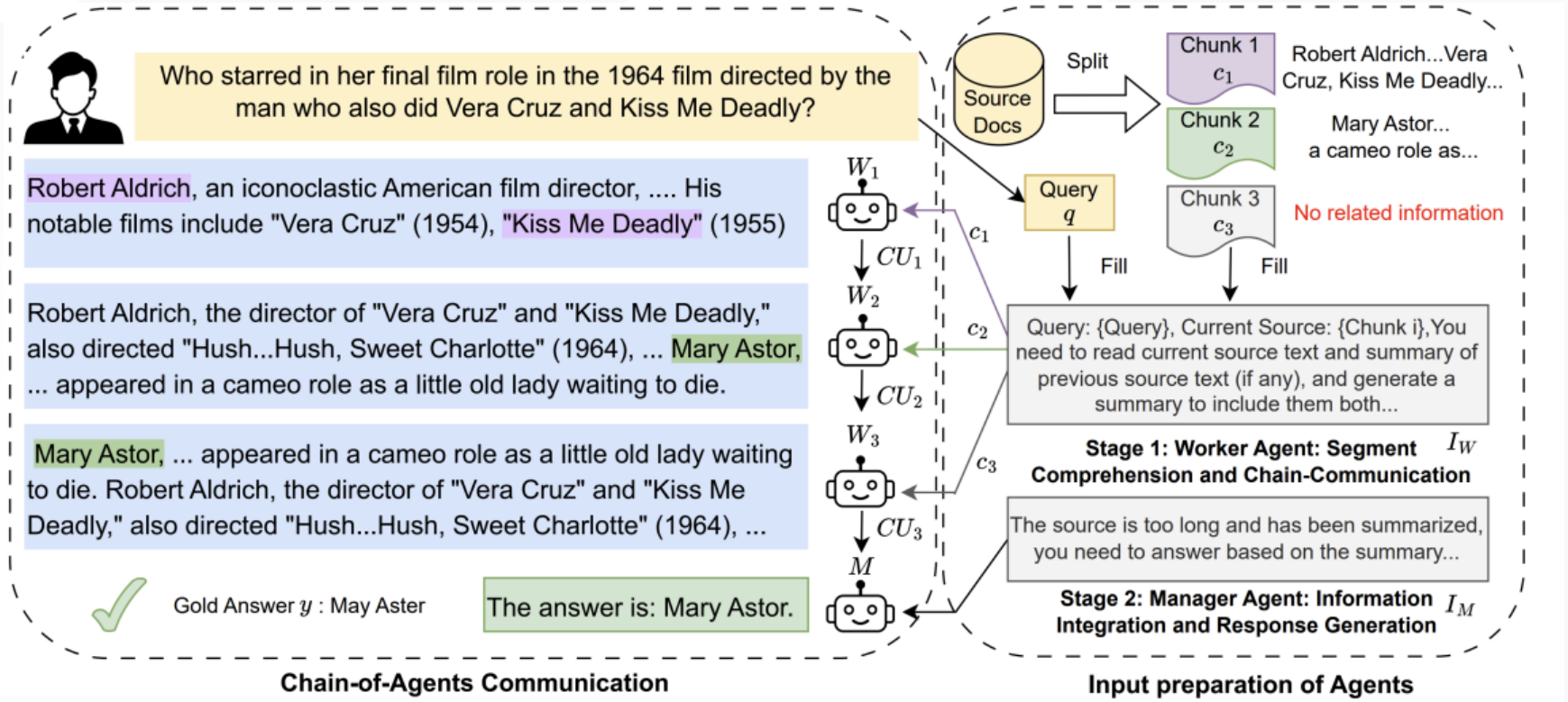
Wang, Qineng, et al. "Rethinking the bounds of LLM reasoning: Are multi-agent discussions the key?." *arXiv preprint arXiv:2402.18272* (2024).

ARCHITECTURE: CONQUER-AND-MERGE DISCUSSION



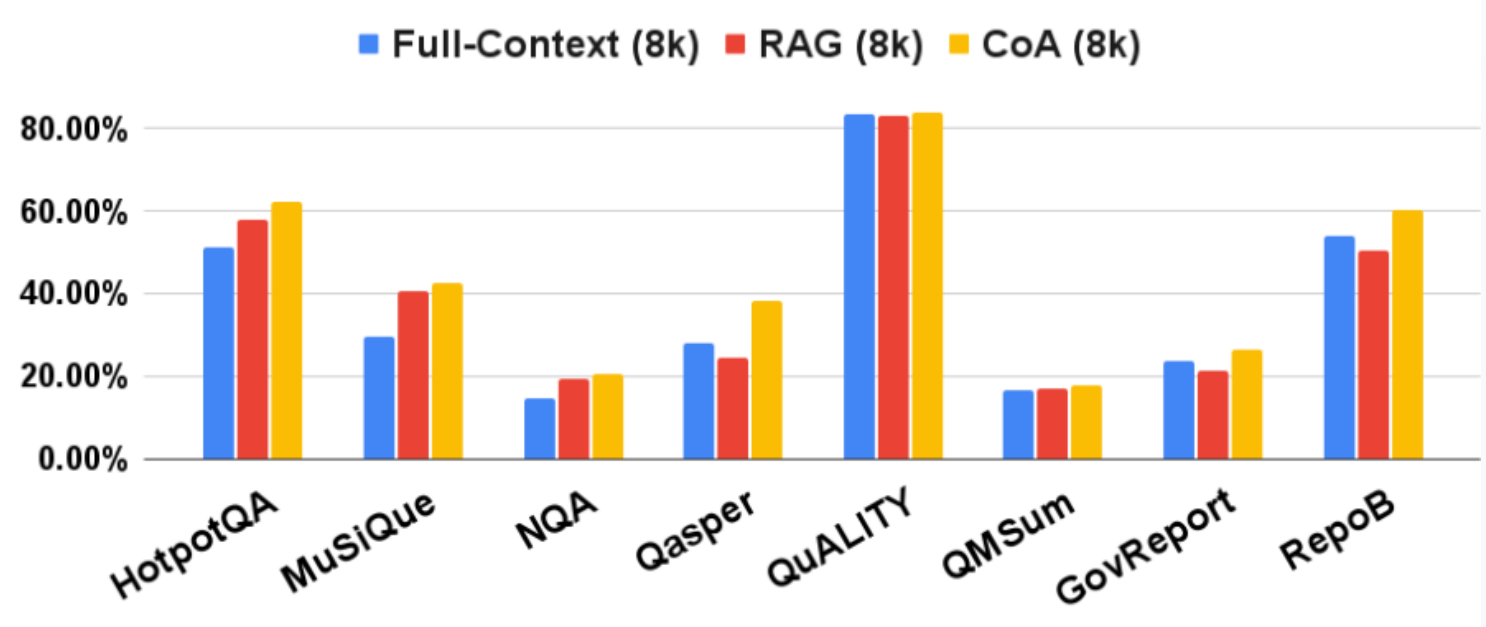
Wang, Qineng, et al. "Rethinking the bounds of LLM reasoning: Are multi-agent discussions the key?." *arXiv preprint arXiv:2402.18272* (2024).

ARCHITECTURE: CHAIN-OF-AGENTS



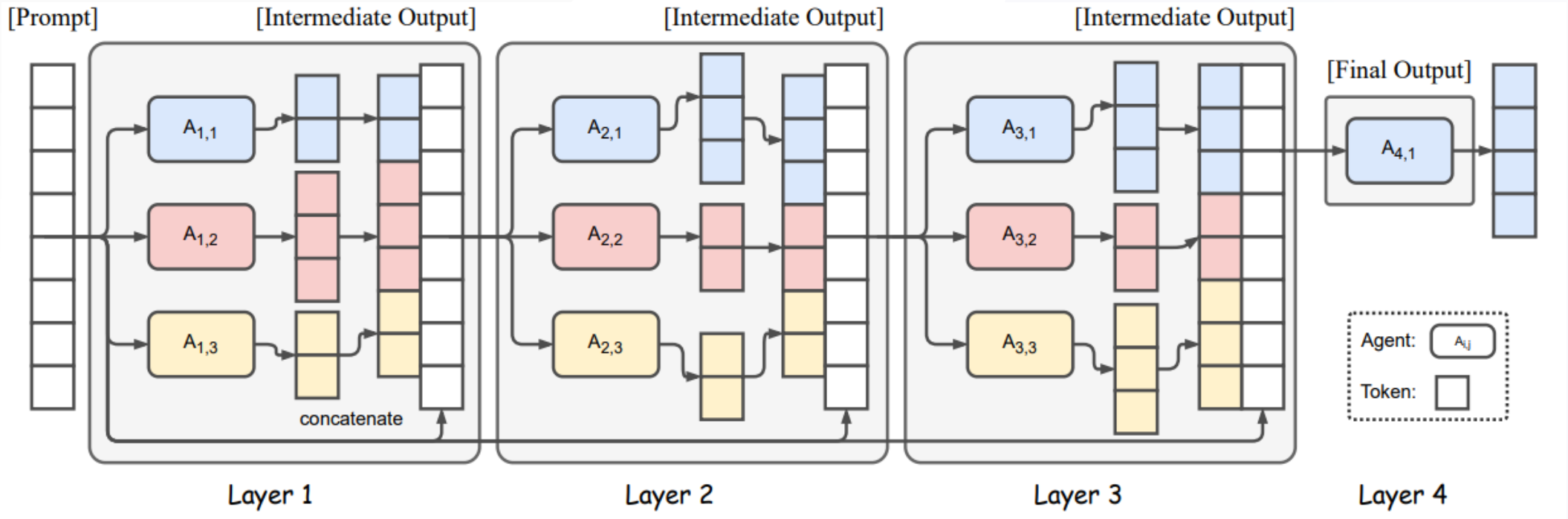
Zhang, Yusen, et al. "Chain of agents: Large language models collaborating on long-context tasks." *Advances in Neural Information Processing Systems* 37 (2024)
https://research.google/blog/chain-of-agents-large-language-models-collaborating-on-long-context-tasks/

ARCHITECTURE: CHAIN-OF-AGENTS



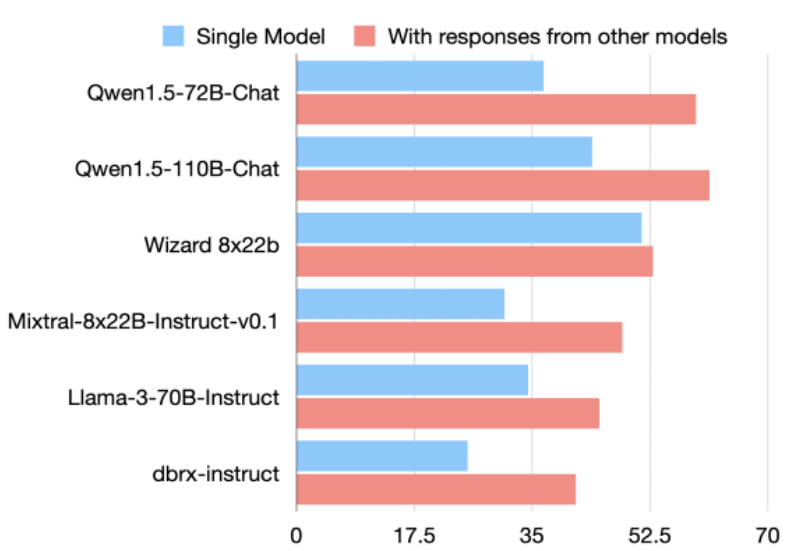
Zhang, Yusen, et al. "Chain of agents: Large language models collaborating on long-context tasks." *Advances in Neural Information Processing Systems* 37 (2024)
https://research.google/blog/chain-of-agents-large-language-models-collaborating-on-long-context-tasks/

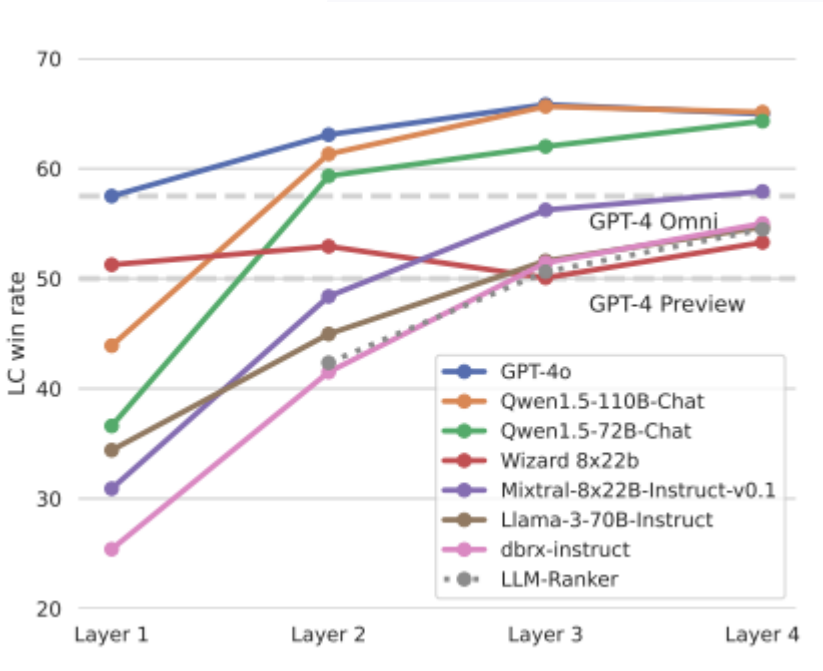
ARCHITECTURE: MIXTURE-OF-AGENTS



Wang, Junlin, et al. "Mixture-of-agents enhances large language model capabilities." *arXiv preprint arXiv:2406.04692* (2024)

ARCHITECTURE: MIXTURE-OF-AGENTS





Wang, Junlin, et al. "Mixture-of-agents enhances large language model capabilities." *arXiv preprint arXiv:2406.04692* (2024)

MEMORY

SHARED MEMORY
Centralized knowledge base accessible by all agents

DISTRIBUTED MEMORY
Agent-specific memory stores for specialized knowledge and local context

CONTEXUAL MEMORY
Stored temporarily in prompt or system message

PERSISTENT MEMORY
External storage (Vector Databases, Files, SQL)

FRAMEWORKS

LangGraph

Graph-based workflows with stateful agent interactions

CrewAI

Role-playing agents collaborating as teams

AutoGen

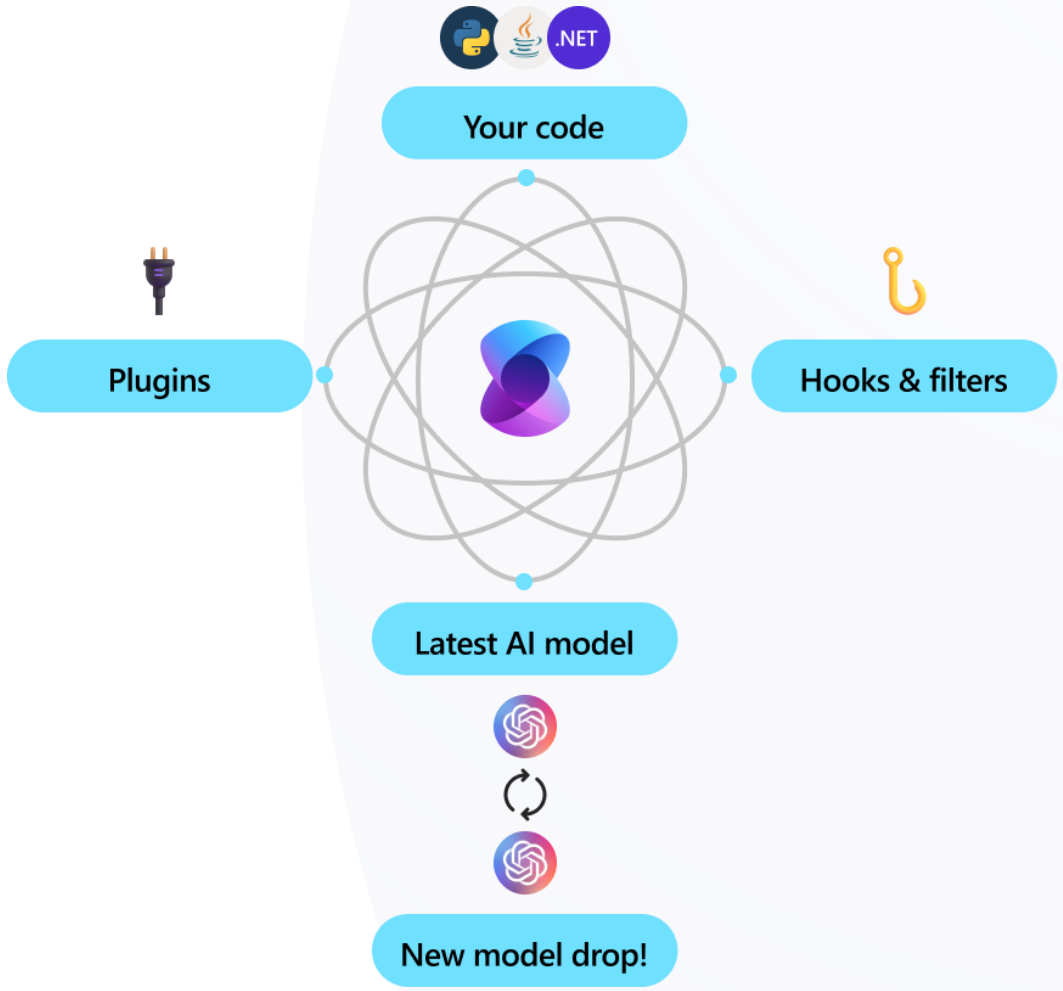
Multi-agent conversations with code execution

Semantic Kernel

Integrating AI into existing enterprise applications

SEMANTIC KERNEL

- Lightweight, open-source development kit that lets you easily build AI agents and integrate the latest AI models into your C#, Python, or Java codebase
- Built-in support for multiple AI services, cloud-based or local
- Built-in support for plugins and memory, with integration with the most popular vector database
- Multi-agent capabilities, with support for local agents, Prompty-based agents and Azure AI Foundry agents



EASY INTEGRATION WITH AI SERVICES

Amazon Bedrock

Gemini

PLUGINS

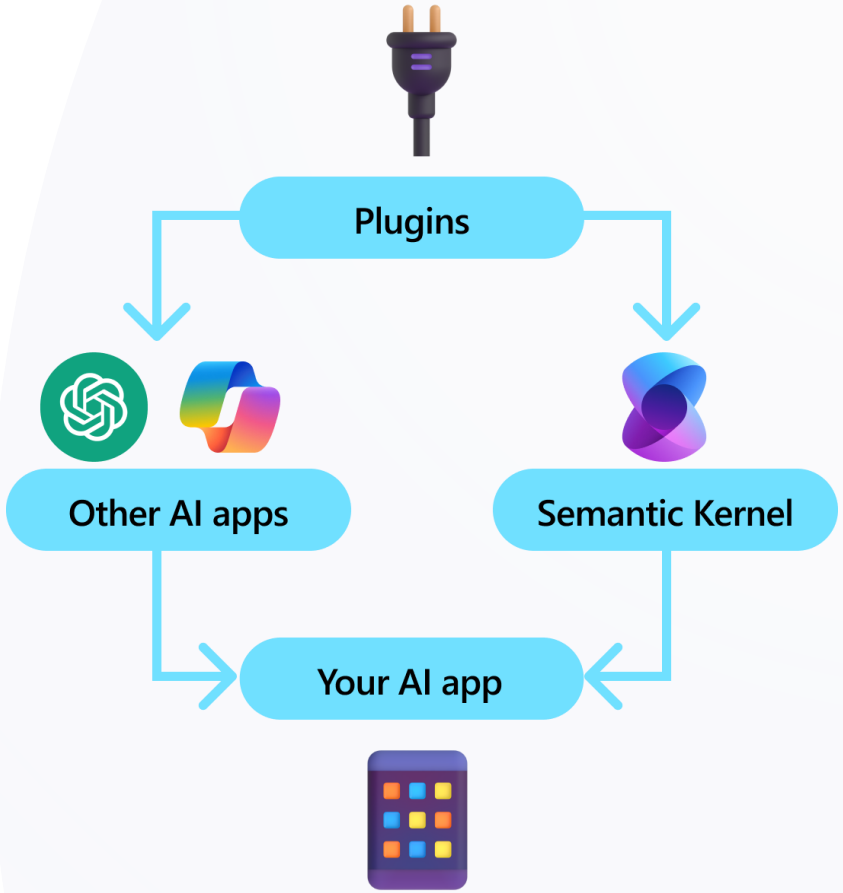
Writer plugin

Function	Description for model
`Brainstorm`	Given a goal or topic description generate a list of ideas.
`EmailGen`	Write an email from the given bullet points.
`ShortPoem`	Turn a scenario into a short and entertaining poem.
`StoryGen`	Generate a list of synopsis for a novel or novella with sub-chapters.
`Translate`	Translate the input into a language of your choice.

Can you write me a short poem about living in Dublin, Ireland and then create a story based on the poem?

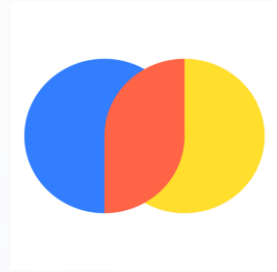
Planner

Copilot
Sure! Here's a story based on living along the Grand Canal in Dublin, Ireland...

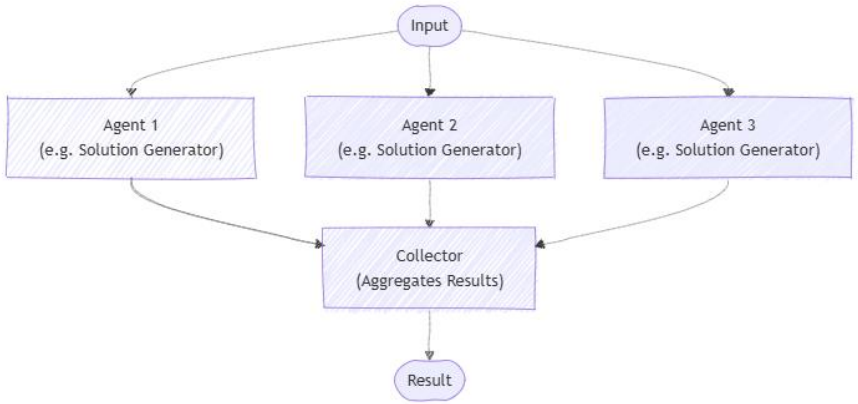


Model-Context-Protocol built-in support

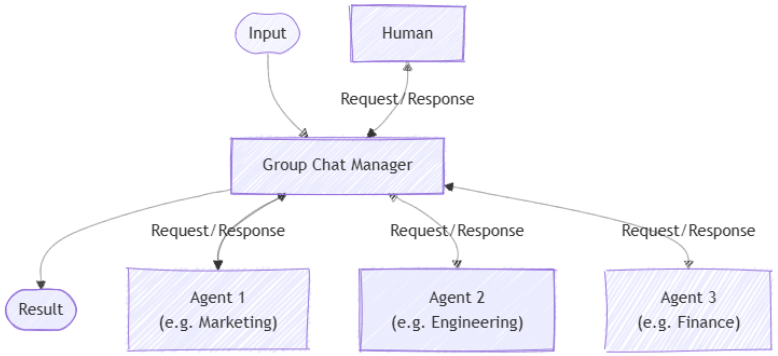
MEMORY



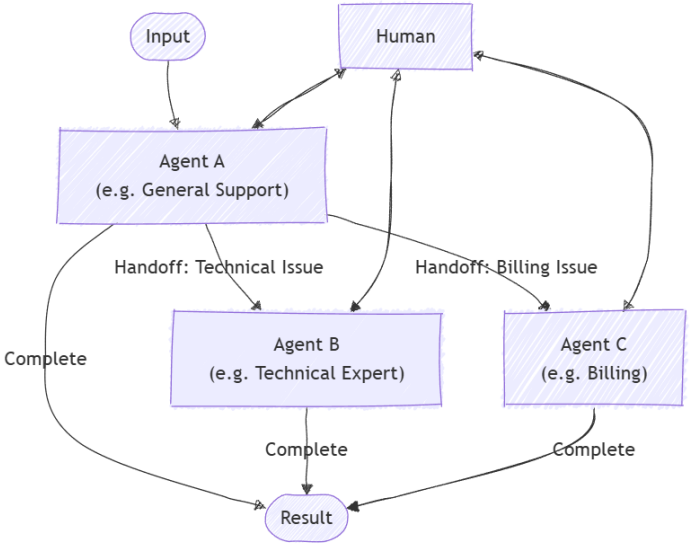
ORCHESTRATION



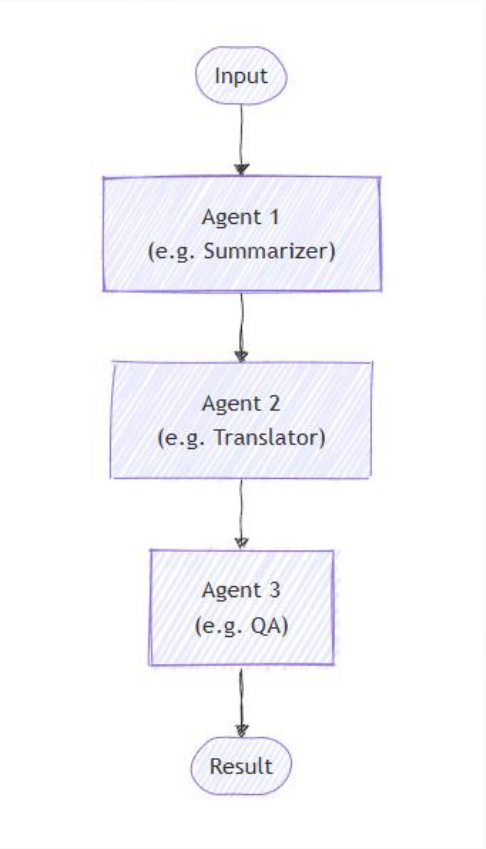
Concurrent



Group Chat



Handoff



Sequential

DOMANDE?

Demo e slide su

https://github.com/qmatteoq/AgentCon2025