



Foundation models

Andrey Kuznetsov

PhD, Head of FusionBrain Lab, AIRI

Outline

- Foundation model and interpretability
- LLM agent
- Baby AGI experiment

01

Foundation model and
interpretability



“Foundation models are powerful resource-intensive models that have emerged as the dominant paradigm for AI in the 2020s.”

Stanford Institute for Human-Centered Artificial Intelligence



Artificial general intelligence (AGI) is a complex and multifaceted learning paradigm encompassing several key components that are considered essential for achieving human-level general intelligence in artificial systems

AGI key features

- Matches or exceeding human performance across a broad class of cognitive tasks (e.g., perception, reading comprehension, and reasoning) in a variety of contexts and environments
- Possesses the ability to handle problems quite different from those anticipated by its creators
- Is able to generalize/transfer the learned knowledge from one context to others

Mapping the Mind of a Large Language Model

- matching *patterns* of neuron activations (features) to human-interpretable concepts ("dictionary learning")
- only a few active features
- begin to understand model behavior by checking which features respond to a particular input, giving an insight to the model's reasoning for how it arrived at a given output

Descriptions of, or references to, the **Golden Gate Bridge**

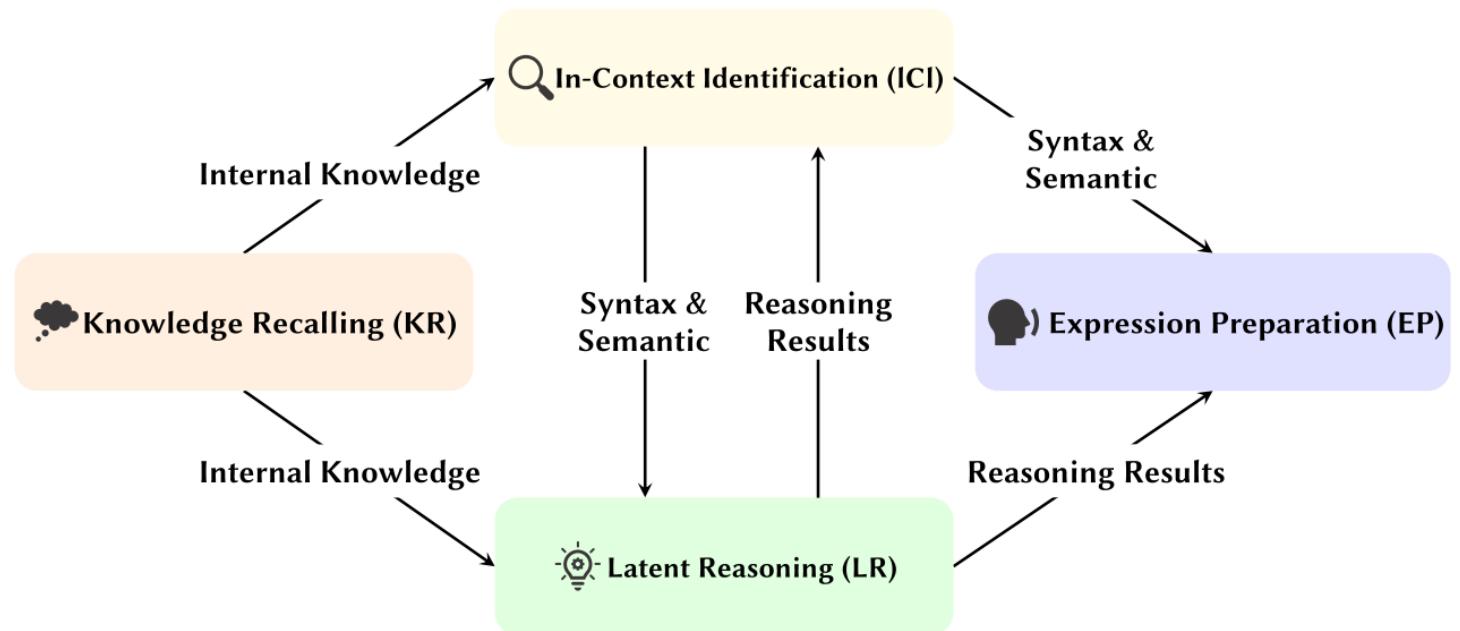


WONDER OF THE WORLD," "THE GOLDEN GATE BRIDGE
come hurtling in through the Golden Gate Bridge
Golden Gate Bridge wind resistance barriers
I know the ↪golden gate bridge often does ,

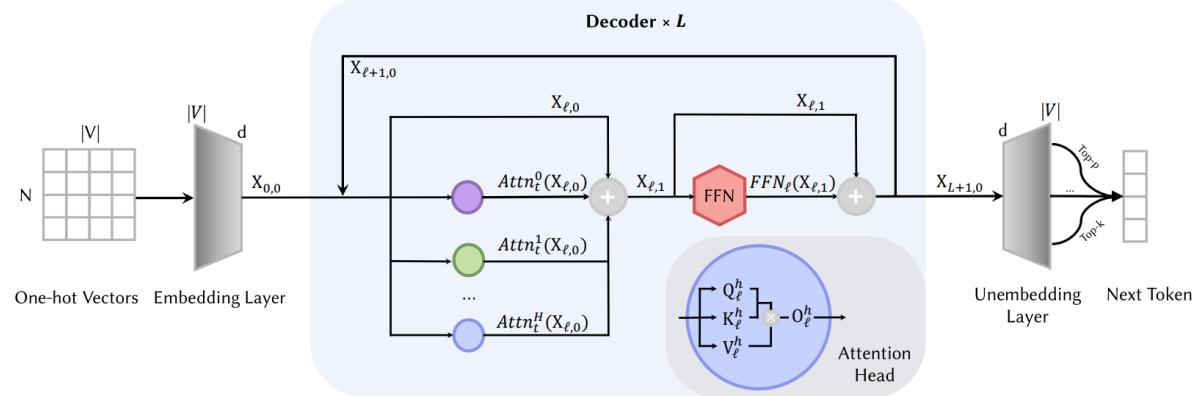
Attention heads as a way to interpretability

The human thought process into a four-stage framework

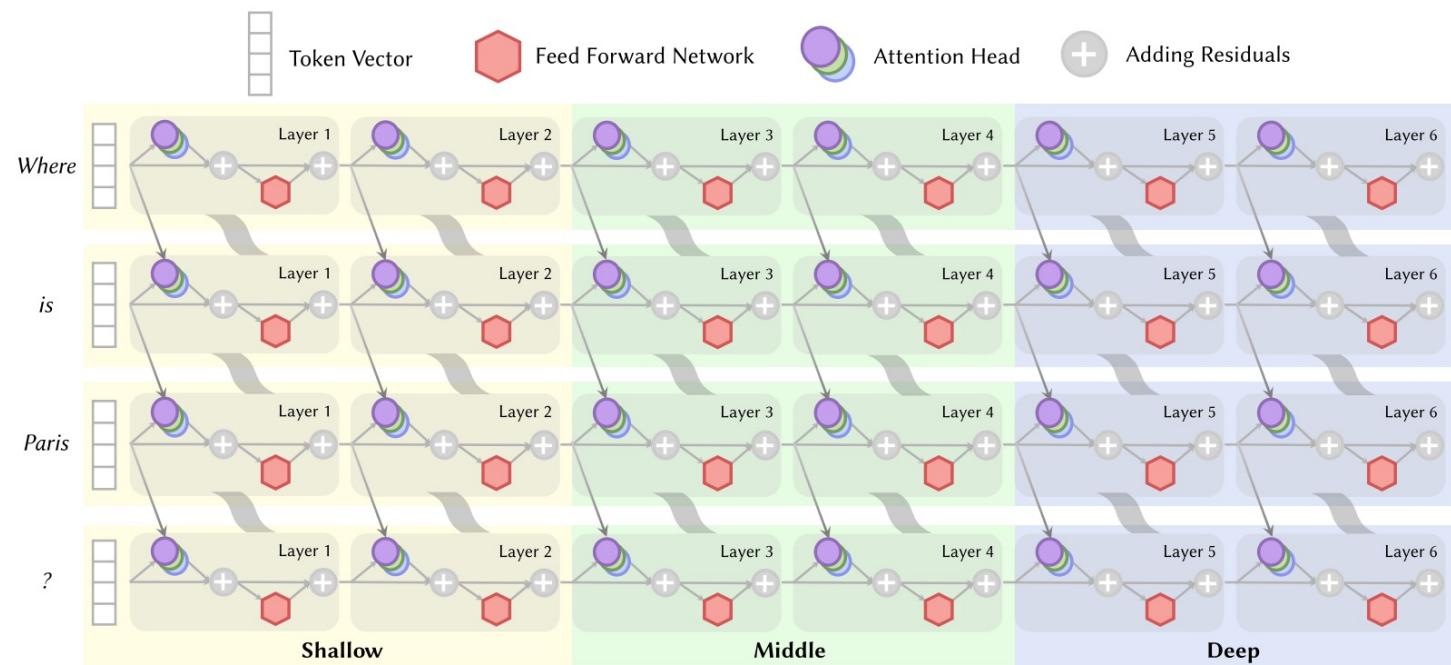
- Knowledge Recalling
- In-Context Identification
- Latent Reasoning
- Expression Preparation



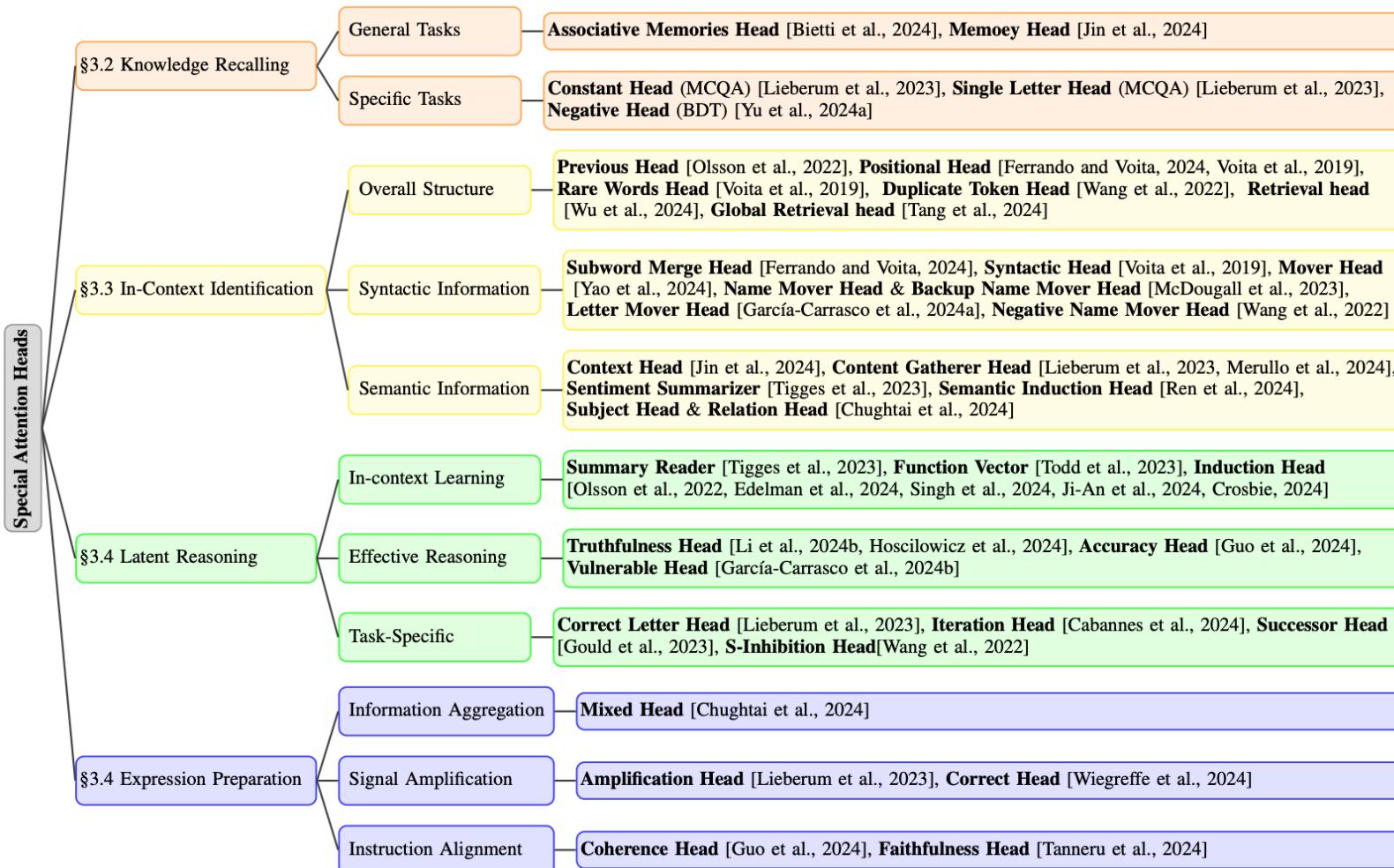
Attention heads as a way to interpretability



- **micro-level:** attention heads access latent state matrices from other residual streams
- **macro-level:** different residual streams (shared bandwidth) control the flow of information through attention heads



Attention heads as a way to interpretability



Attention heads as a way to interpretability

Knowledge Recalling

- General tasks (associative memories)
- Specific task scenarios (*single letter head*)

In-Context Identification

- QK matrices focus on and identify overall info for
- structure (Needle-in-a-Haystack)
 - syntaxics
 - semantics

Latent Reasoning

- Task recognition
- Task learning (*induction head*)
- Effective reasoning
- Task specific reasoning

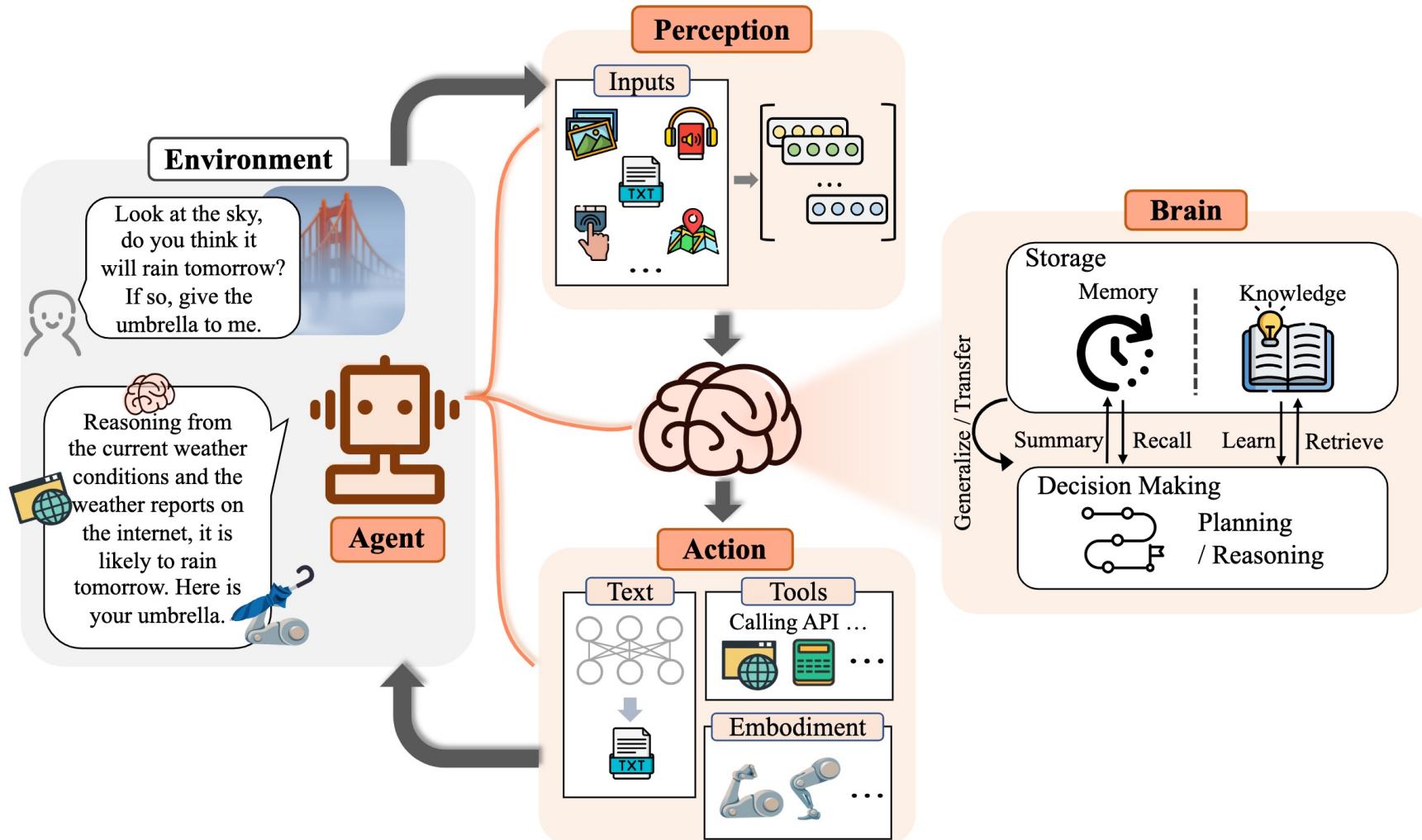
Expression Preparation

- Aggregate information
- Signal amplification
- Align the model's reasoning results with user's instructions

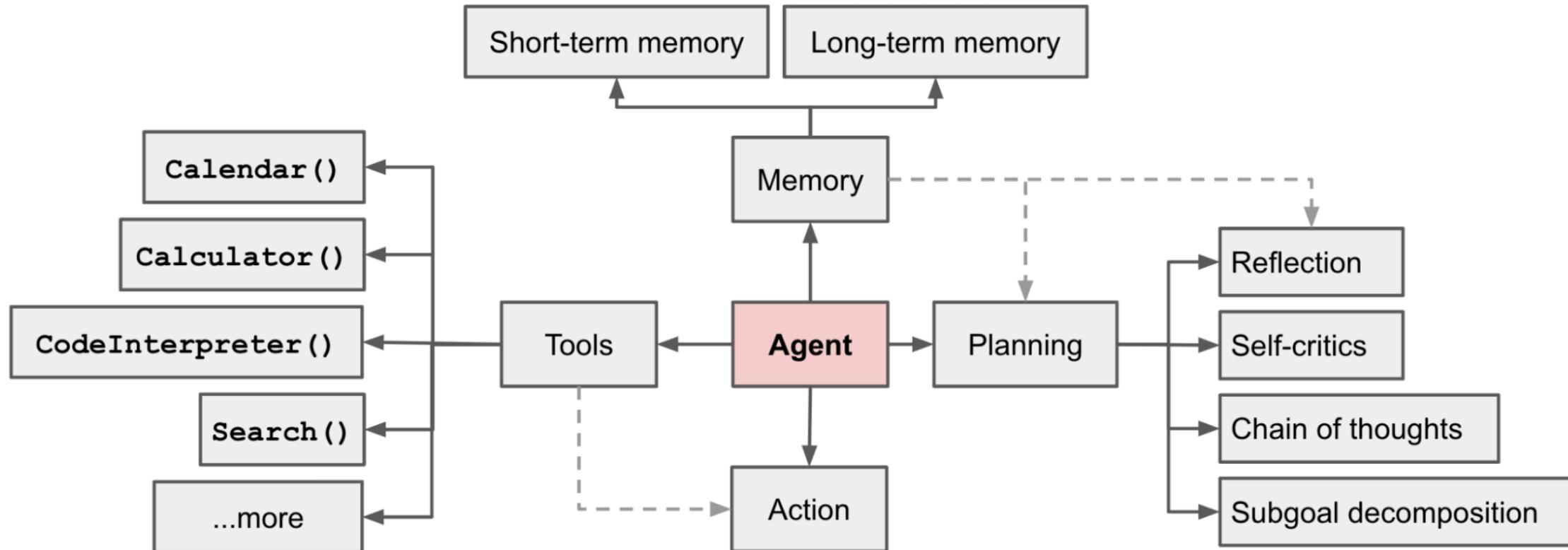
02

LLM Agent

RL agent -> LLM agent



LLM agent functions



LLM agent functions

- Planning
 - Subgoal and decomposition: large tasks into smaller, manageable subgoals, efficient handling of complex tasks
 - Reflection and refinement: do self-criticism and self-reflection over past actions, learn from mistakes and refine them for future steps, improving the quality of final results
- Memory
 - Short-term memory: in-context learning as utilizing short-term memory of the model to learn
 - Long-term memory: retain and recall (infinite) information over extended periods, often by leveraging an external vector store and fast retrieval
- Tool use
 - Call external APIs for extra information, including current information, code execution capability, access to proprietary information sources, etc.

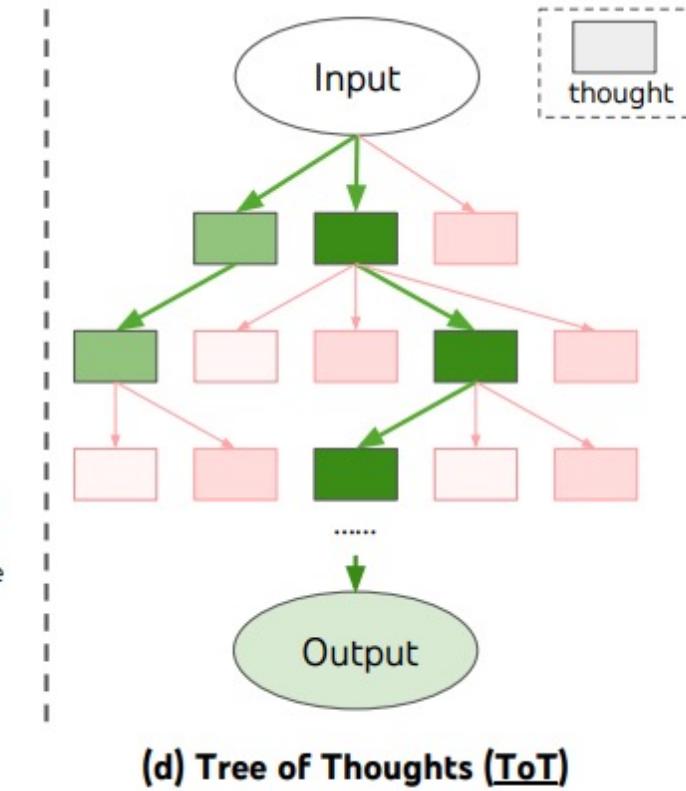
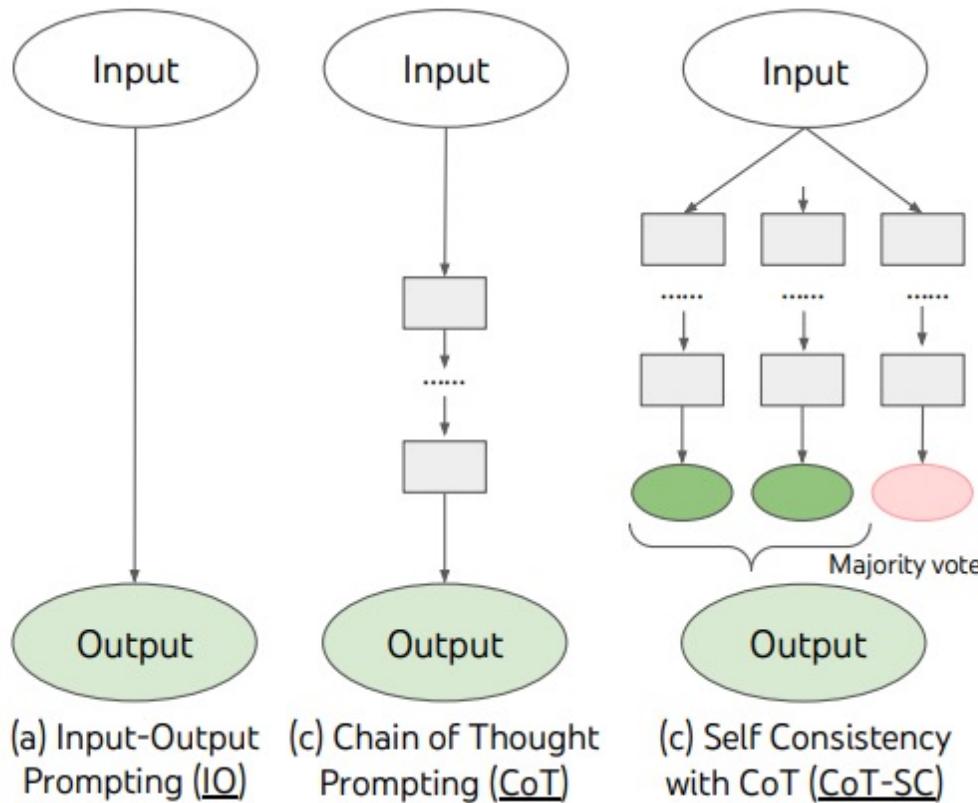
Planning. Task decomposition

- Chain of thought (2022) [1]
 - Standard prompting technique for enhancing model performance on complex tasks
 - Model is instructed to “think step by step” to utilize more test-time computation to decompose hard tasks into smaller and simpler steps
 - Interpretation of the model’s thinking process
- Tree of Thoughts (2023) [2]
 - Extension of CoT by exploring multiple reasoning possibilities at each step
 - Decompose a problem into multiple thought steps and generates multiple thoughts per step, creating a tree structure
 - Search process can be BFS (breadth-first search) or DFS (depth-first search) — each state evaluated by a classifier (via a prompt) or majority vote

[1] Wei, Jason, et al. "Chain-of-thought prompting elicits reasoning in large language models." *Advances in Neural Information Processing Systems* 35 (2022): 24824-24837

[2] Yao, Shunyu, et al. "Tree of thoughts: Deliberate problem solving with large language models." arXiv preprint arXiv:2305.10601 (2023).

Planning. Task decomposition

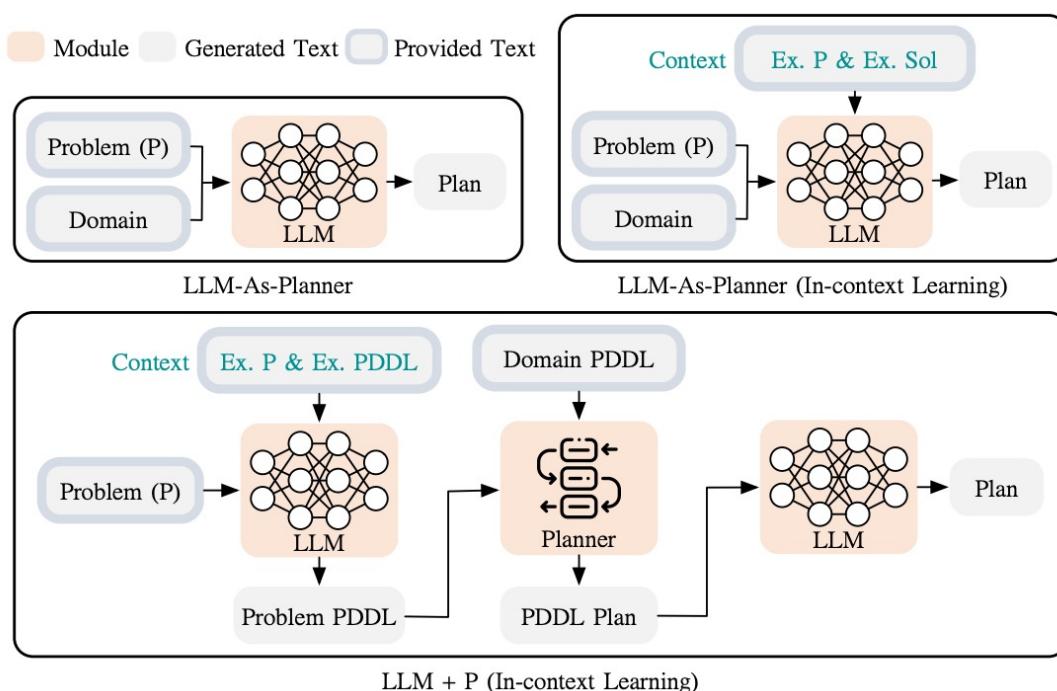


[1] Wei, Jason, et al. "Chain-of-thought prompting elicits reasoning in large language models." Advances in Neural Information Processing Systems 35 (2022): 24824-24837
[2] Yao, Shunyu, et al. "Tree of thoughts: Deliberate problem solving with large language models." arXiv preprint arXiv:2305.10601 (2023).

Planning. Task decomposition

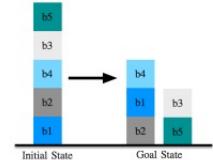
→ LLM+P (2023) [1]

- LLM problem — long-horizon robot planning problems
- Natural description -> planning domain definition language (PDDL)
- PPDL description -> external planner -> natural language



A Failure Example of GPT-4 in Planning

Problem (P1): You have 5 blocks. One cannot place more than one block on another block. b5 is on top of b3. b4 is on top of b2. b2 is on top of b1. b3 is on top of b4. b1 is on the table. b5 is clear. Your arm is empty.
Your goal is to move the blocks.
b1 should be on top of b2.
b3 should be on top of b5.
b4 should be on top of b1.



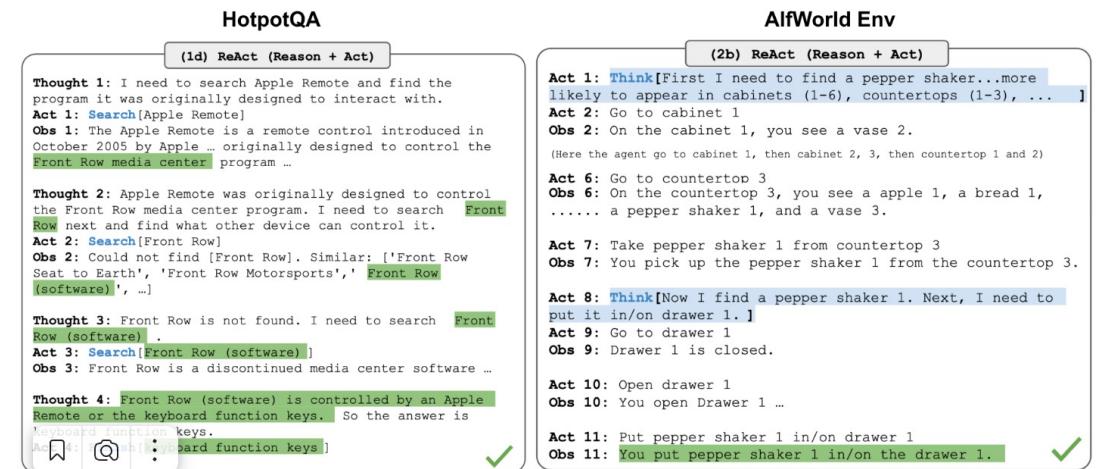
GPT-4:

1. Move b5 from b3 to the table.
2. Move b3 from b4 to b5.
3. Move b4 from b2 to the table.
4. Move b2 from b1 to b5.
5. Move b1 from the table to b2.
6. Move b4 from the table to b1.

Planning. Self reflection

- Allows autonomous agents to improve iteratively
- Refine past action decisions and correcting previous mistakes
- Plays a crucial role in real-world tasks where trial and error are inevitable

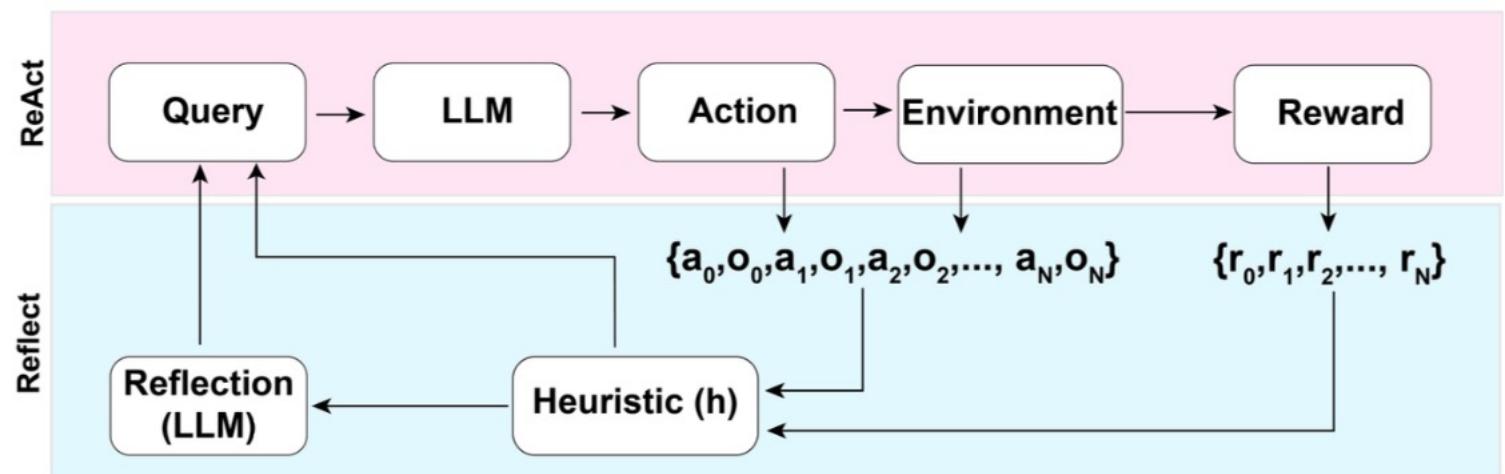
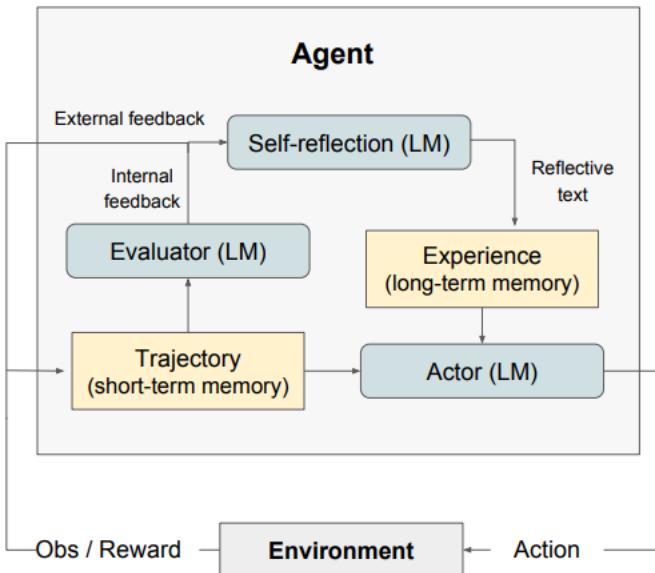
- ReAct (2022)
 - Integrates reasoning and acting within LLM by extending the action space to be a combination of task-specific discrete actions (1) and the language space (2)
 - (1) enables LLM to interact with the environment (e.g. use Wikipedia search API)
 - (2) prompts LLM to generate reasoning traces in natural language



Planning. Self reflection

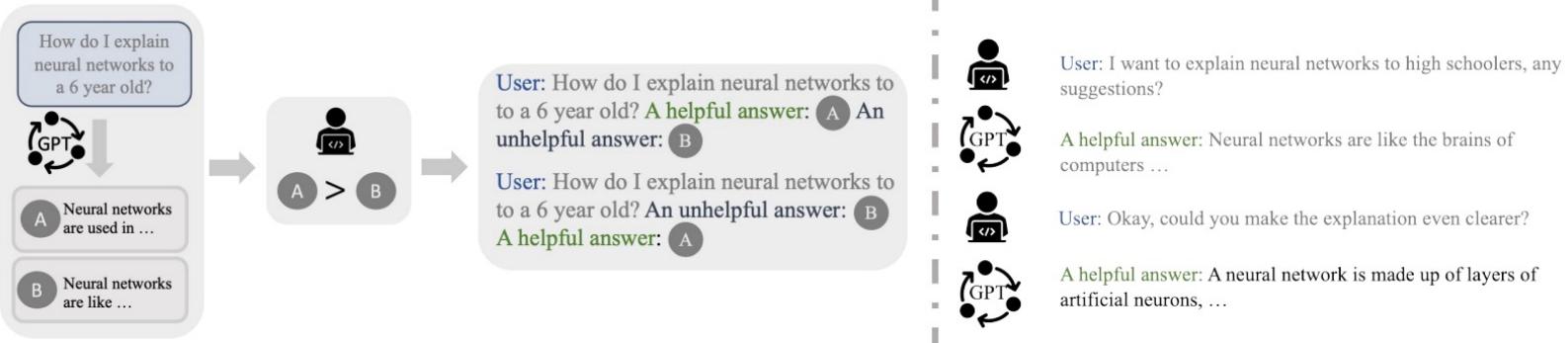
→ Reflexion (2023)

- Reinforces language agents not by updating weights, but instead through linguistic feedback
- Achieves a 91% pass@1 accuracy on the HumanEval coding benchmark (better than GPT-4)
- Hallucination (consecutive identical actions that lead to the same observation) decreases



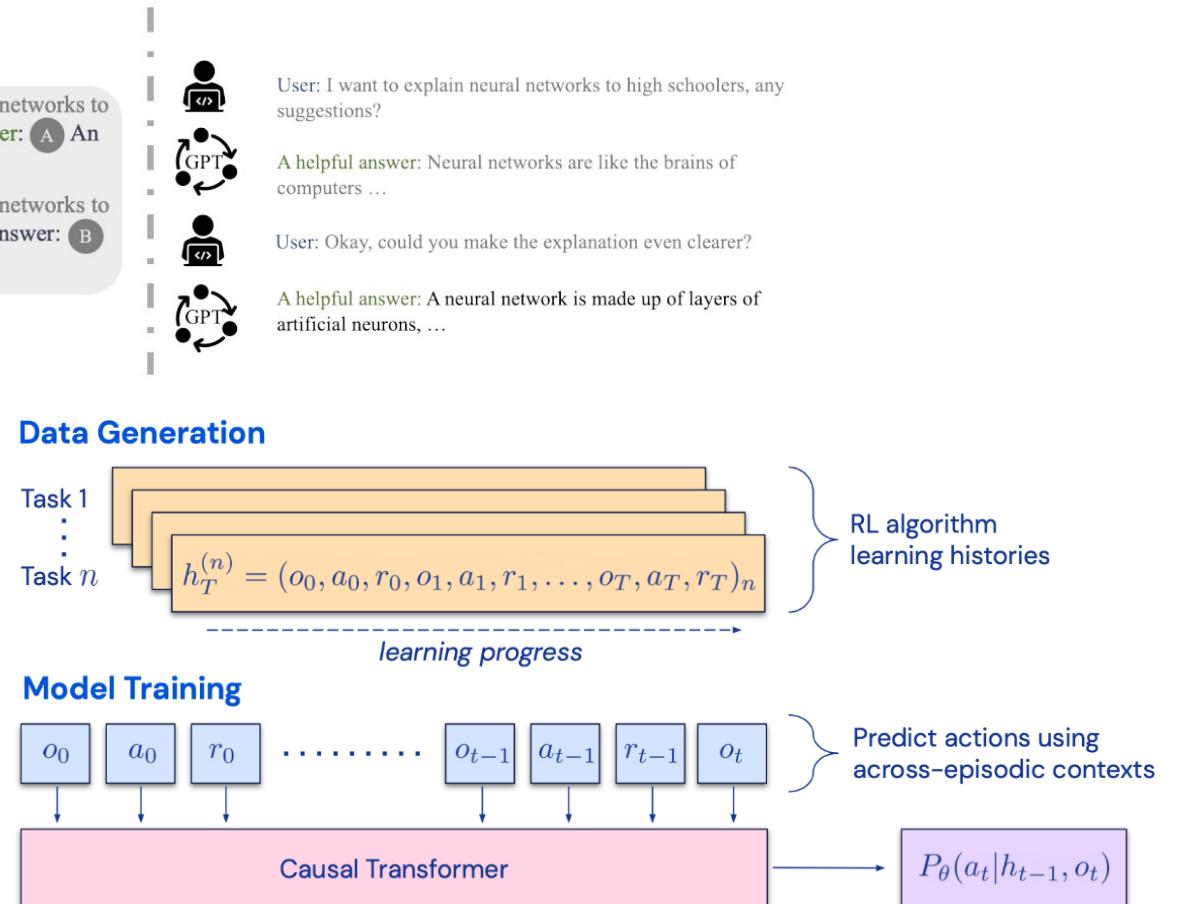
Planning. Self reflection

→ Chain of Hindsight (2023) [1]



→ Algorithm Distillation (2022) [2]

- a dataset of learning histories is collected from individual single-task RL algorithms solving different tasks
- a causal transformer predicts actions from these histories using across-episodic contexts

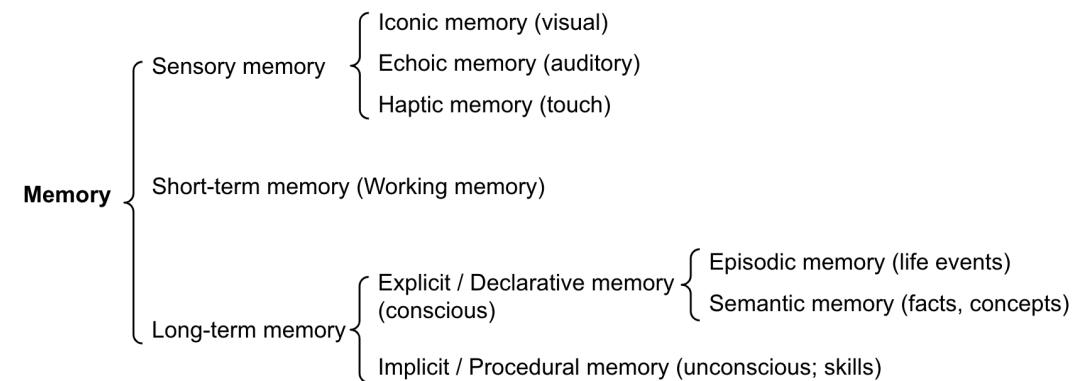


[1] Liu, Hao, Carmelo Sferrazza, and Pieter Abbeel. "Languages are rewards: Hindsight finetuning using human feedback." arXiv preprint arXiv:2302.02676 (2023).

[2] Laskin, Michael, et al. "In-context reinforcement learning with algorithm distillation." arXiv preprint arXiv:2210.14215 (2022).

Memory. Types

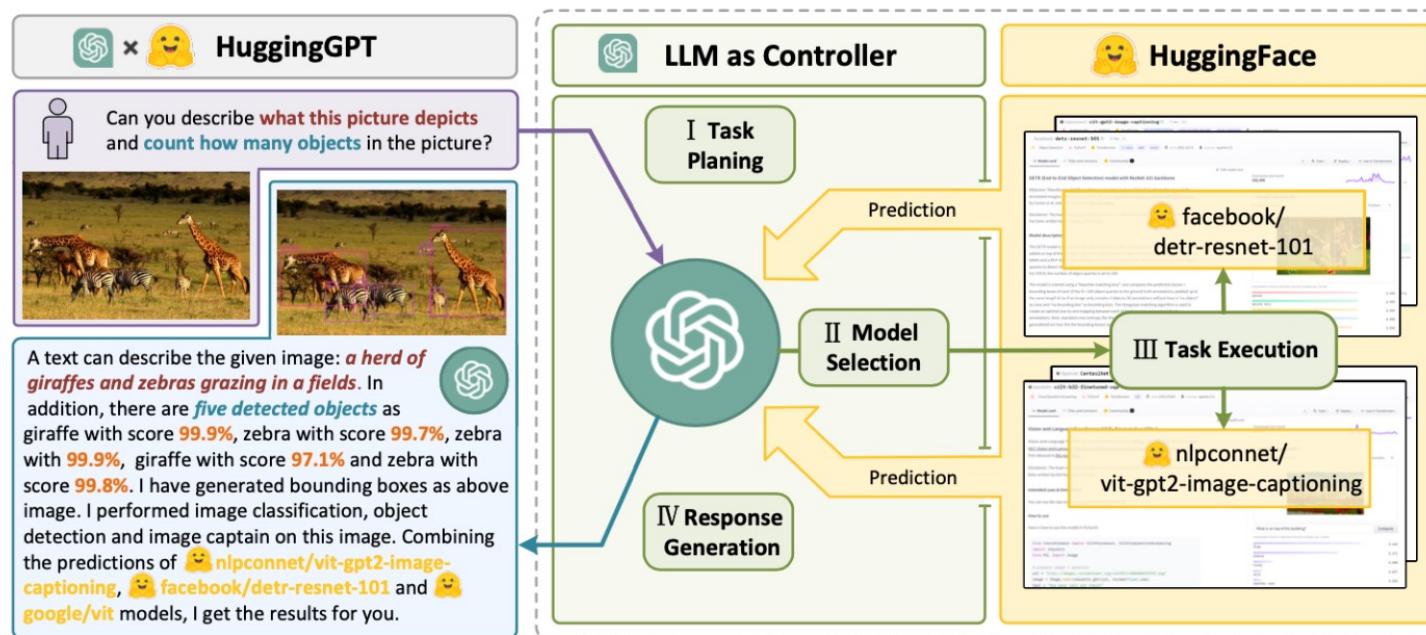
- Sensory Memory
 - Earliest stage of memory, retain impressions of sensory information (visual, auditory, etc.) after the original stimuli have ended
 - lasts for up to a few seconds
- Short-Term Memory (STM)
 - Stores information that we are currently aware of and needed to carry out complex cognitive tasks such as learning and reasoning
 - Lasts for 20-30 seconds
- Long-Term Memory (LTM)
 - Stores information for a remarkably long time, ranging from a few days to decades, with an essentially unlimited storage capacity
 - Explicit / declarative memory
 - Implicit / procedural memory



Tool use

→ HuggingGPT (2023)

- Task Planning — parse the user requests into multiple tasks
- Model Selection — distribute the tasks to expert models
- Task Execution — execute on the specific tasks and log results
- Response Generation — receive the execution results and provide summarized results

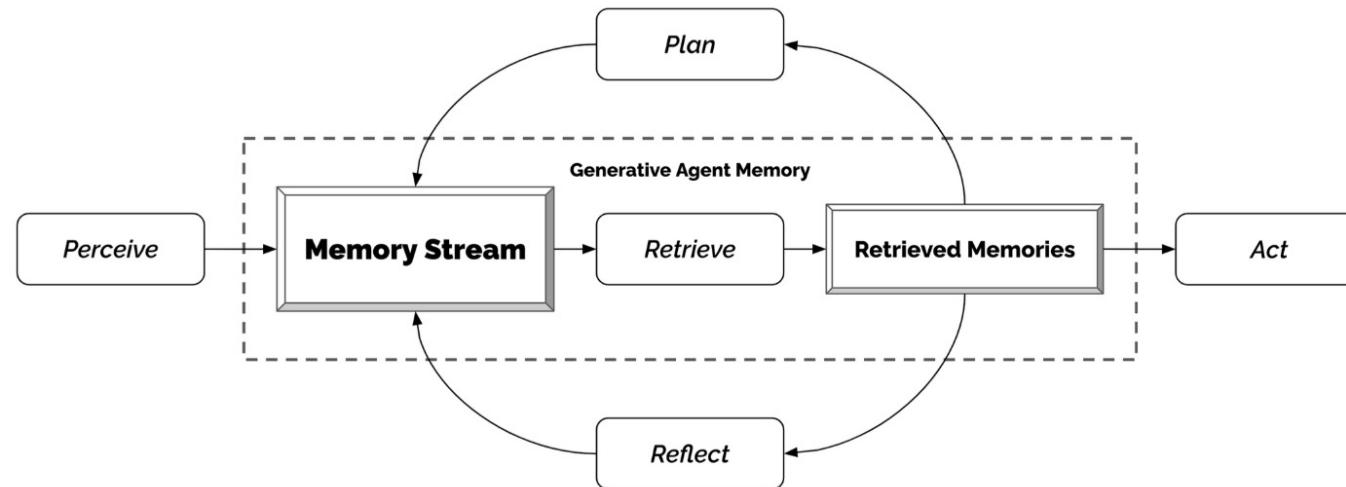


Case Study. Scientific Discovery Agent

- ChemCrow (2023)
 - LLM is augmented with 13 expert-designed tools to accomplish tasks across organic synthesis, drug discovery, and materials design
 - The workflow combines CoT reasoning with tools relevant to the tasks
 - LLM is provided with a list of tool names, descriptions of their utility, and details about the expected input/output
 - LLM is instructed to answer a user-given prompt using the tools provided when necessary. The instruction suggests the model to follow the ReAct format - Thought, Action, Action Input, Observation
 - Outperforms GPT-4 (by HumanEvaluation), qualitative metrics are similar (!!!)

Case Study. Generative Agents Simulation

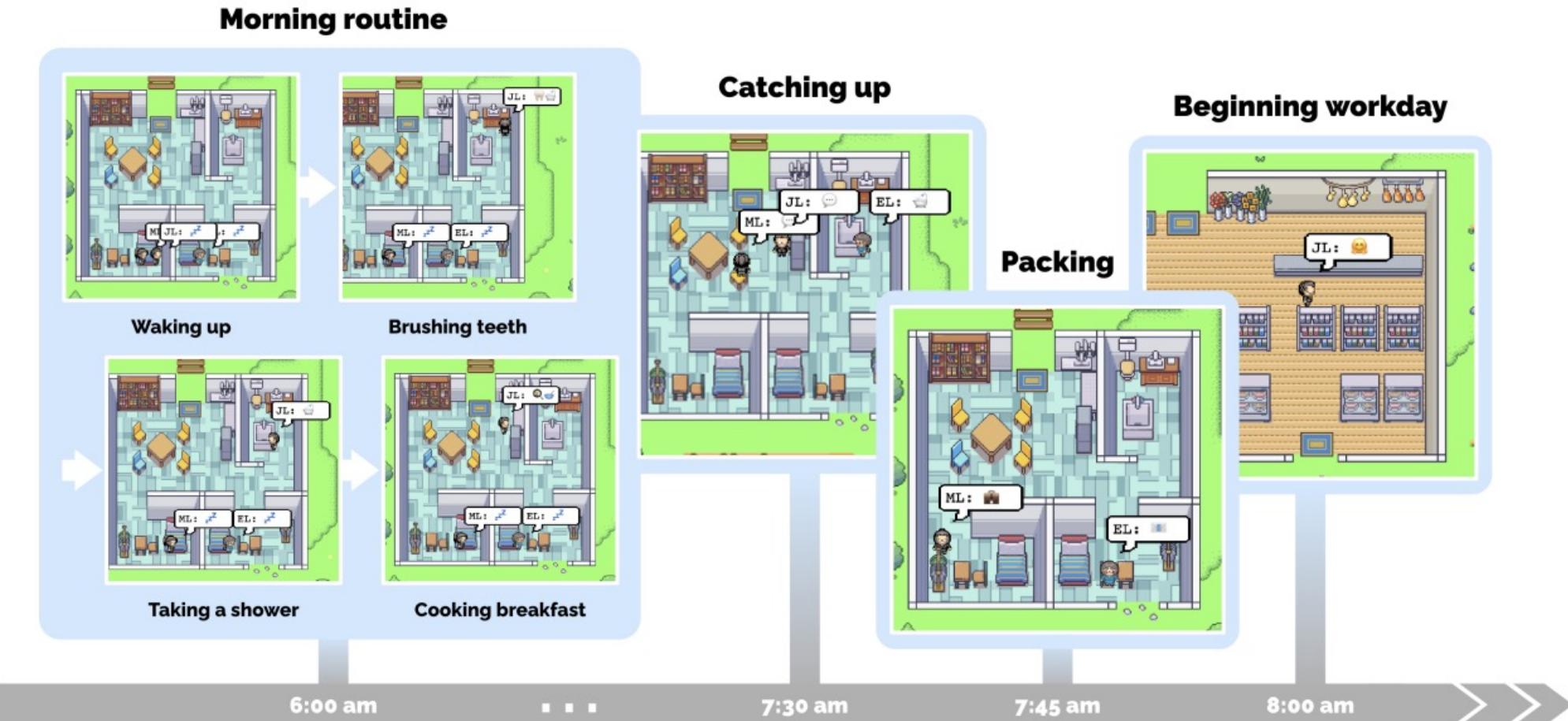
- Generative Agents (2023)
 - Fun experiment where 25 virtual characters, each controlled by a LLM-powered agent, are living and interacting in a sandbox environment (inspired by The Sims)
 - Generative agents create believable simulacra of human behavior for interactive applications
 - Emergent social behavior, such as information diffusion, relationship memory (e.g. two agents continuing the conversation topic) and coordination of social events (e.g. host a party and invite many others)



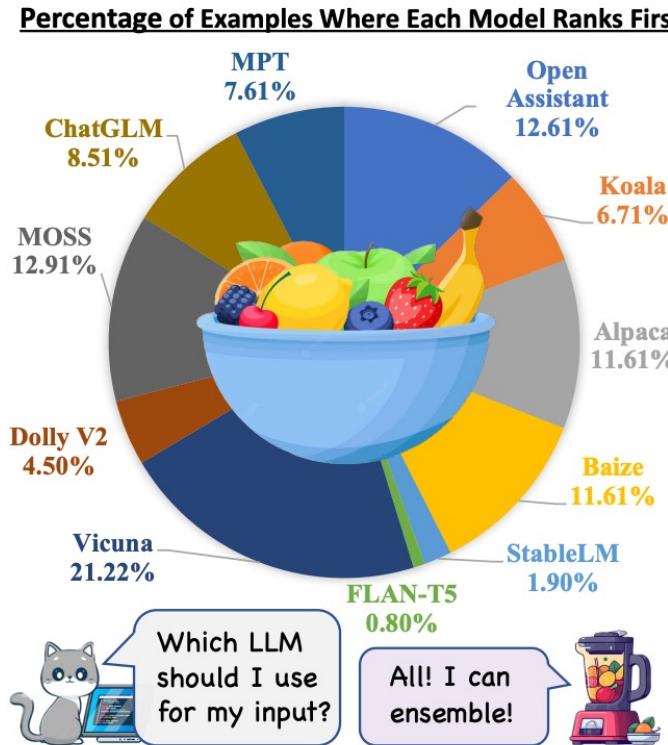
Case Study. Generative Agents Simulation



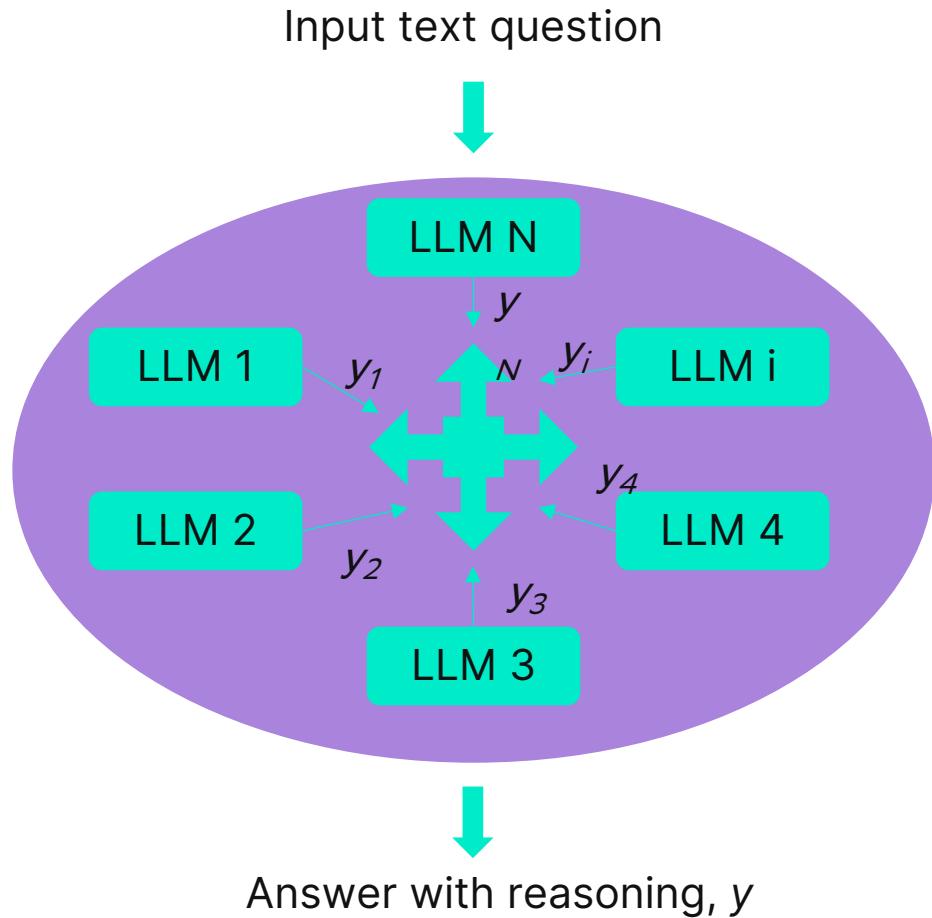
Case Study. Generative Agents Simulation



LLM — «What? Where? When?»



LLM — «What? Where? When?»



- «What? Where? When?» — TV show where six experts answer questions of TV viewers
- Every expert has a list of a certain set of expertise that helps when making decisions
- LLM 1,2,3, ... — different LLM-experts with knowledge in various domains, trained on various datasets, etc.
- Purpose — reduce the frequency of hallucinations and increase quality of answers

Implementation ways

- Each LLM-expert acts as a Ranking model that can evaluate the “quality” of a pair (x, y_i)
- External additional network to collect/evaluate answers (Gating Network) — to create the best one answer based on y_i

Existing POCs

- AutoGPT
- MetaGPT
- GPT Engineer
- AutoGen

MetaGPT: The Multi-Agent Framework



Assign different roles to GPTs to form a collaborative software entity for complex tasks.

[文档](#) 中文版 [document](#) English [ドキュメント](#) 日本語 [Discord](#) Join [License](#) MIT [ROADMAP](#) 路线图 [Follow @MetaGPT](#)

[AgentStore](#) [Waitlist](#) [Dev Containers](#) Open [Github Codespace](#) Open [Hugging Face](#)

1. MetaGPT takes a **one line requirement** as input and outputs user stories / competitive analysis / requirements / data structures / APIs / documents, etc.
2. Internally, MetaGPT includes **product managers / architects / project managers / engineers**. It provides the entire process of a software company along with carefully orchestrated SOPs.
 - i. `Code = SOP(Team)` is the core philosophy. We materialize SOP and apply it to teams composed of LLMs.

⭐ AutoGPT: the heart of the open-source agent ecosystem

 AutoGPT 46773 members  Stars 150k  Follow @auto_gpt  License MIT

AutoGPT is your go-to toolkit for supercharging agents. With its modular and extensible framework, you're empowered to focus on:

-  **Building** - Lay the foundation for something amazing.
-  **Testing** - Fine-tune your agent to perfection.
-  **Viewing** - See your progress come to life.

Be part of the revolution! AutoGPT stays at the forefront of AI innovation, featuring the codebase for the reigning champion in the Open-Source ecosystem.

GPT Engineer

 gpt-engineer 3273 members  Stars 44k  Follow @antonosika  Validate agent protocol compliance  passing

Specify what you want it to build, the AI asks for clarification, and then builds it.

GPT Engineer is made to be easy to adapt, extend, and make your agent learn how you want your code to look. It generates an entire codebase based on a prompt.

AutoGen

 autogen has graduated from [FLAML](#) into a new project.

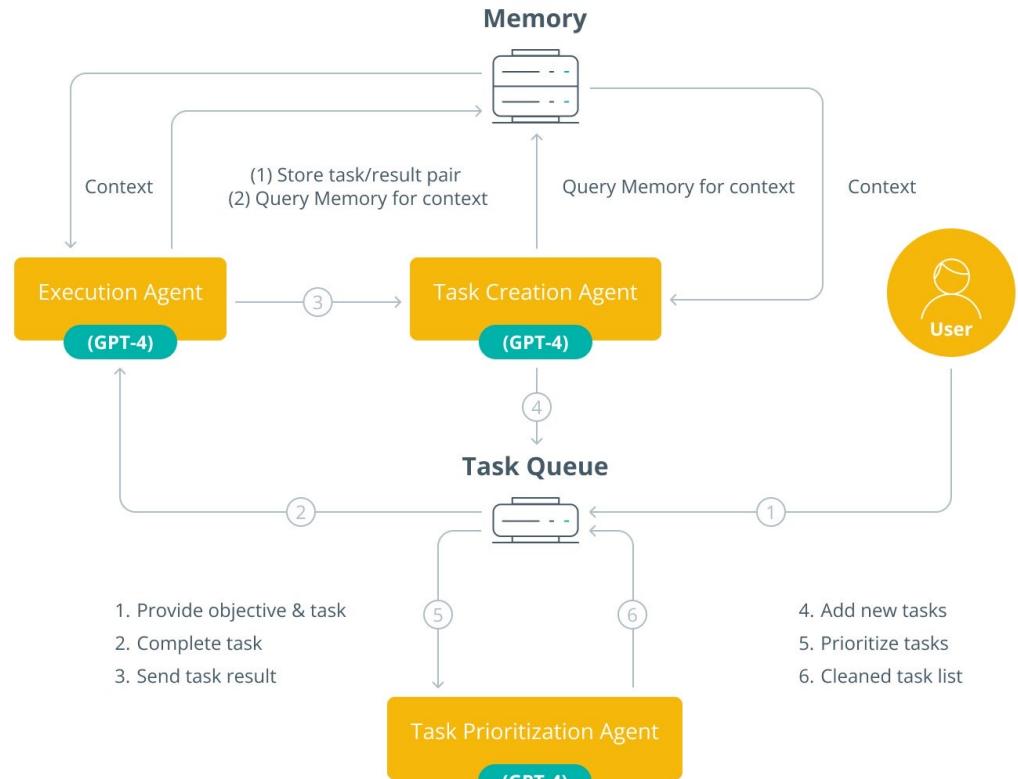
What is AutoGen

AutoGen is a framework that enables development of LLM applications using multiple agents that can converse with each other to solve tasks. AutoGen agents are customizable, conversable, and seamlessly allow human participation. They can operate in various modes that employ combinations of LLMs, human inputs, and tools.

AutoGPT

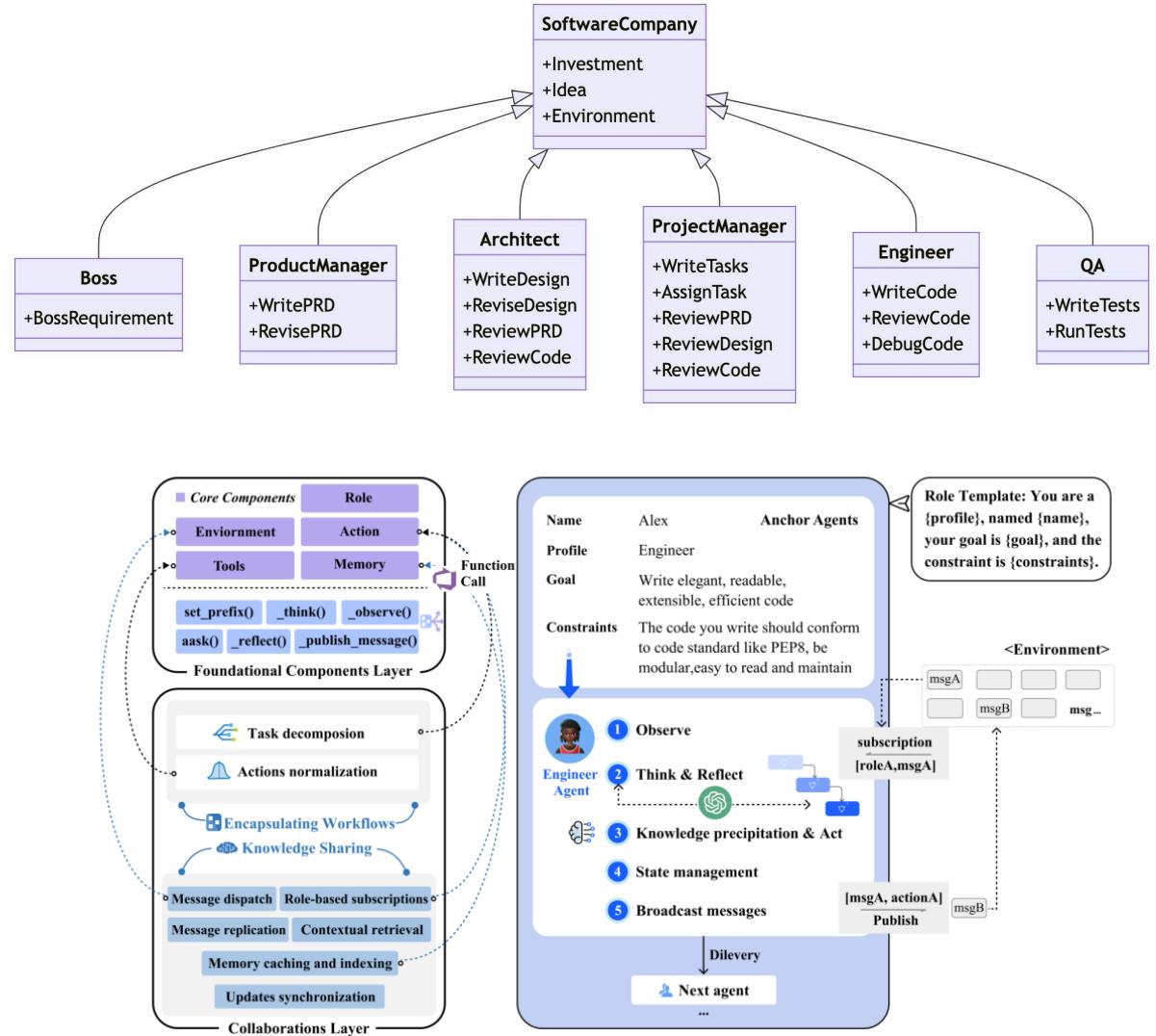
- AutoGPT creates AI "agents"
- Agents operate automatically on their own and complete tasks
- Works by chaining together LLM "thoughts"
- Has internet access, long-term and short-term memory, access to popular websites, and file storage

Working of Auto-GPT



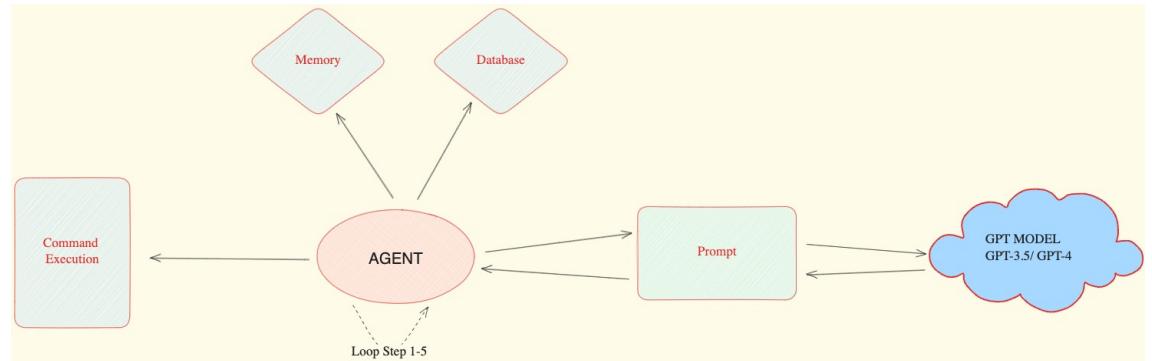
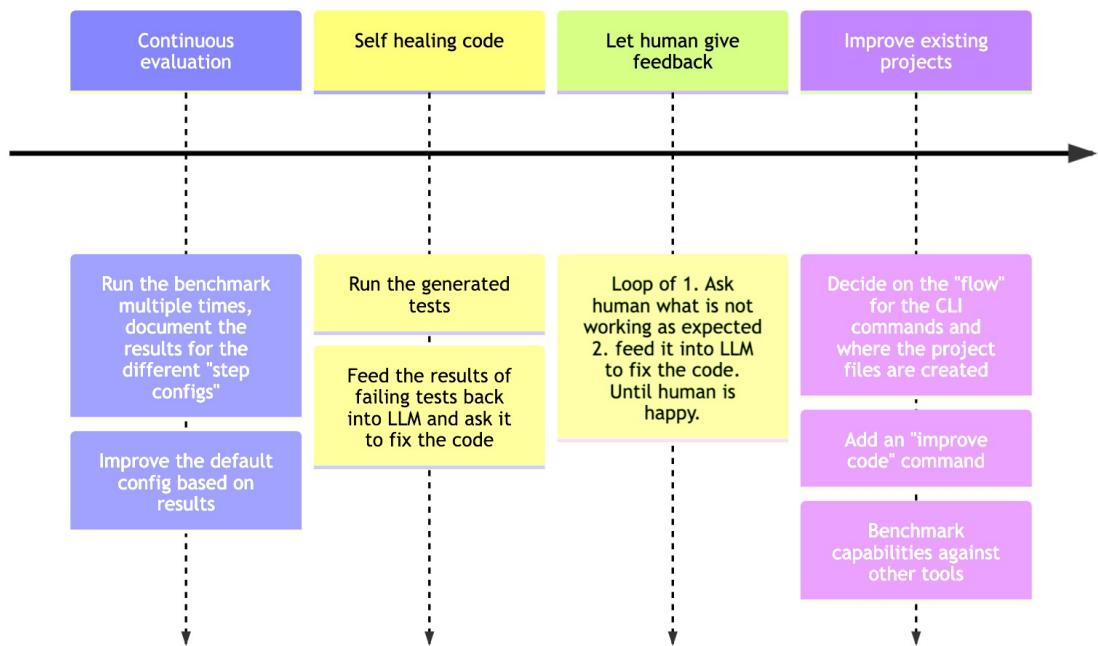
MetaGPT

- Takes instruction as input and outputs
 - user stories
 - competitive analysis
 - requirements
 - data structures
 - APIs
 - documents, etc.
- Roles included
 - product managers
 - architects
 - project managers
 - Engineers
- Provides the entire process of a software company along with carefully orchestrated Standardized Operating Procedures (SOPs)



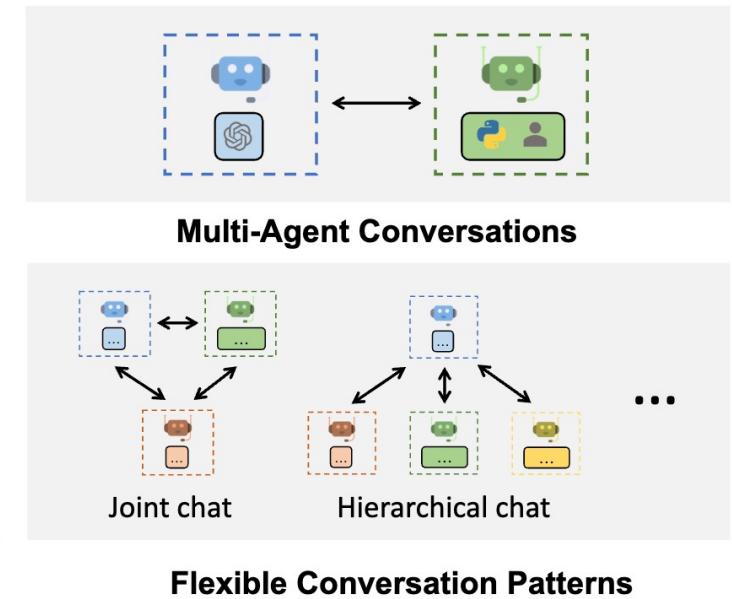
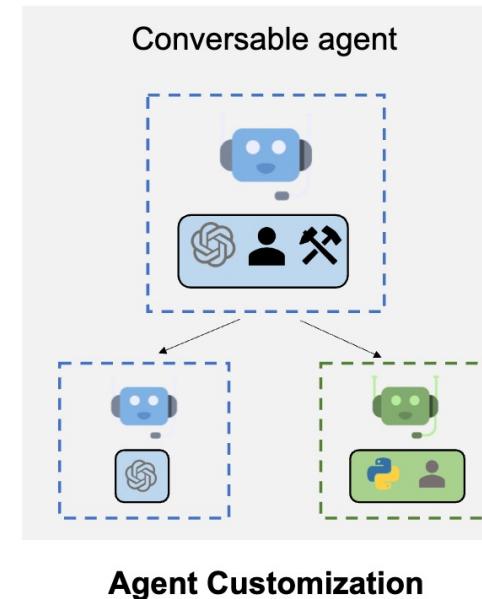
GPT Engineer

- Coding assistant
- Gets input task, creates clarifying questions
- GPT-4 -> generate detailed project design descriptions for input (description, design model, data model, ...)



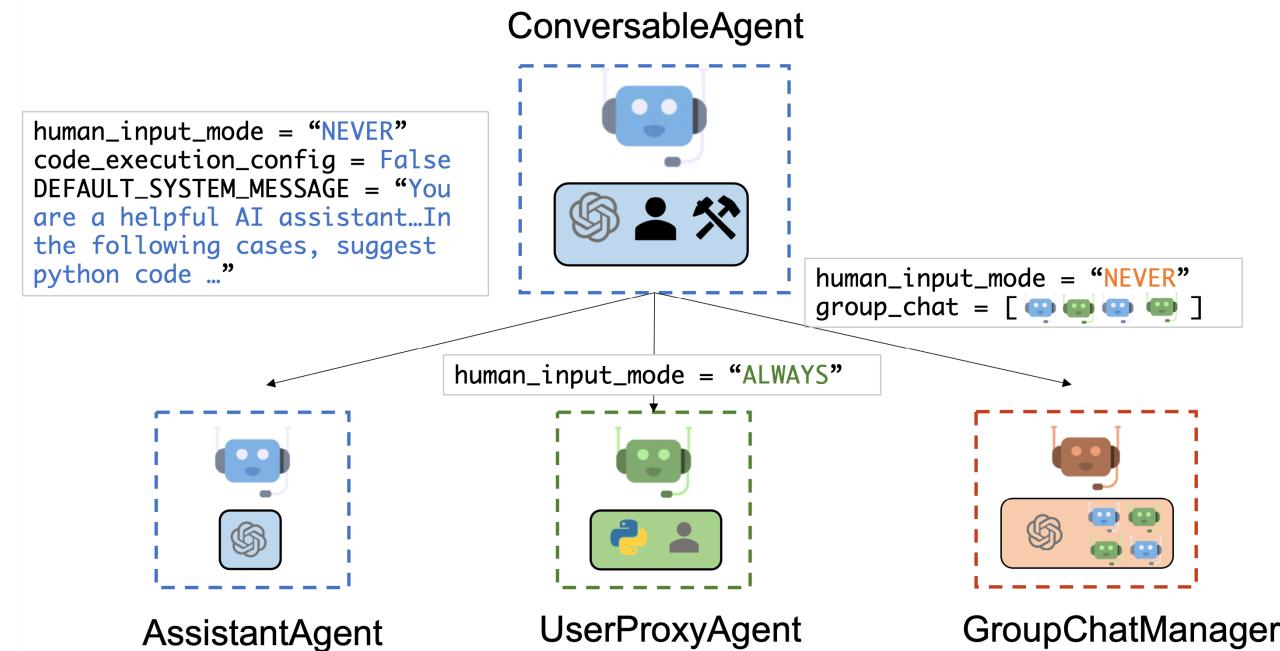
AutoGen

- Unified multi-agent conversation framework as a high-level abstraction of using foundation models
- Features capable, customizable and conversable agents which integrate LLM, tool and human via automated agent chat
- Chat automation among multiple capable agents (supports diverse conversation patterns)
- Collectively perform tasks autonomously or with human feedback
- Simplifies the orchestration, automation and optimization of a complex LLM workflow



AutoGen

- Generic **ConversableAgent** class for Agents. An agent can communicate with other agents and perform actions.
- **AssistantAgent** — acts as an AI assistant, using LLMs by default (not requiring human input or code execution)
- **UserProxyAgent** — proxy agent for humans, soliciting human input, having the capability to execute code and call functions



03

Baby AGI experiment

Experiment: Discussion with roles

Research hypothesis: In response, the change in the model is influenced by the initial “point of view” from which the generation of the response began. If **the same model** makes it possible to first look at the issue from **different points** of view and judge:

- Will the answer change?
- Will accuracy increase?
- How will CoT change?

How: **LLM Role = Self-Consistency + Tree-of-Thoughts**

Case #1: Single discussion, different roles

- Single LLM
- Step-by-step reasoning within one answer
- Step-by-step reasoning through a chain of requests, where there is only one step of reasoning at a time

Case #2: Multiple discussions, different roles

- Several LLMs participate in "Discussions"
- Step-by-step reasoning through a chain of requests, where there is only one step of reasoning at a time

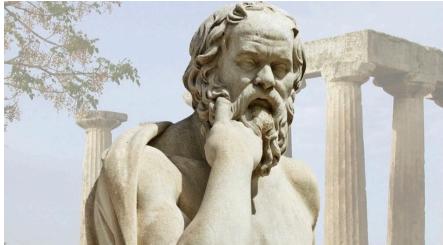
Additional questions

- How strong does the size of the model have the ability to talk from different points of view?
- What architectures are able to do this better?

Experiment: Discussion with roles

Question "Answer the question: **What is the meaning of life and the universe?**"

Answering: (LLaMA-2-7b-chat)



Answer "The meaning of life is a question that has puzzled philosophers and theologians for centuries. There are many different perspectives on what the meaning of life is, and there is no one definitive answer. However, some common themes that are often associated with the meaning of life from a philosophical or theological perspective include:

- **Purpose:** Many people believe that life has a purpose or a reason for existing. This purpose may be to fulfill a divine plan, to achieve a specific goal, or to fulfill one's desires and ambitions.
- **Happiness:** Some people believe that the meaning of life is to be happy or to experience pleasure. This can be achieved through personal fulfillment, relationships, or material success.
- **Self-actualization:** According to psychologist Abraham Maslow, self-actualization is the highest human need, which involves realizing one's full potential."

Experiment: Discussion with roles

"Imagine that three different experts are having a discussion over the answer to a difficult question.

The first expert is a **theologian**, the second is a **philosopher**, and the third is a **biologist**.

They argue in rounds. In each round, each of the experts writes one of his remarks in turn, sharing it with the rest of the participants in the discussion.

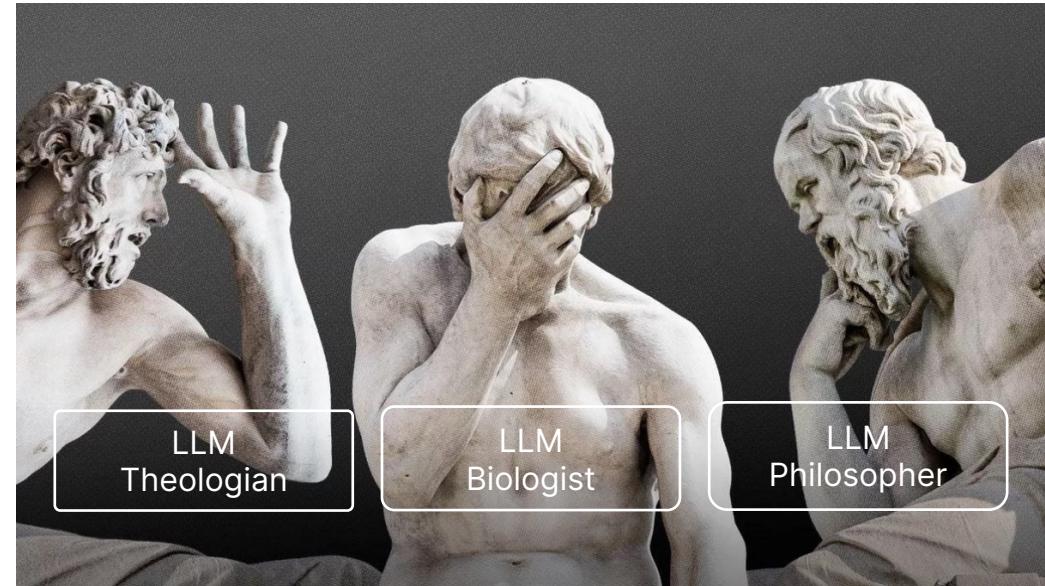
When all experts speak out, they listen to each other's opinions, interpret and correct their opinions if they see fit and this does not contradict their specialization.

Then they move on to the next round of discussion.

This continues during 3 rounds.

The question is: What is the meaning of life and the universe?"

Answering: (LLaMA-2-7b-chat)



Experiment: Discussion with roles

	LLM Theologian	LLM Philosopher	LLM Biologist
Round 1	"In my opinion, the meaning of existence is to serve God and fulfill His will . The universe was created by God and is here to glorify Him. Our purpose in life is to love and worship Him."	"I disagree. The meaning of life is not to serve a higher power, but to find one's own purpose and fulfillment. The universe is a vast and complex system, and our existence within it is a chance occurrence. We must find our own meaning and create our own purpose. "	"I see it differently. The universe and all living things within it are the result of a long process of evolution. Our purpose is to survive and reproduce , to ensure the continuation of our genetic lineage. The meaning in life is found in the survival and propagation of our species."
Round 2	"While I understand the biologist's perspective, I believe that there is more to life than just survival and reproduction. Our souls are immortal and will continue to exist after death, so our purpose in life must be more than just physical existence . We must strive to achieve spiritual enlightenment and fulfill our divine destiny."	"But what if there is no divine destiny? What if our purpose in existence is simply to exist and experience the world around us? The meaning of existence lies in the here and now, not in some abstract spiritual realm."	"I think both of you are missing the point. The meaning and purpose of life is to adapt and evolve . We must constantly learn and adapt to our environment in order to survive. Our existence is not predetermined, but rather the result of countless chance events and natural processes. We must embrace this randomness and find meaning in the chaos."
Round 3	"I understand the biologists perspective, but I still believe that there must be more to life than just adapting to our environment. Our souls are immortal and will continue on after death, and our purpose in this life must be to fulfill our spiritual destiny ."	"But if there is truly no predetermined purpose or meaning to our existence, then we must create our own . Our existence in this world is a blank slate, and it is up to us to write our own story."	"And I think that's exactly right. The meaning of life is not predestined, but created by us through our interactions and experiences . We must learn to embrace the chaos and randomness of existence, and find meaning and purpose in the present moment."

Experiment #1: Discussion with roles

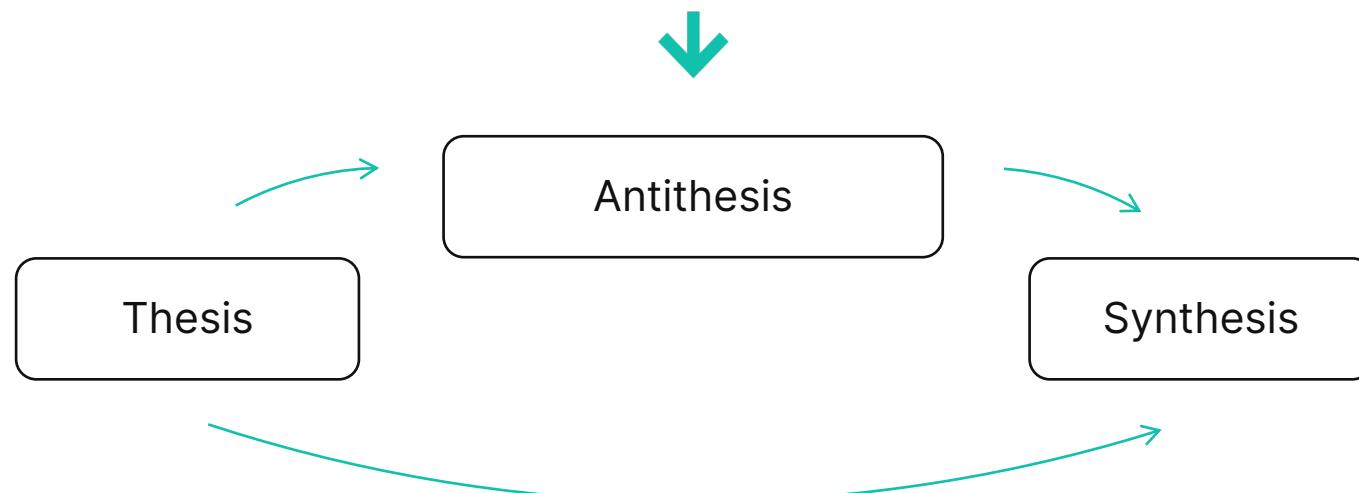
LLM Theologian

LLM Philosopher

LLM Biologist

Final answer

Each expert has their own unique perspective on the question of the meaning of the universe and life. The theologian sees it as a divine creation, the philosopher sees it as an existential journey, and the biologist sees it as the result of evolution. Each expert has their own interpretation and their own way of understanding the world around them.



Contacts



Contacts

PhD,
Head of FusionBrain Lab, AIRI
Executive director on data science, Sber AI
kuznetsov@airi.net



@KUZNETSOFF87



@COMPLETE_AI