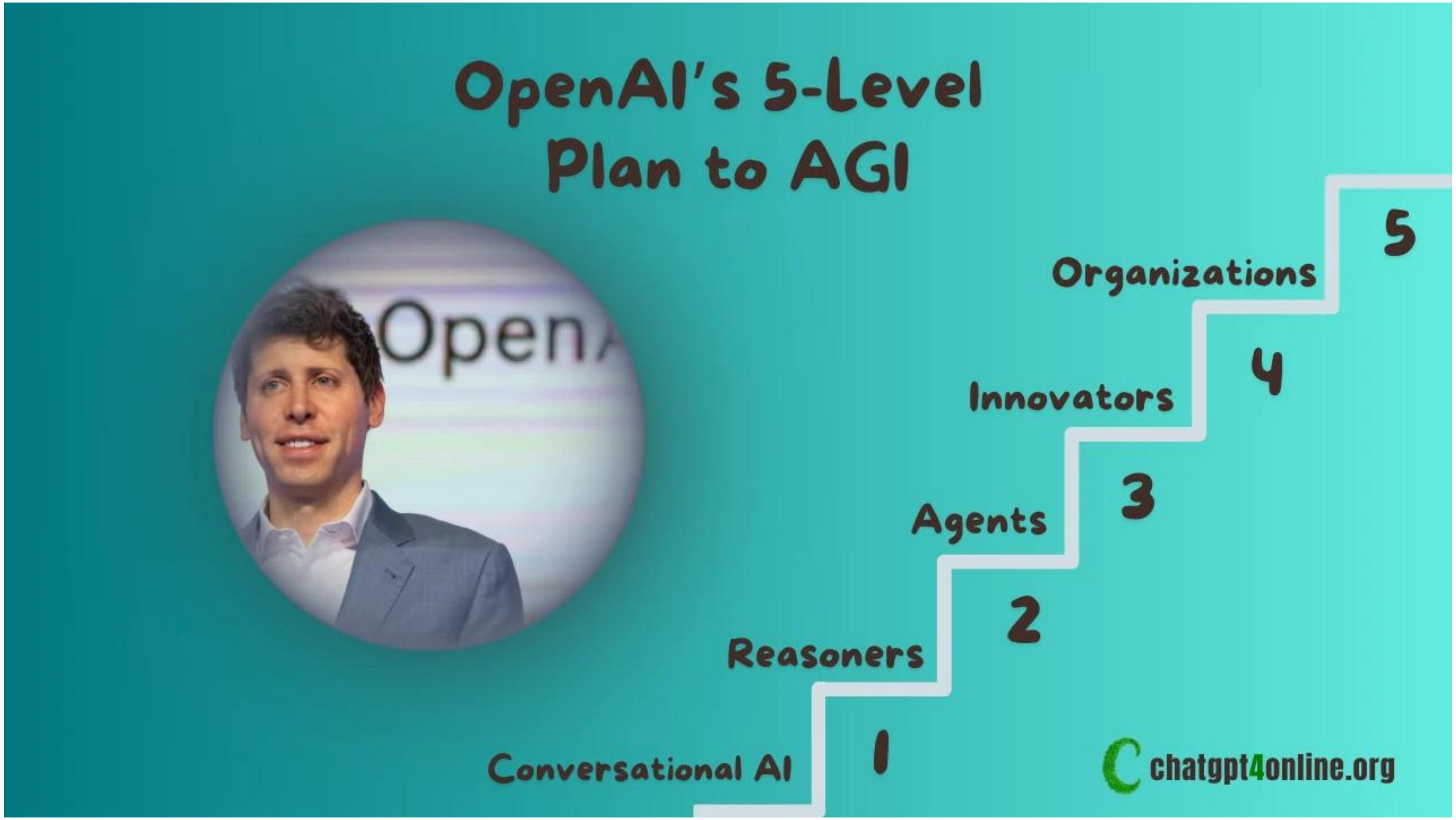
AI Agent

Xu Tan/谭旭

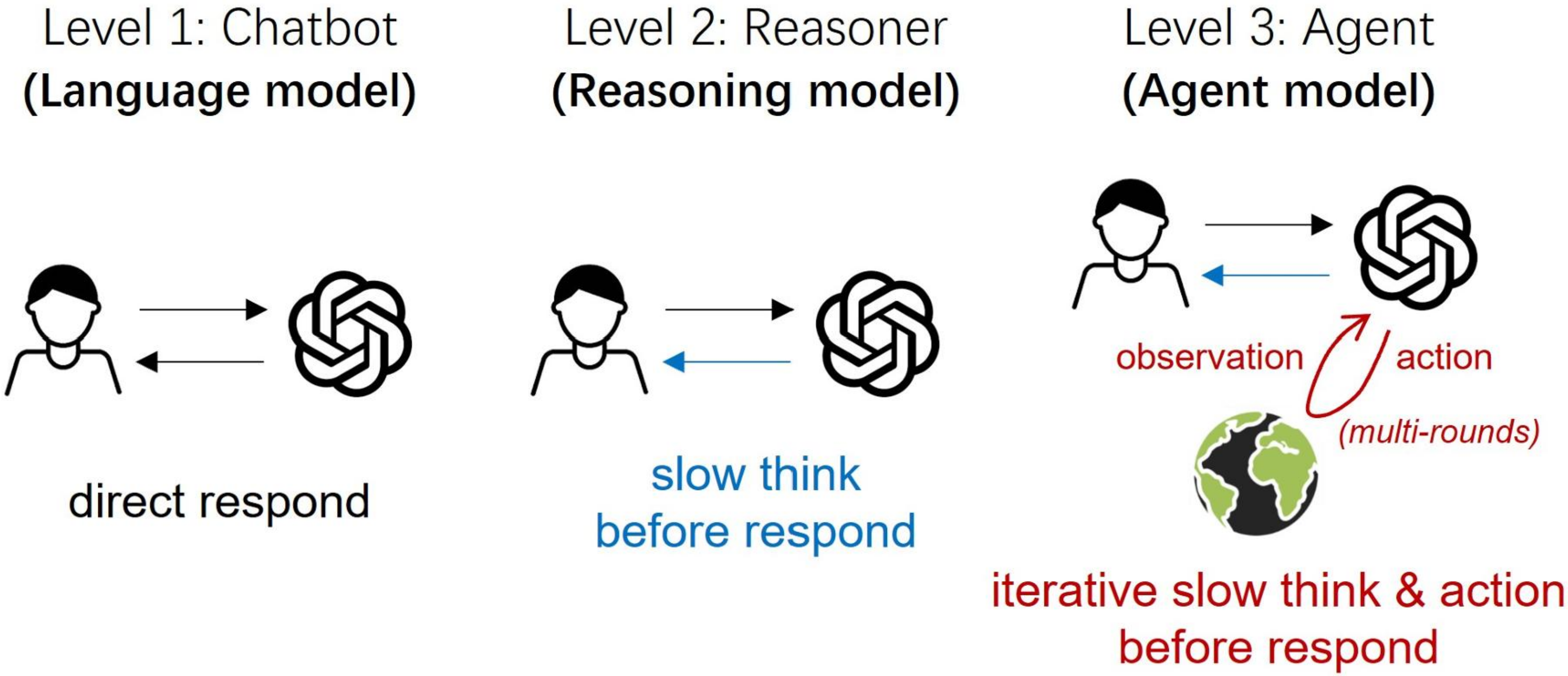
Outline

- Background
  - Agent & AGI
  - What is Agent
- Agent: Foundations
  - Key Components: Reasoning, Memory, Tool Use
  - Agentic Workflow vs Large Agent Model
- Agent: Applications
  - Search/Research Agent
  - Computer-Using Agent
  - Other Vertical Agents
- Challenges & Future Trends

AGI Roadmap


OpenAI's 5-Level Plan to AGI
OpenA
5
Organizations
4
Innovators
3
Agents
2
Reasoners
1
Conversational AI
chatgpt4online.org

AGI Roadmap: From Chatbot to Reasoner to Agent



Level 1: Chatbot

(Language model)

direct respond

Level 2: Reasoner

(Reasoning model)

slow think
before respond

Level 3: Agent

(Agent model)

observation

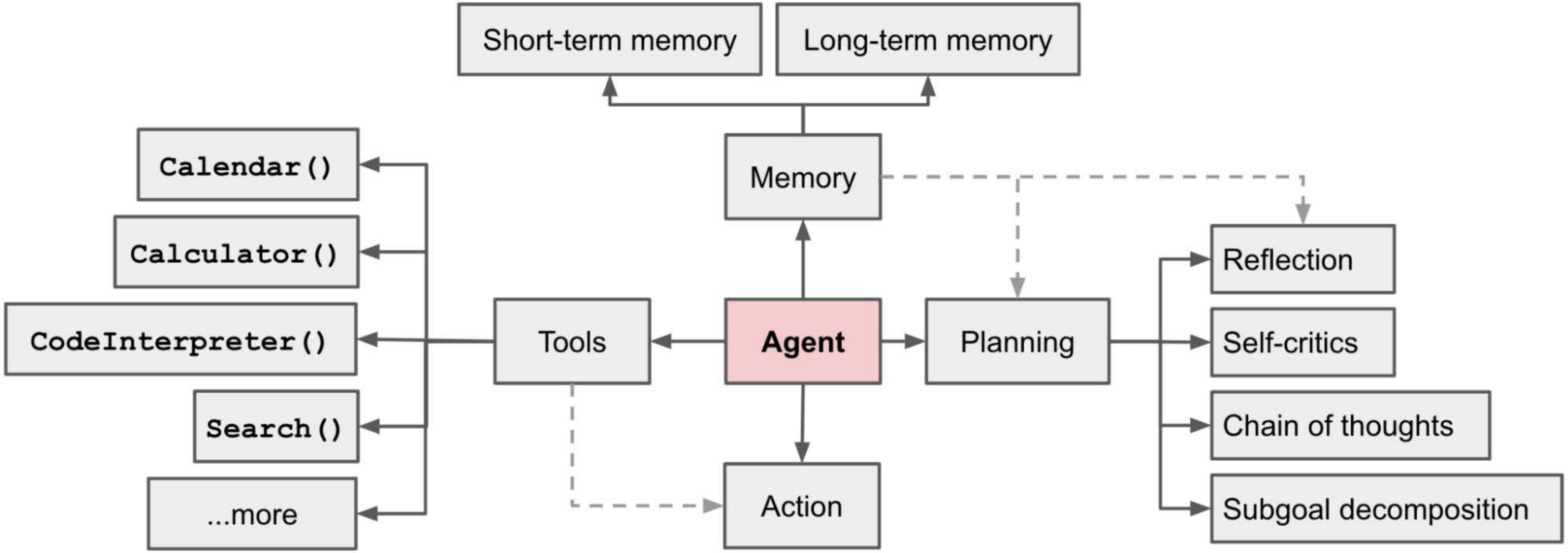
action

(multi-rounds)

iterative slow think & action
before respond

From 2503.06580

What is AI Agent

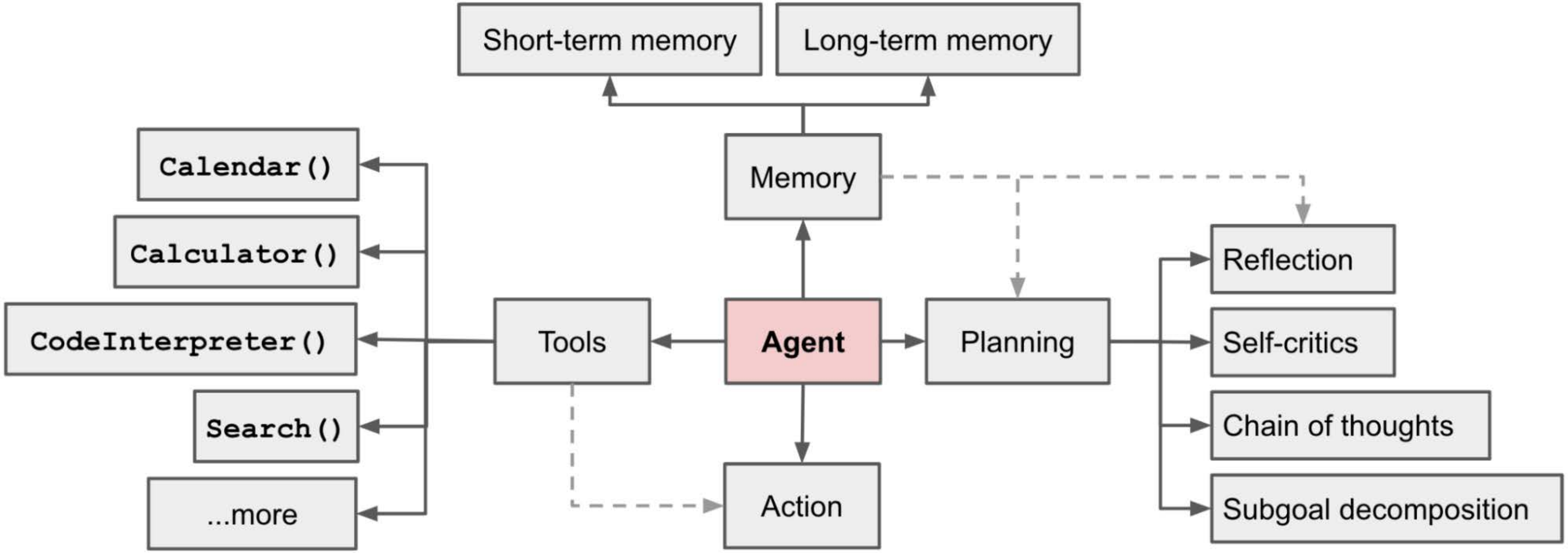


From Lilian Weng's blog

Outline

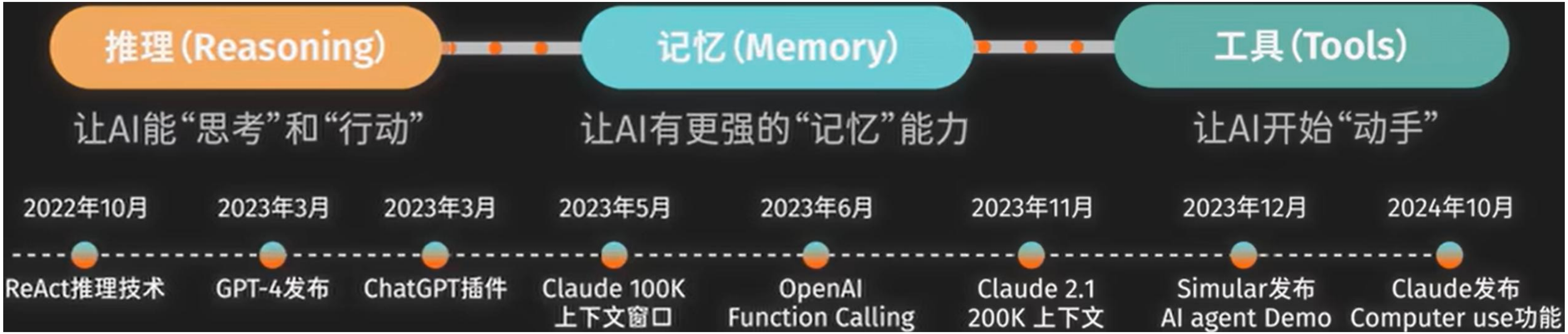
- Background
  - Agent & AGI
  - What is Agent
- Agent: Foundations
  - Key Components: Reasoning, Memory, Tool Use
  - Agentic Workflow vs Large Agent Model
- Agent: Applications
  - Search/Research Agent
  - Computer-Using Agent
  - Other Vertical Agents
- Challenges & Future Trends

Key Components of AI Agent



From Lilian Weng's blog

Progress on Reasoning/Memory/Tool



推理(Reasoning)

让AI能“思考”和“行动”

记忆(Memory)

让AI有更强的“记忆”能力

工具(Tools)

让AI开始“动手”

2022年10月 ReAct推理技术

2023年3月 GPT-4发布

2023年3月 ChatGPT插件

2023年5月 Claude 100K 上下文窗口

2023年6月 OpenAI Function Calling

2023年11月 Claude 2.1 200K 上下文

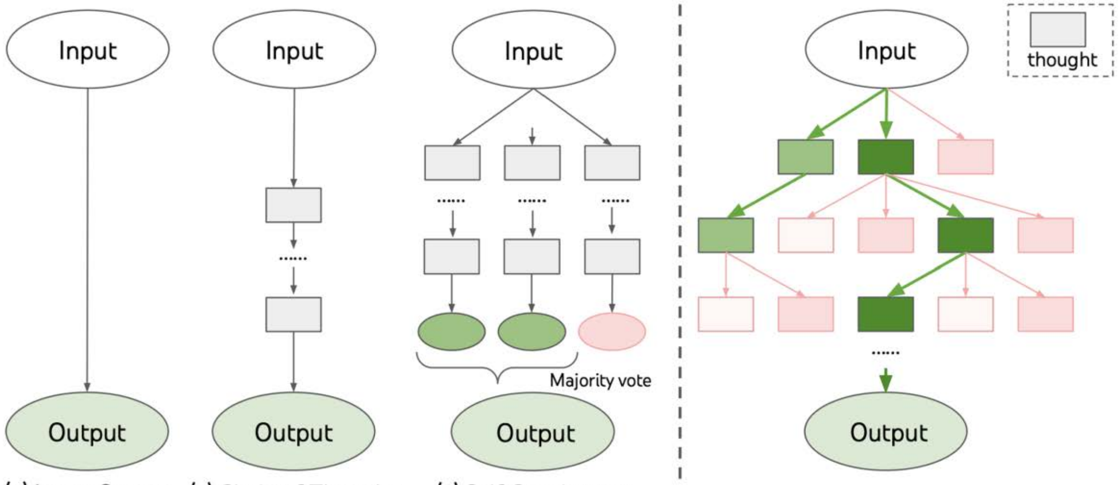
2023年12月 Simular发布 AI agent Demo

2024年10月 Claude发布 Computer use功能

From 硅谷101

Key Components of AI Agent: Reasoning

- Planning, task decomposition, chain-of-thought, reflection, reasoning
  - Multi-step reasoning: CoT step-by-step, search tree
  - Self-refection: ReAct: Synergizing reasoning and acting; Reflexion: reinforce language agents by linguistic feedback
  - Intrinsic Reasoning: OpenAI O1 & DeepSeek R1

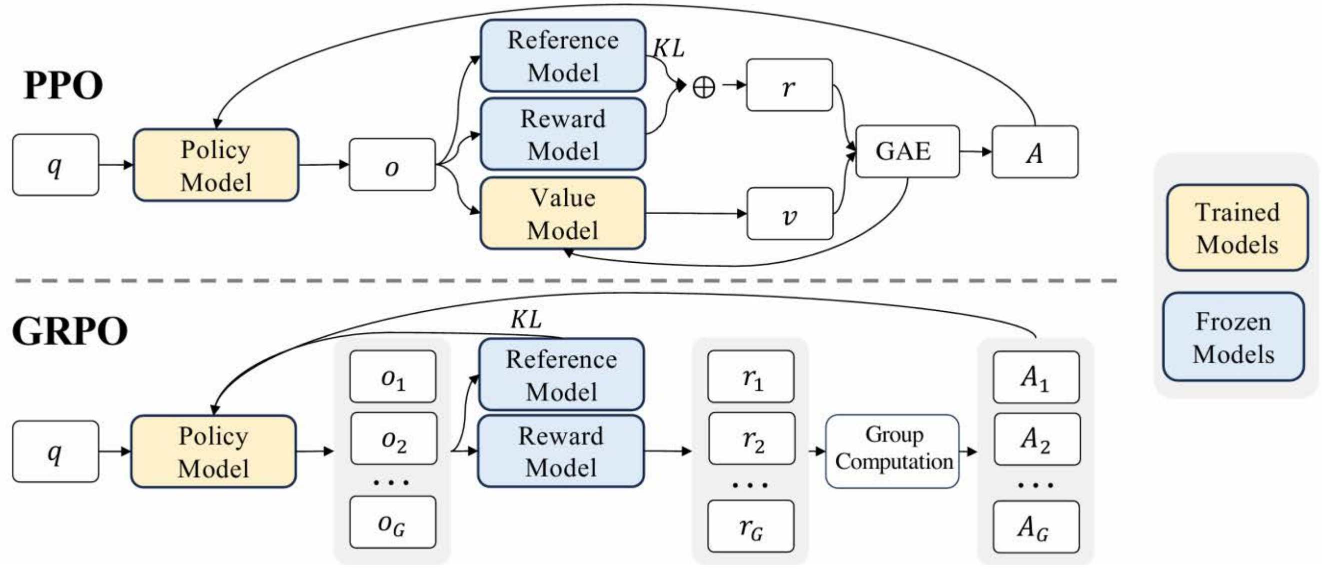


(a) Input-Output Prompting (IO)

(c) Chain of Thought Prompting (CoT)

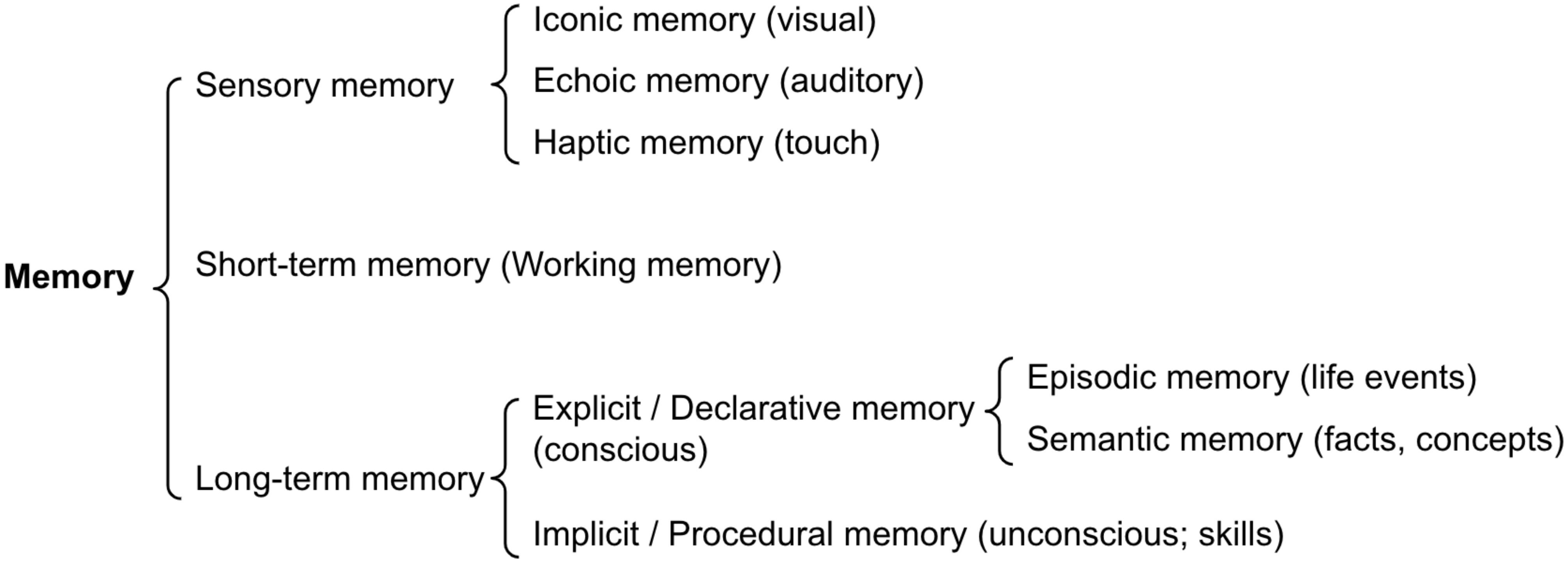
(c) Self Consistency with CoT (CoT-SC)

(d) Tree of Thoughts (ToT)



From 2305.10601 2402.03300

Key Components of AI Agent: Memory



Memory

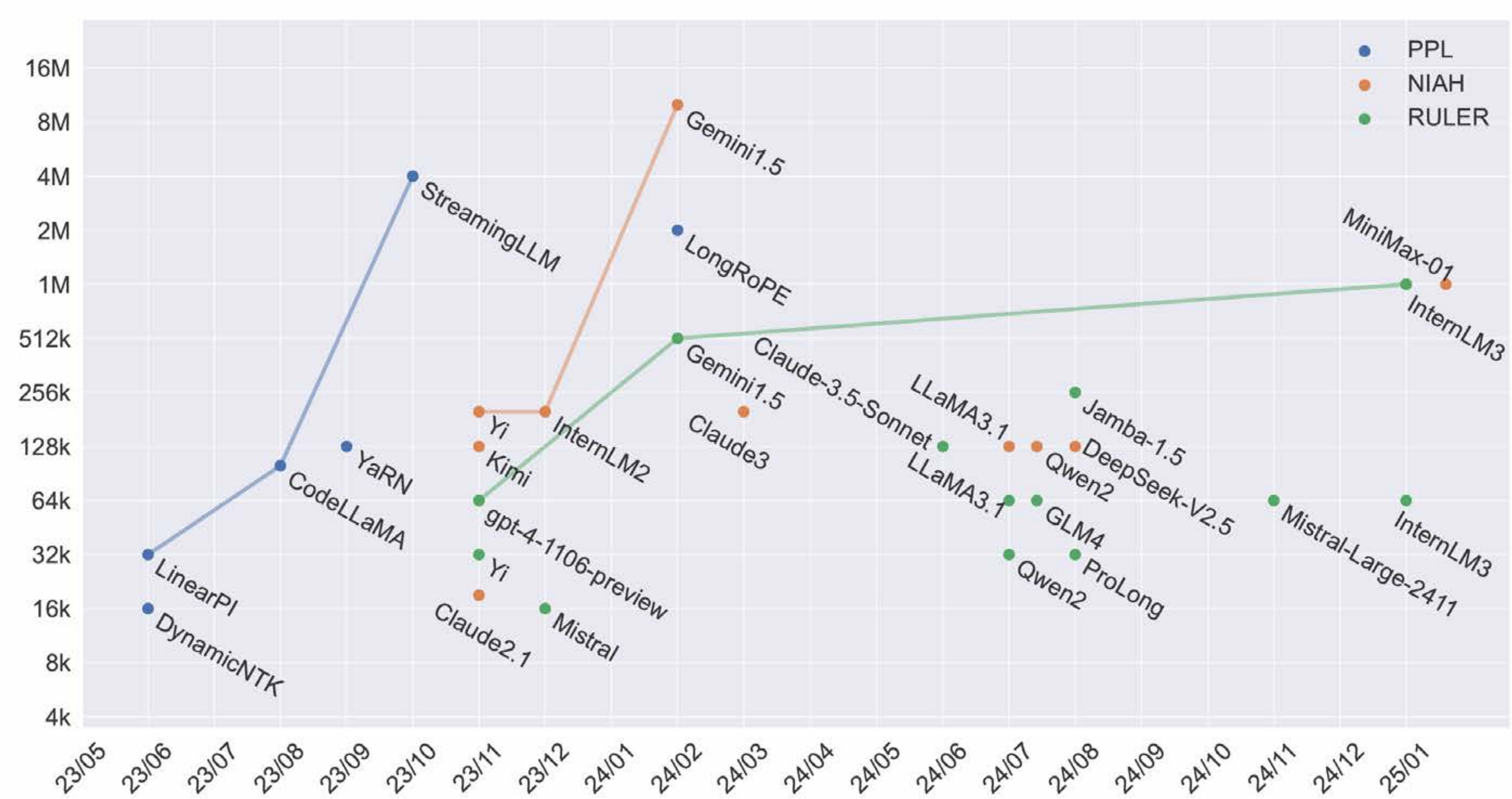
- Sensory memory
  - Iconic memory (visual)
  - Echoic memory (auditory)
  - Haptic memory (touch)
- Short-term memory (Working memory)
- Long-term memory
  - Explicit / Declarative memory (conscious)
    - Episodic memory (life events)
    - Semantic memory (facts, concepts)
  - Implicit / Procedural memory (unconscious; skills)

- Sensory memory: embedding input
- Short-term memory: in-context learning
- Long-term memory: retrieval-augmented generation (RAG)

From Lilian Weng's blog

Key Components of AI Agent: Memory

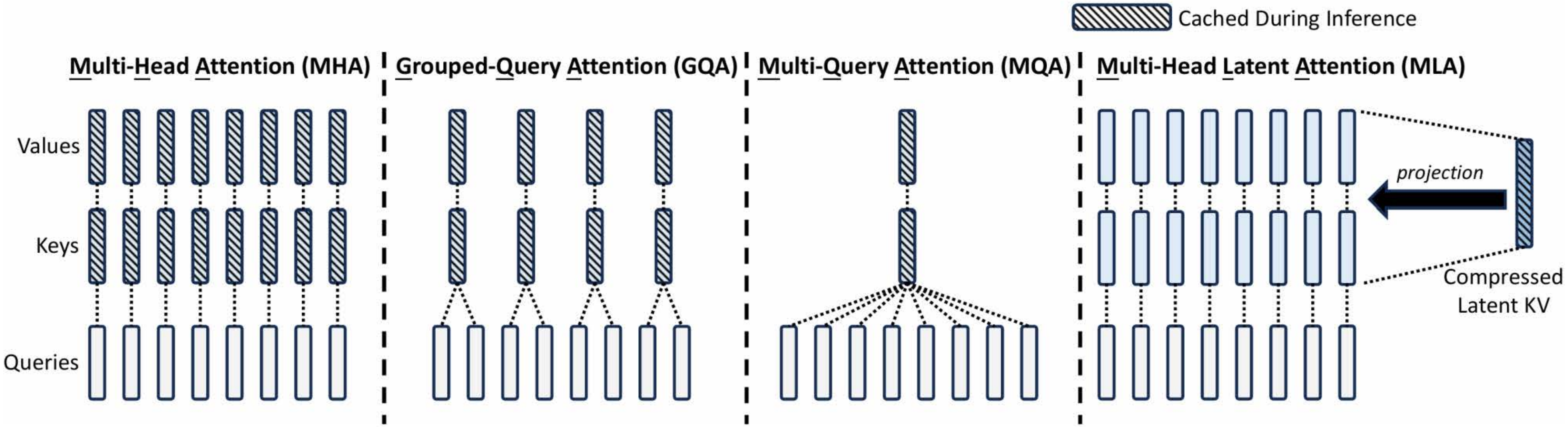
- Progress on Long-Context



From 2502.17129

Key Components of AI Agent: Memory

- Storage: Reduce KV cache memory

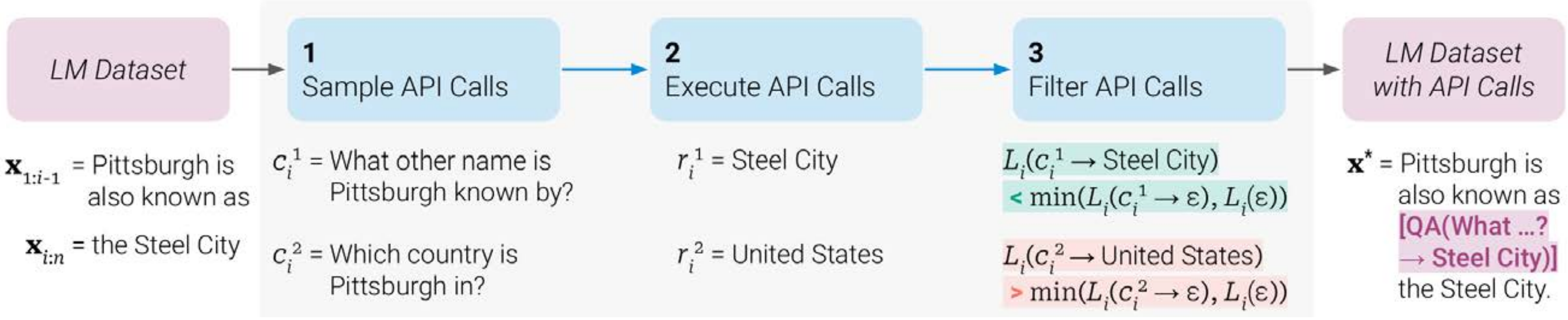

▨ Cached During Inference
Multi-Head Attention (MHA)
Grouped-Query Attention (GQA)
Multi-Query Attention (MQA)
Multi-Head Latent Attention (MLA)
Values
Keys
Queries
projection
Compressed Latent KV

- Computation: Sparse/linear attention
  - Native sparse attention, mixture of block attention
  - Mamba

From 2405.04434 2502.11089 2502.13189 2312.00752

Key Components of AI Agent: Tool

- Toolformer



LM Dataset

1 Sample API Calls

2 Execute API Calls

3 Filter API Calls

LM Dataset with API Calls

$\mathbf{x}_{1:i-1}$ = Pittsburgh is also known as

$\mathbf{x}_{i:n}$ = the Steel City

c_i^1 = What other name is Pittsburgh known by?

c_i^2 = Which country is Pittsburgh in?

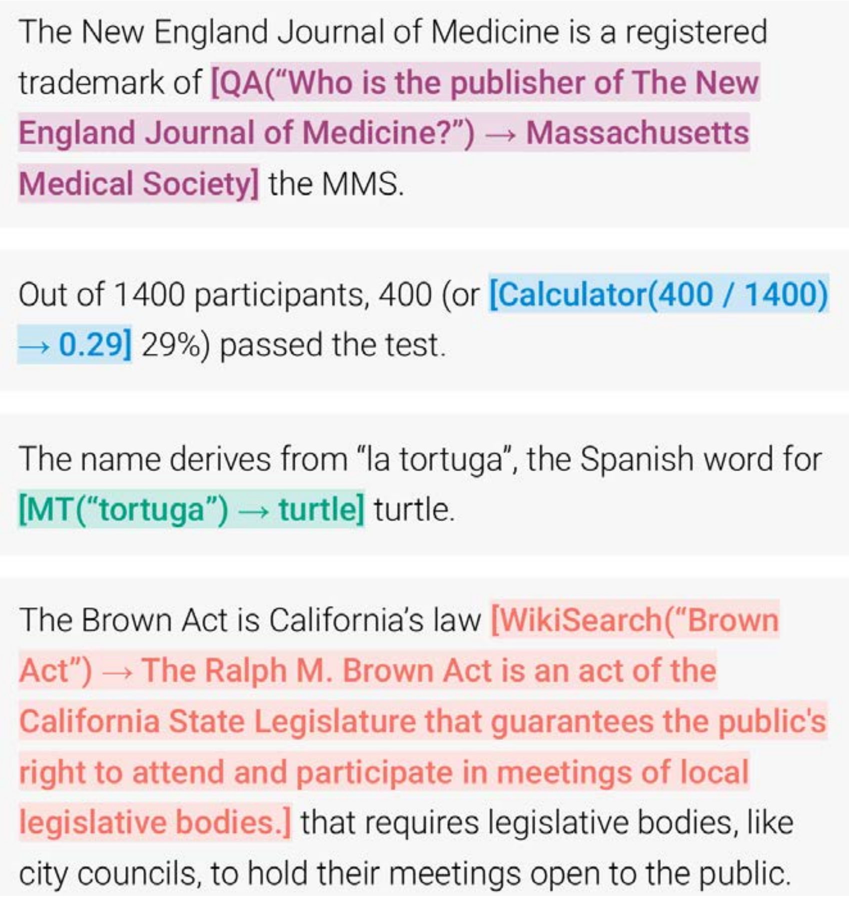
r_i^1 = Steel City

r_i^2 = United States

$L_i(c_i^1 \rightarrow \text{Steel City})$ $< \min(L_i(c_i^1 \rightarrow \varepsilon), L_i(\varepsilon))$

$L_i(c_i^2 \rightarrow \text{United States})$ $> \min(L_i(c_i^2 \rightarrow \varepsilon), L_i(\varepsilon))$

$\mathbf{x}^*$ = Pittsburgh is also known as [QA(What …? → Steel City)] the Steel City.



The New England Journal of Medicine is a registered trademark of [QA("Who is the publisher of The New England Journal of Medicine?") → Massachusetts Medical Society] the MMS.

Out of 1400 participants, 400 (or [Calculator(400 / 1400) → 0.29] 29%) passed the test.

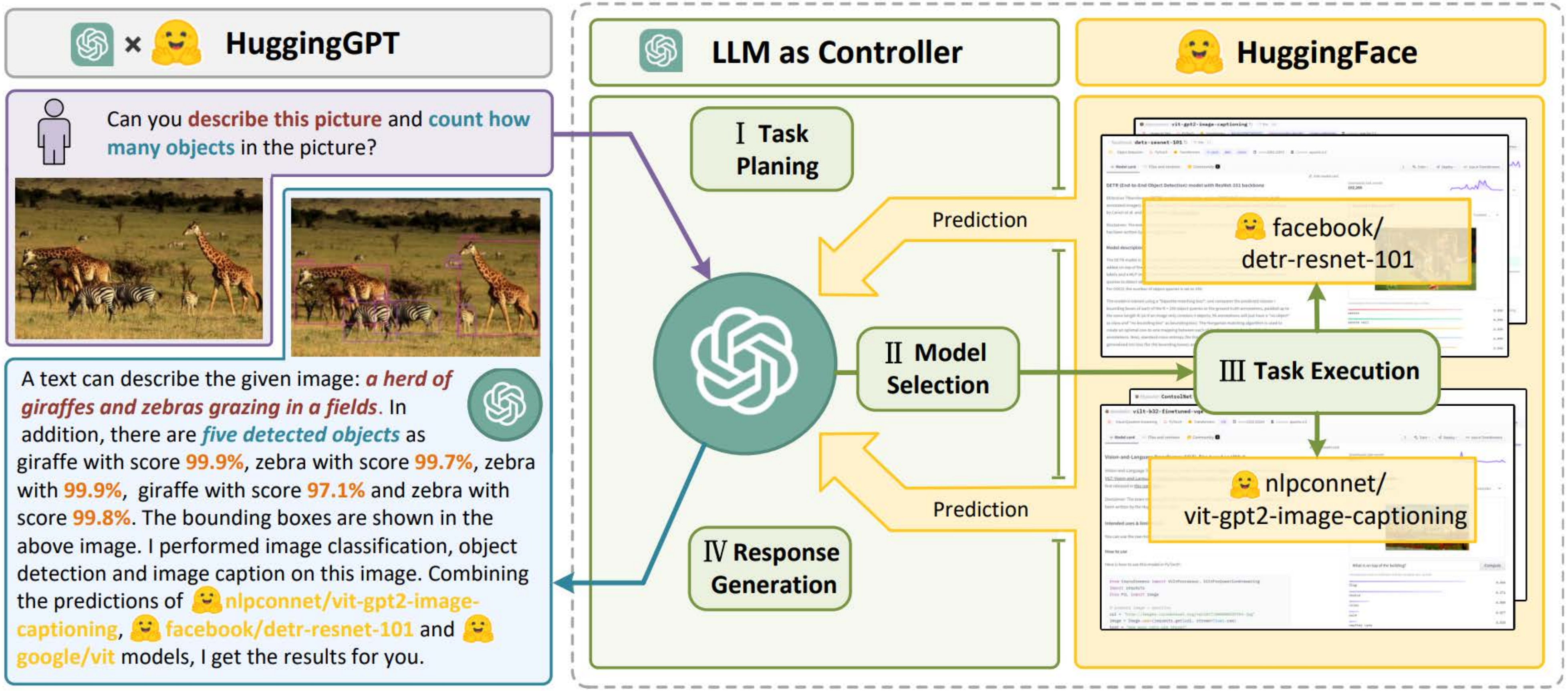
The name derives from "la tortuga", the Spanish word for [MT("tortuga") → turtle] turtle.

The Brown Act is California's law [WikiSearch("Brown Act") → The Ralph M. Brown Act is an act of the California State Legislature that guarantees the public's right to attend and participate in meetings of local legislative bodies.] that requires legislative bodies, like city councils, to hold their meetings open to the public.

From 2302.04761

Key Components of AI Agent: Tool

- HuggingGPT
  - Task planning, model selection, task execution, response generation



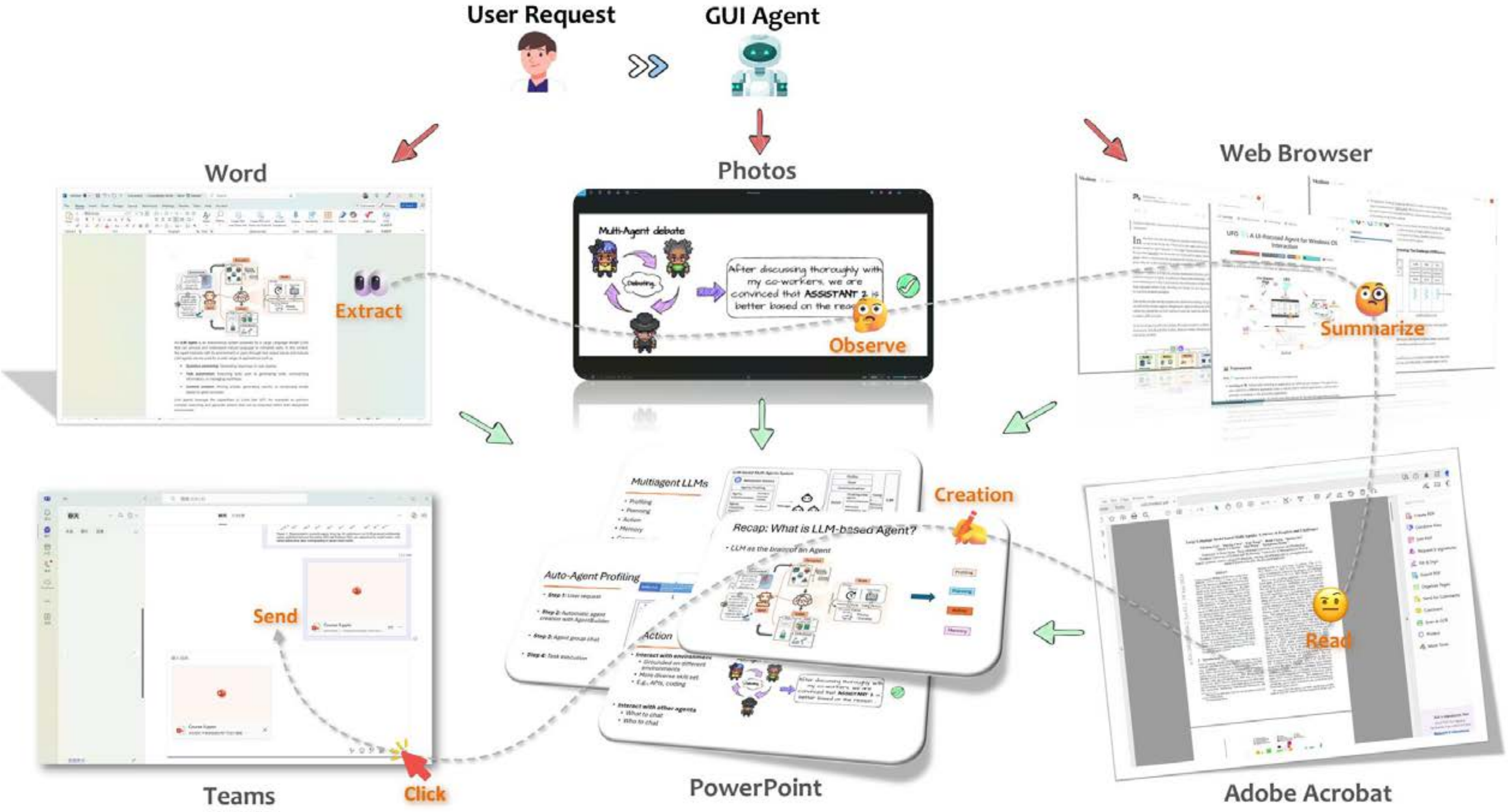
Can you **describe this picture** and **count how many objects** in the picture?

A text can describe the given image: ***a herd of giraffes and zebras grazing in a fields***. In addition, there are ***five detected objects*** as giraffe with score 99.9%, zebra with score 99.7%, zebra with 99.9%, giraffe with score 97.1% and zebra with score 99.8%. The bounding boxes are shown in the above image. I performed image classification, object detection and image caption on this image. Combining the predictions of nlpconnet/vit-gpt2-image-captioning, facebook/detr-resnet-101 and google/vit models, I get the results for you.

From 2303.17580

Key Components of AI Agent: Tool

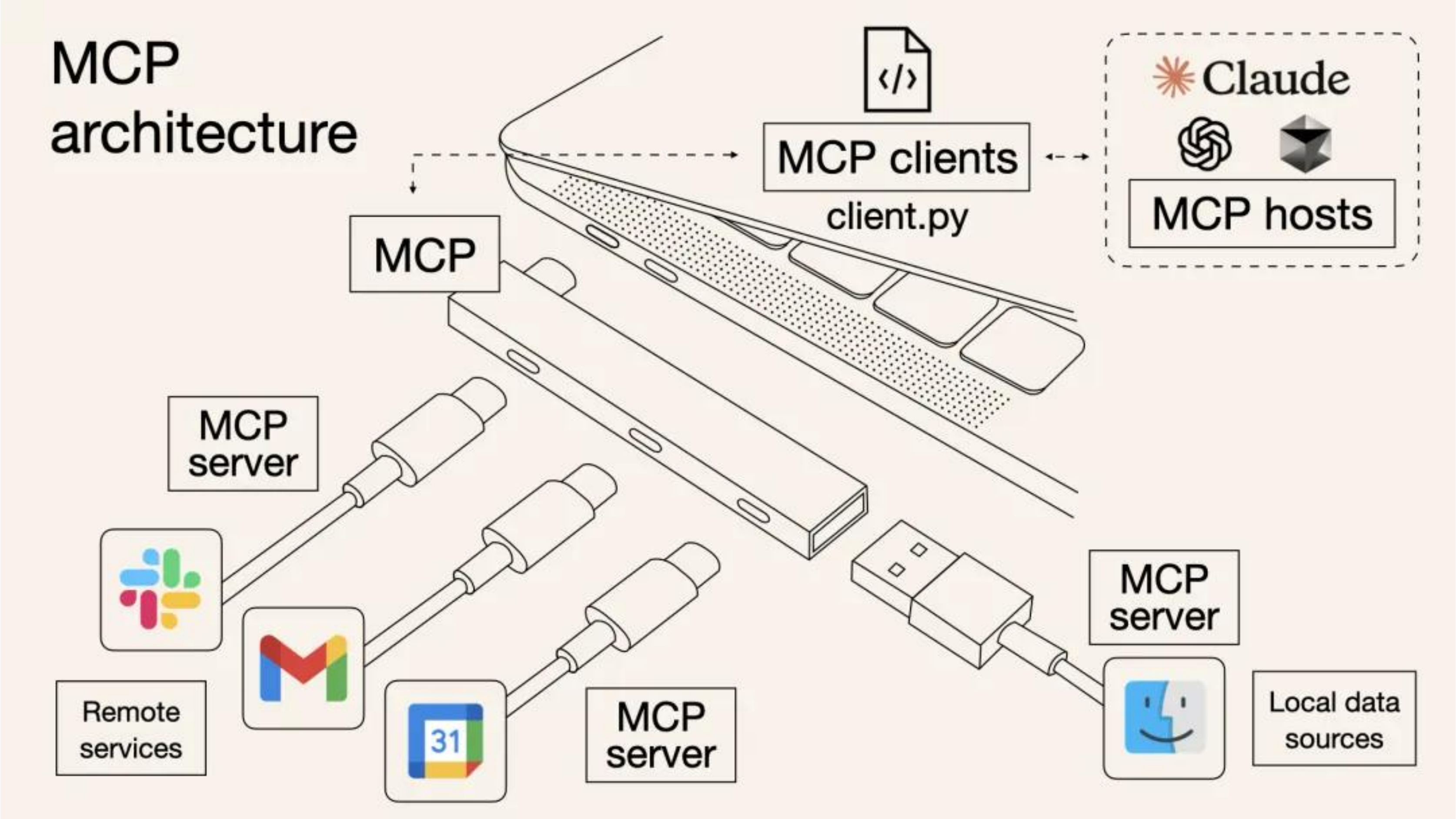
- Browsers and computers
  - Use browsers and computers through API or GUI



From 2411.18279

Key Components of AI Agent: Tool

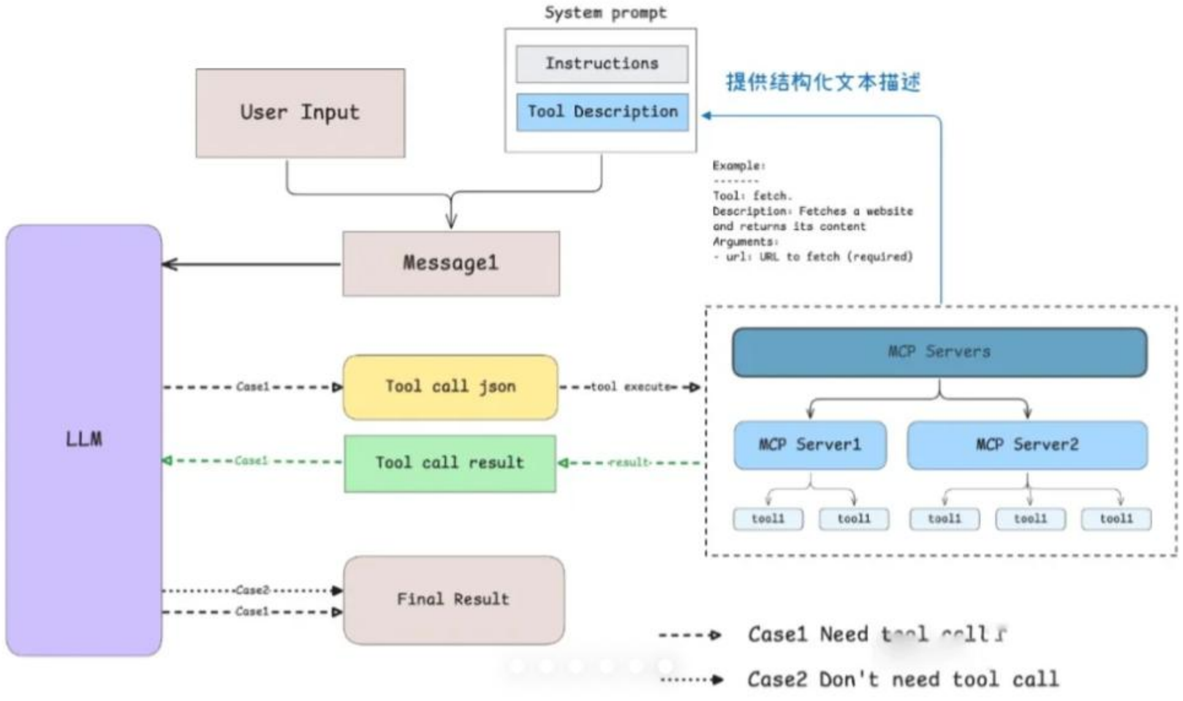
- Model Context Protocol (MCP)



Key Components of AI Agent: Tool

- Model Context Protocol (MCP)

1. 客户端（Claude Desktop / Cursor）将你的问题发送给 Claude。
2. Claude 分析可用的工具，并决定使用哪一个（或多个）。
3. 客户端通过 MCP Server 执行所选的工具。
4. 工具的执行结果被送回给 Claude。
5. Claude 结合执行结果构造最终的 prompt 并生成自然语言的回应。
6. 回应最终展示给用户！



Key Components of AI Agent: Tool

- Model Context Protocol (MCP)
  - An example: LLM call Ableton to create music
    - LLM: Claude
    - LLM Client: Claude Desktop
    - MCP server: AbletonMCP Server
      - AbletonMCP server controls Ableton through API

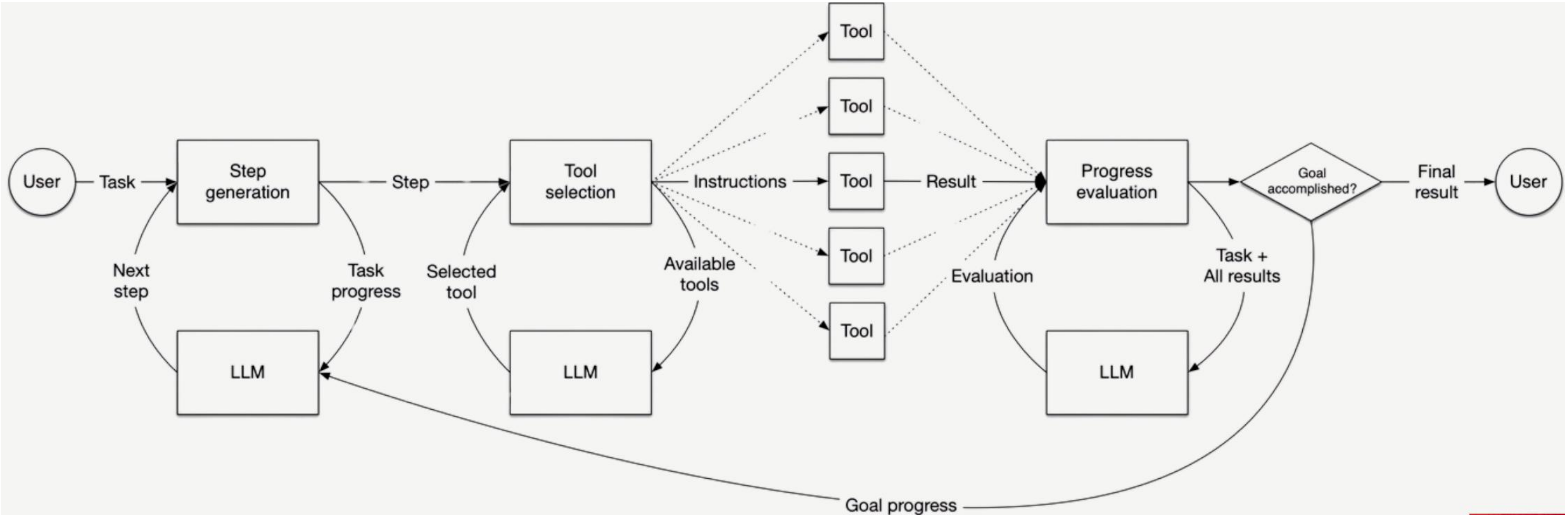
From ahujasid/ableton-mcp

Outline

- Background
  - Agent & AGI
  - What is Agent
- Agent: Foundations
  - Key Components: Reasoning, Memory, Tool Use
  - Agentic Workflow vs Large Agent Model
- Agent: Applications
  - Search/Research Agent
  - Computer-Using Agent
  - Other Vertical Agents
- Challenges & Future Trends

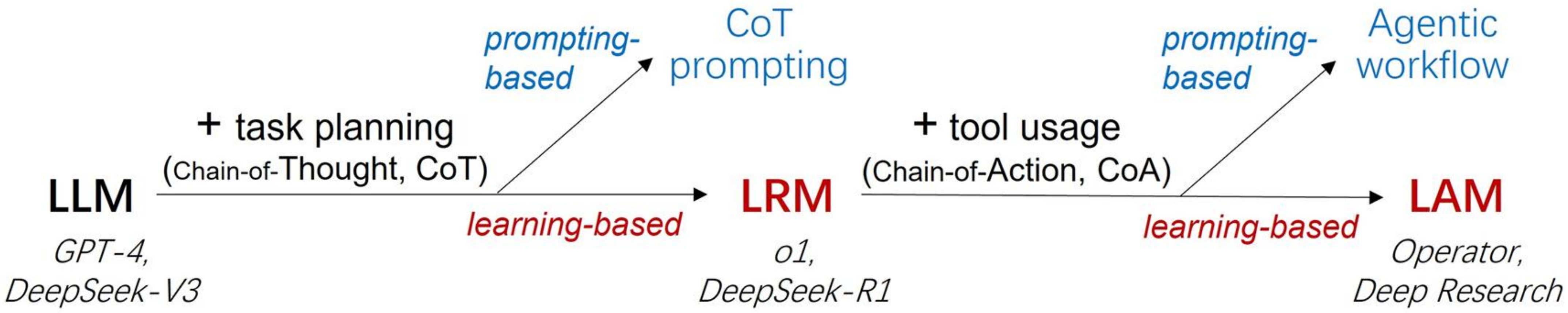
Agentic Workflow vs Large Agent Model

- Agentic workflow
  - Prompting-based, predefined workflow



Agentic Workflow vs Large Agent Model

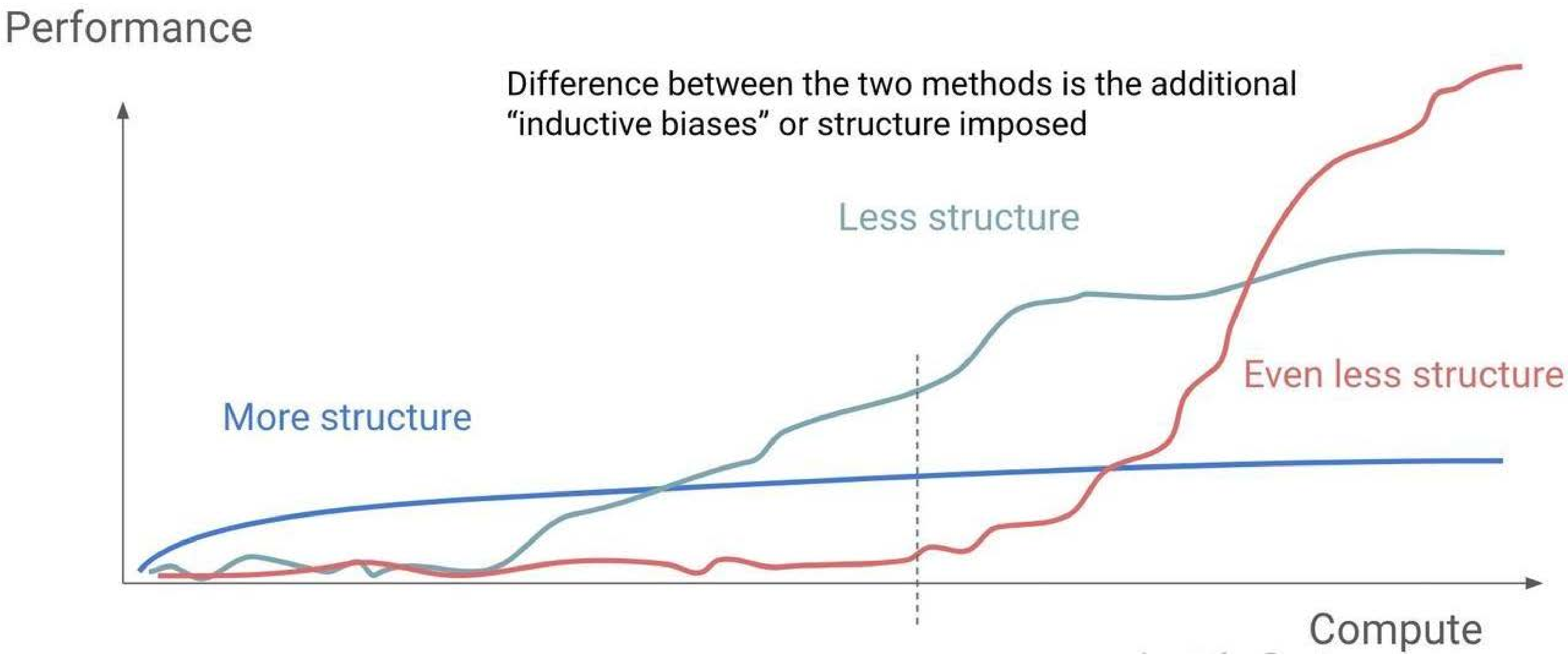
- From prompting to learning



From 2503.06580

Agentic Workflow vs Large Agent Model

- Less structure, more intelligence



Difference between the two methods is the additional "inductive biases" or structure imposed

Agentic Workflow vs Large Agent Model

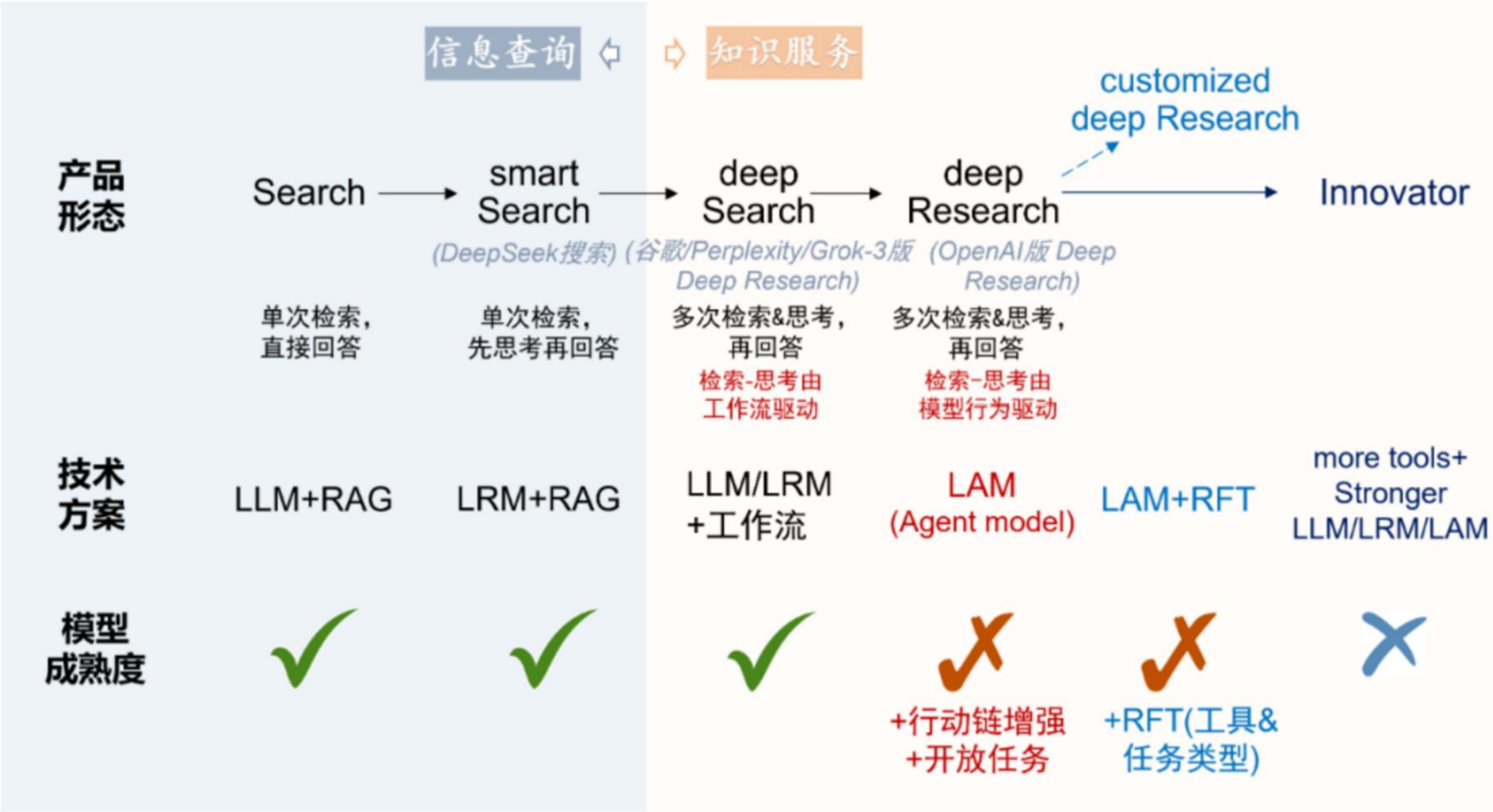
- Large agent model (LAM)
  - Internalize: learn the logics between thinking steps and internalize the CoT generation capability as a “active” model behavior
  - Through SFT and/or RL
    - Data/environment preparation
      - e.g., simulate web search environment
      - <question><think><action><observation>…<think><action><observation><answer>
    - SFT: Warm start the model
    - Prompt and verifier/reward
      - How to speedup rollout process
      - How to get verifier/reward
    - RL: PPO/GRPO
    - Repeat

Outline

- Background
  - Agent & AGI
  - What is Agent
- Agent: Foundations
  - Key Components: Reasoning, Memory, Tool Use
  - Agentic Workflow vs Large Agent Model
- Agent: Applications
  - Search/Research Agent
  - Computer-Using Agent
  - Other Vertical Agents
- Challenges & Future Trends

Agent Application: Search/Research Agent

- The roadmap of search/research agent



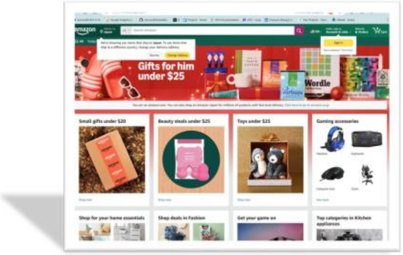
信息查询 ⇦ ⇨ 知识服务

产品形态	Search →	smart Search →	deep Search →	deep Research →	customized deep Research	Innovator
		(DeepSeek搜索)	(谷歌/Perplexity/Grok-3版 Deep Research)	(OpenAI版 Deep Research)		
	单次检索，直接回答	单次检索，先思考再回答	多次检索&思考，再回答 检索-思考由工作流驱动	多次检索&思考，再回答 检索-思考由模型行为驱动		
技术方案	LLM+RAG	LRM+RAG	LLM/LRM +工作流	LAM (Agent model)	LAM+RFT	more tools+ Stronger LLM/LRM/LAM
模型成熟度	✓	✓	✓	✗ +行动链增强 +开放任务	✗ +RFT(工具&任务类型)	✗

Agent Application: Computer-using Agent

- Computer-using Agent
  - Agents using computing devices (e.g., computers and mobile phones) by operating within the environments and interfaces (e.g., Graphical User Interface (GUI), Application Programming Interface (API)) provided by operating systems (OS) to automate tasks

  - Computing devices: Computers, mobile phones, servers
  - OS Agent (windows, mac, android, ios), Web Agent
  - API agent vs GUI Agent

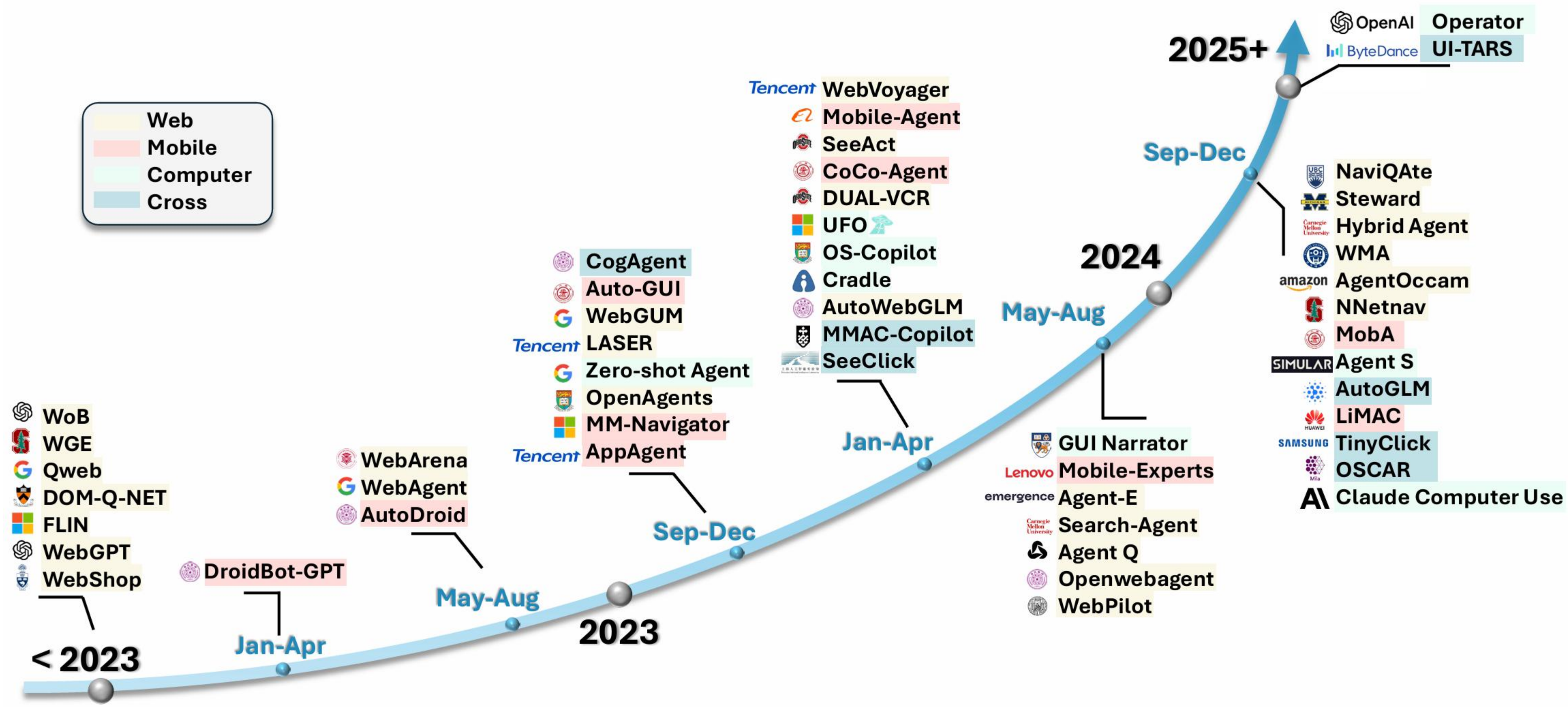


(a) Web GUI

(b) Mobile GUI

(c) Computer GUI

Agent Application: Computer-using Agent



Web
Mobile
Computer
Cross

< 2023
- WoB
- WGE
- Qweb
- DOM-Q-NET
- FLIN
- WebGPT
- WebShop

Jan-Apr
- DroidBot-GPT

May-Aug
- WebArena
- WebAgent
- AutoDroid

2023

Sep-Dec
- CogAgent
- Auto-GUI
- WebGUM
- LASER
- Zero-shot Agent
- OpenAgents
- MM-Navigator
- AppAgent

Jan-Apr
- WebVoyager
- Mobile-Agent
- SeeAct
- CoCo-Agent
- DUAL-VCR
- UFO
- OS-Copilot
- Cradle
- AutoWebGLM
- MMAC-Copilot
- SeeClick

2024

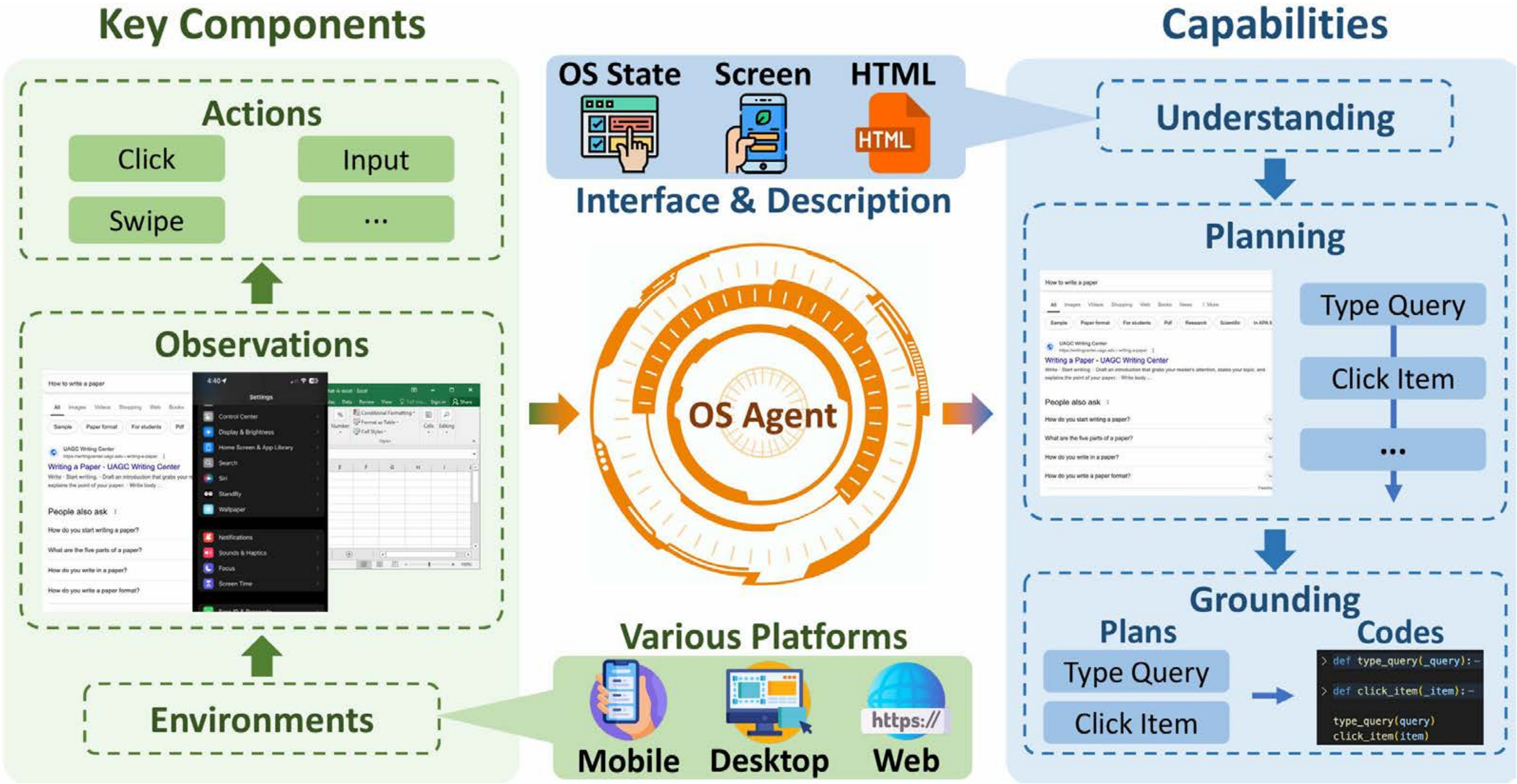
May-Aug
- GUI Narrator
- Mobile-Experts
- Agent-E
- Search-Agent
- Agent Q
- Openwebagent
- WebPilot

Sep-Dec
- NaviQAte
- Steward
- Hybrid Agent
- WMA
- AgentOccam
- NNetnav
- MobA
- Agent S
- AutoGLM
- LiMAC
- TinyClick
- OSCAR
- Claude Computer Use

2025+
- OpenAI Operator
- ByteDance UI-TARS

From 2411.18279

Agent Application: Computer-using Agent



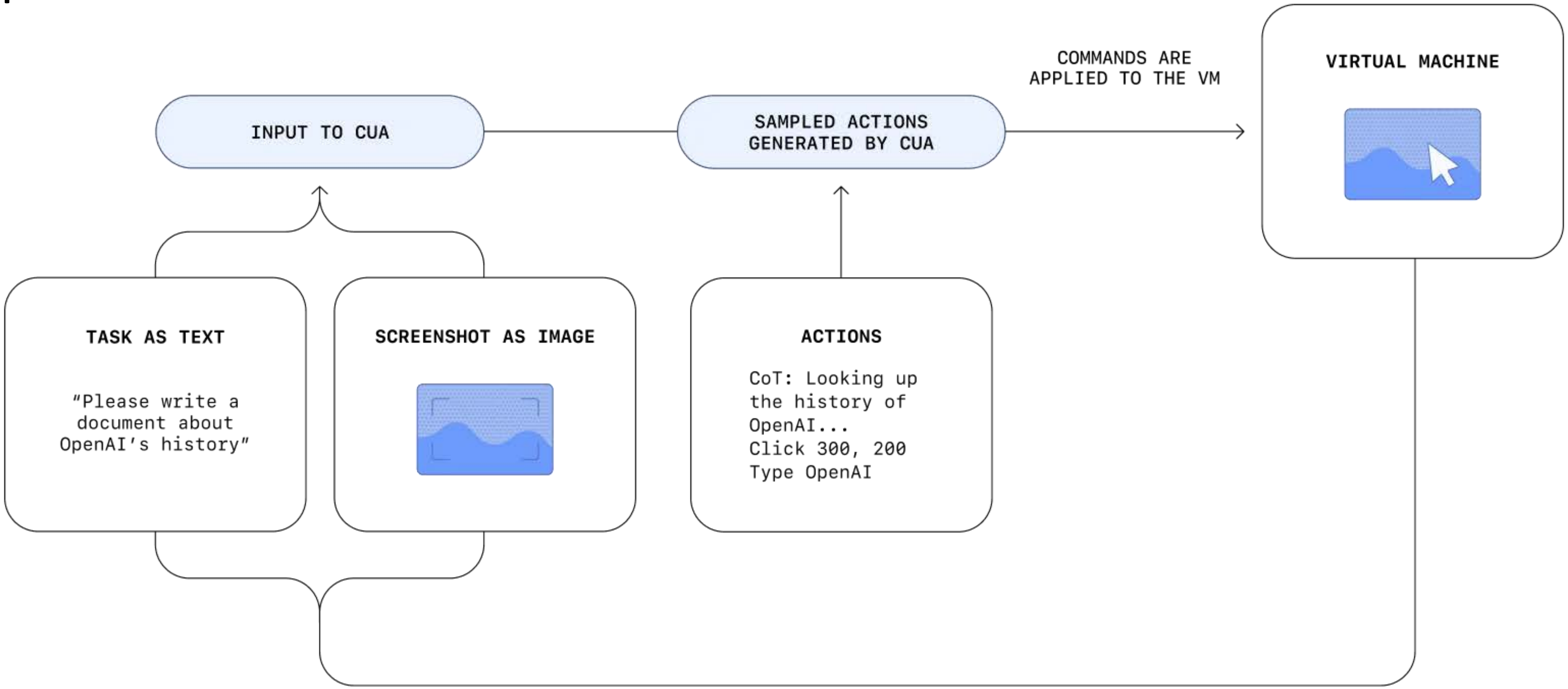
```
> def type_query(_query): ...
> def click_item(_item): ...

type_query(query)
click_item(item)
```

From OS Agents

Agent Application: Computer-using Agent

- Operator



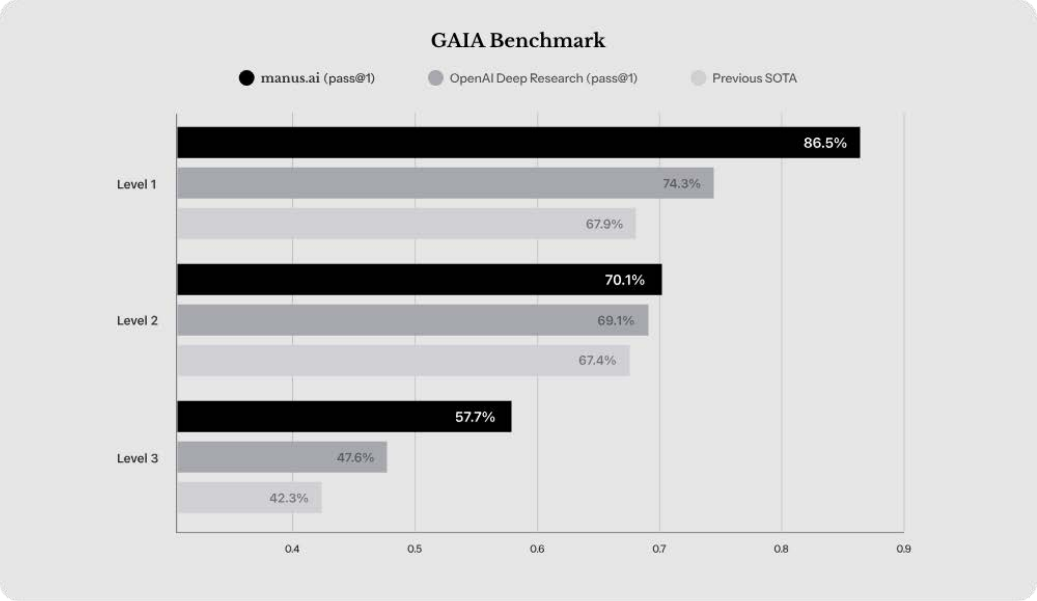
"Please write a document about OpenAI's history"

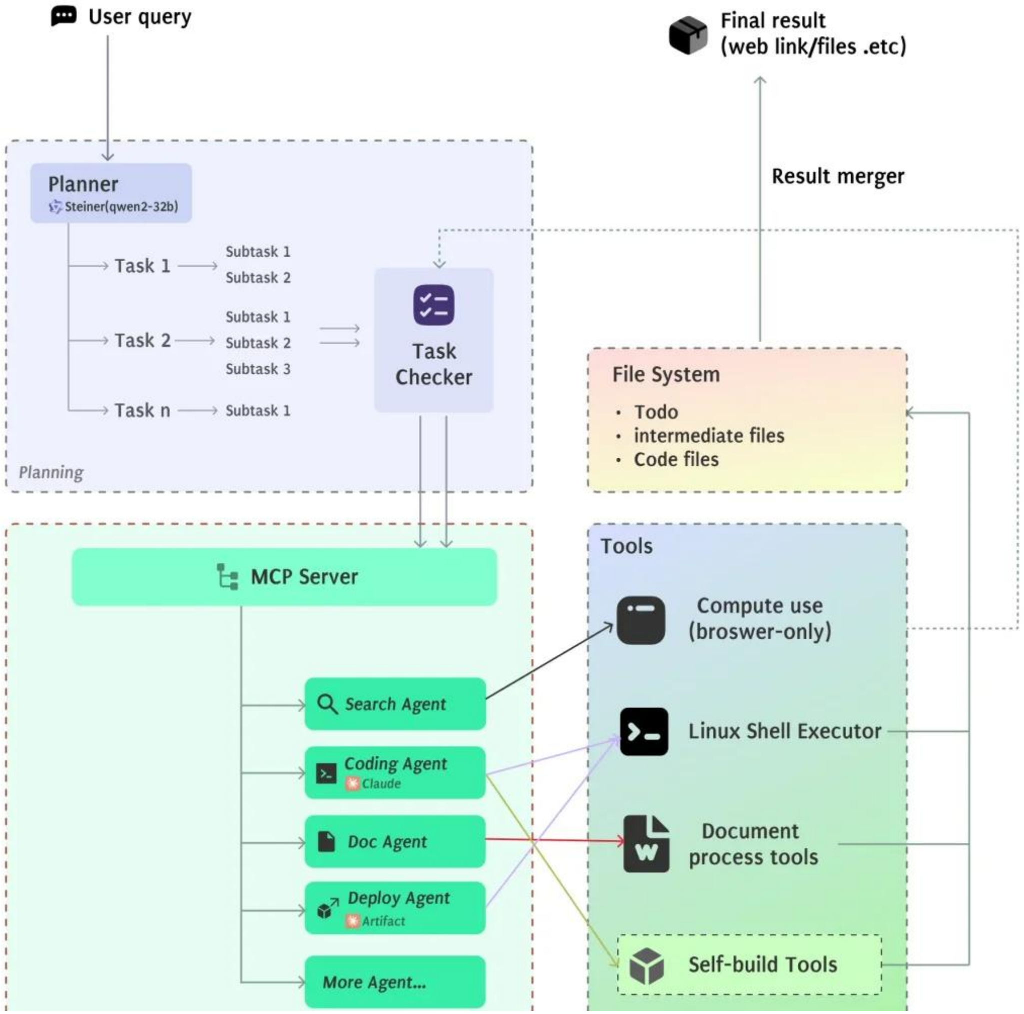
CoT: Looking up the history of OpenAI...
Click 300, 200
Type OpenAI

From Introducing Operator | OpenAI

Agent Application: Computer-using Agent

- Manus
  - Operator + DeepResearch





Agent Application: Vertical

- Coding & Development
  - Devin, Cursor, Replit, Windsurf, Trae
- Customer service and Sales
  - Decagon, Clay
- Business research
  - Hebbia
- Scientific research
  - Elicit
- Supply chain
  - Palantir
- Healthcare
  - Epic

Agent Evaluation

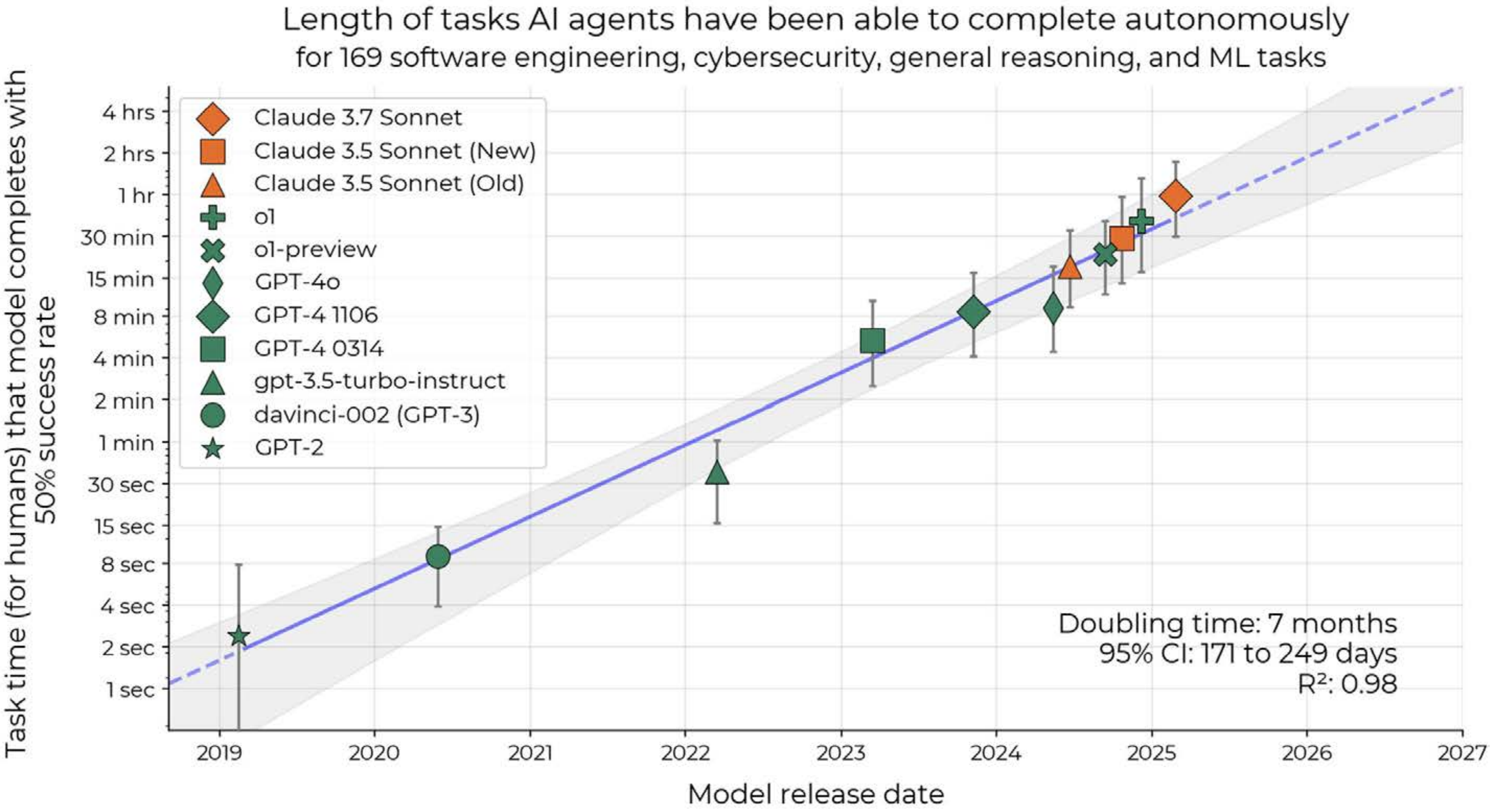
- Benchmark
  - OSWorld
  - WebArena
  - WebVoyoger
  - GAIA
  - Humanity’s last exam

Outline

- Background
  - Agent & AGI
  - What is Agent
- Agent: Foundations
  - Key Components: Reasoning, Memory, Tool Use
  - Agentic Workflow vs Large Agent Model
- Agent: Applications
  - Search/Research Agent
  - Computer-Using Agent
  - Other Vertical Agents
- Challenges & Future Trends

Agent Moore Law

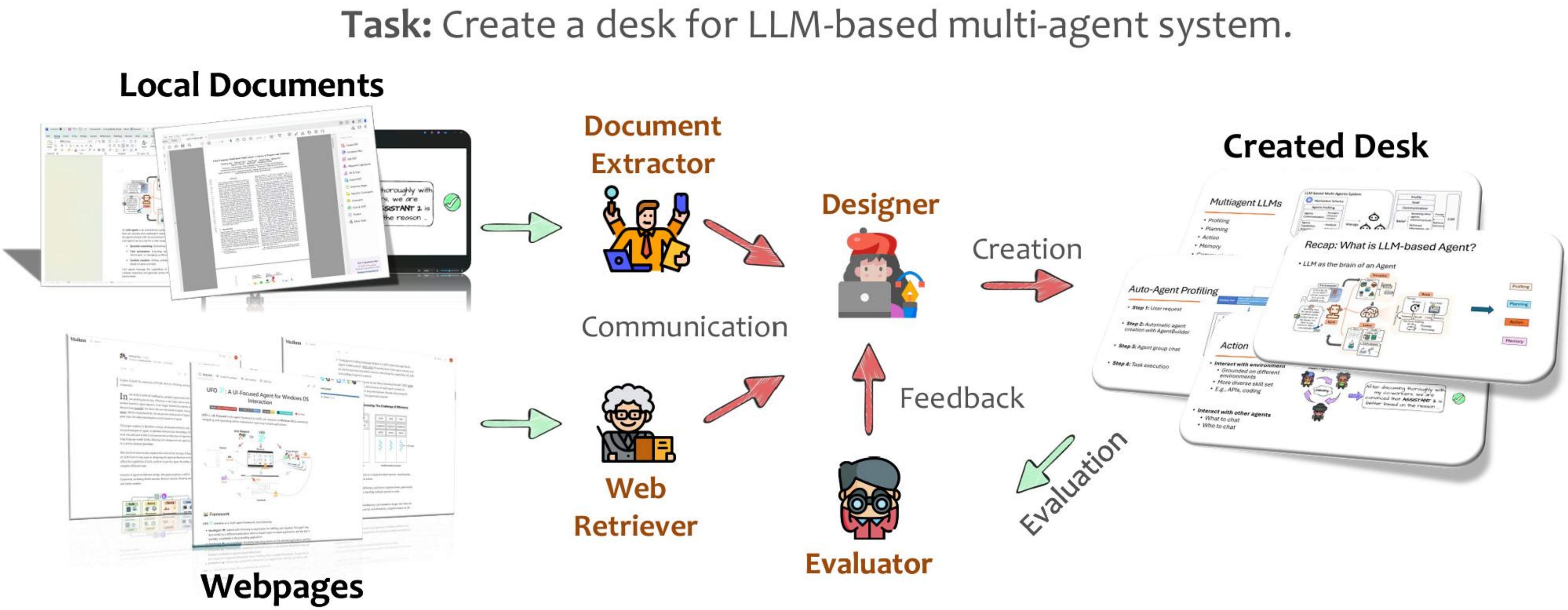
- Agentic time double every 7 months



From 2503.14499

Multi-Agent System

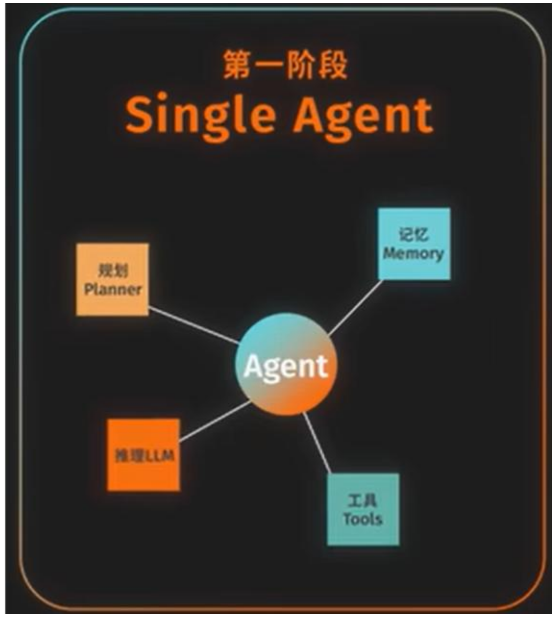
- An example

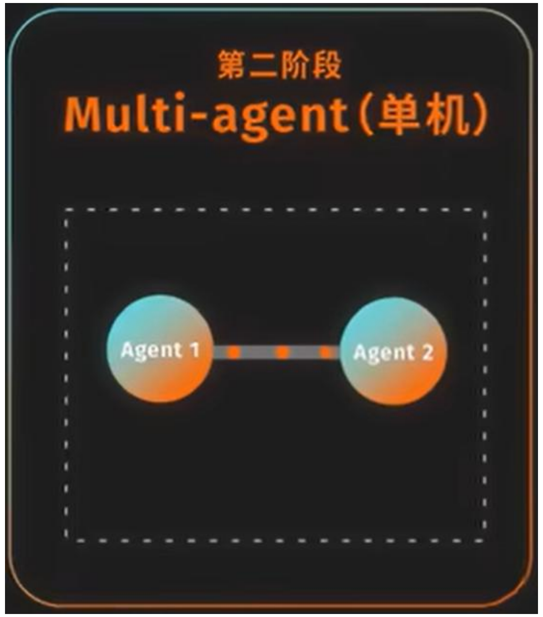


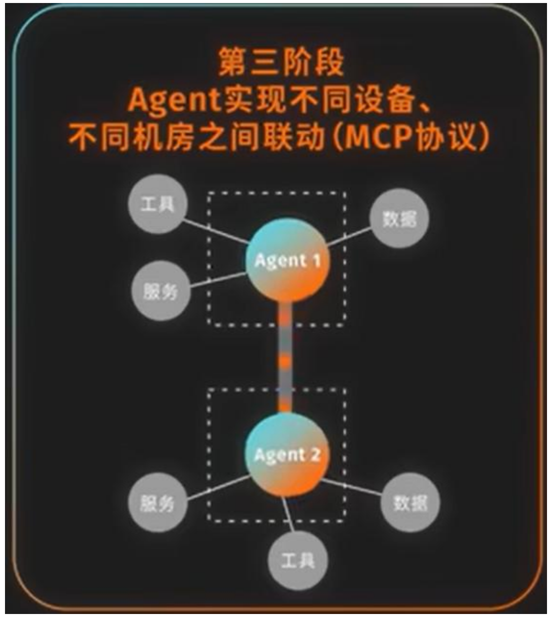
From 2411.18279

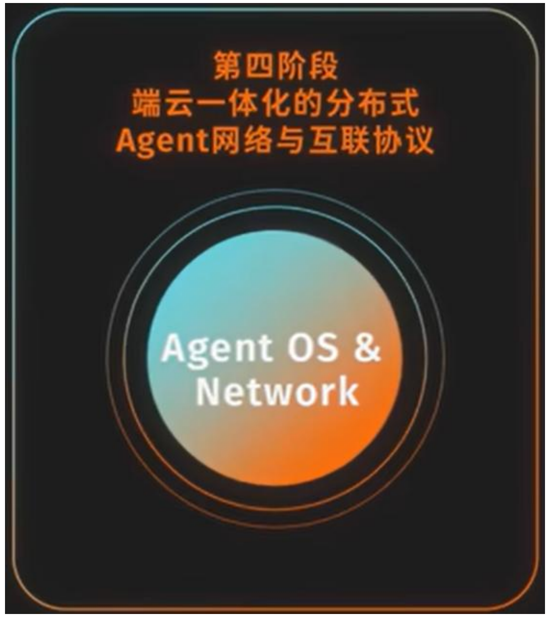
Multi-Agent System

- Communication in multi-agent system



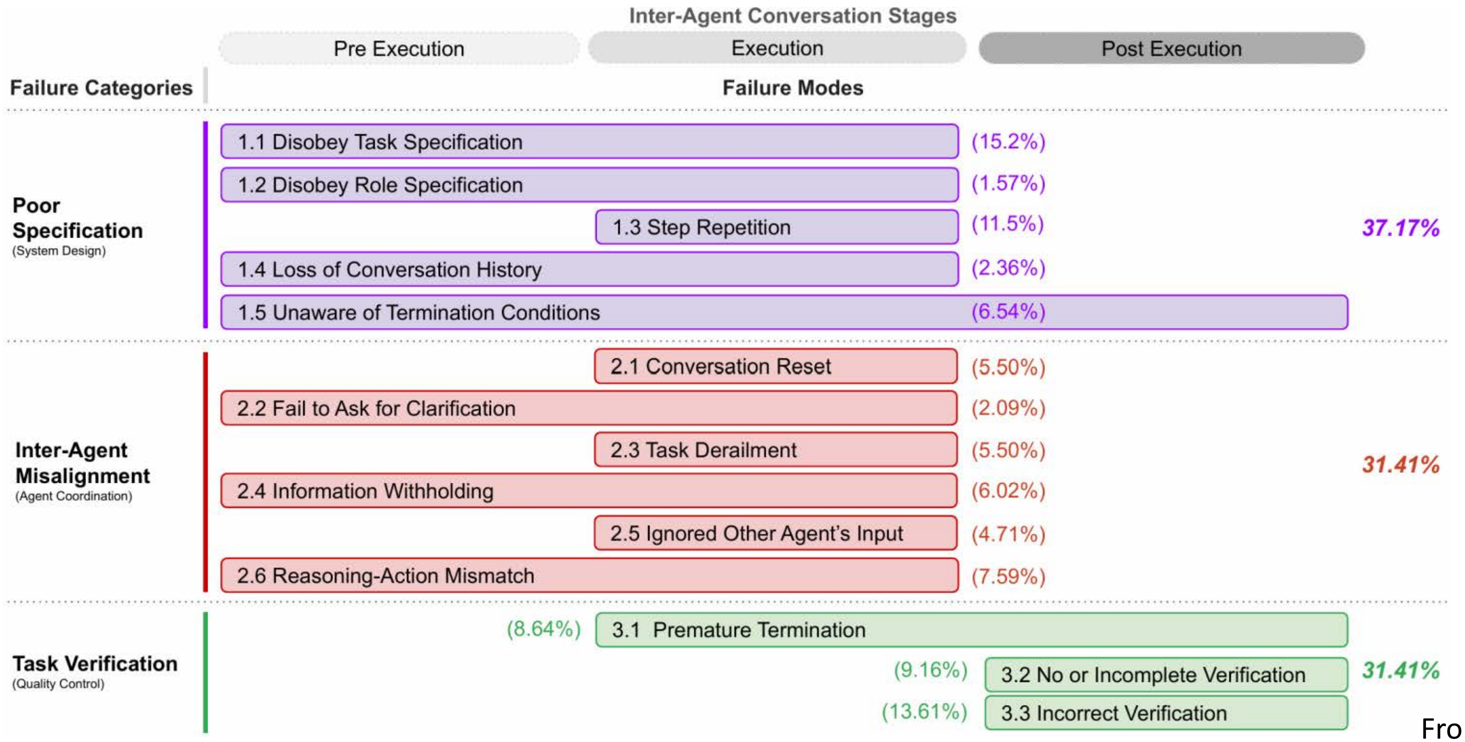






Multi-Agent System

- Why multi-agent system fails?
  - Planning: task decomposition, roll assignment
  - Coordination
  - Verification



Inter-Agent Conversation Stages

Pre Execution | Execution | Post Execution

Failure Categories | Failure Modes

Poor Specification (System Design) — 37.17%

- 1.1 Disobey Task Specification (15.2%)
- 1.2 Disobey Role Specification (1.57%)
- 1.3 Step Repetition (11.5%)
- 1.4 Loss of Conversation History (2.36%)
- 1.5 Unaware of Termination Conditions (6.54%)

Inter-Agent Misalignment (Agent Coordination) — 31.41%

- 2.1 Conversation Reset (5.50%)
- 2.2 Fail to Ask for Clarification (2.09%)
- 2.3 Task Derailment (5.50%)
- 2.4 Information Withholding (6.02%)
- 2.5 Ignored Other Agent's Input (4.71%)
- 2.6 Reasoning-Action Mismatch (7.59%)

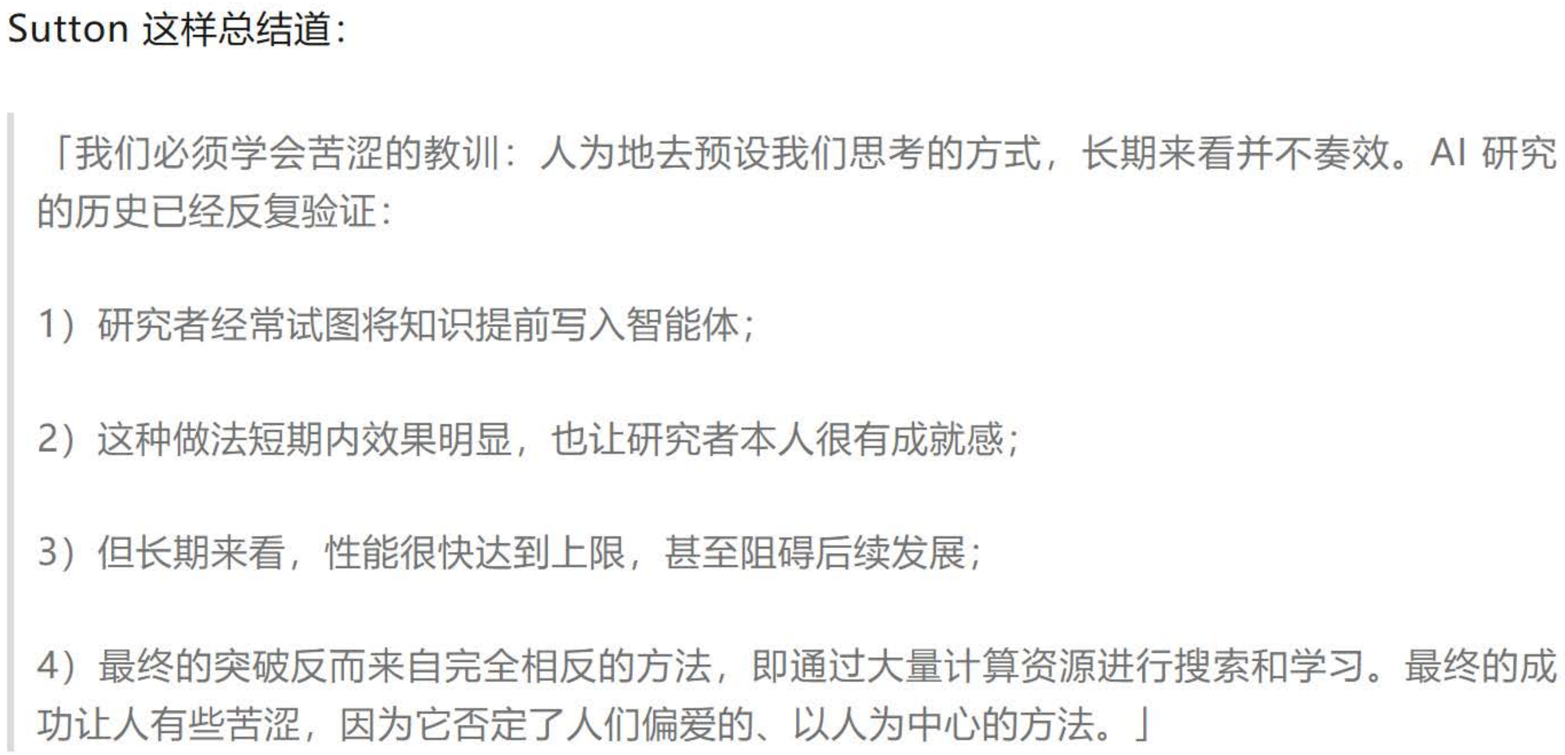
Task Verification (Quality Control) — 31.41%

- (8.64%) 3.1 Premature Termination
- (9.16%) 3.2 No or Incomplete Verification
- (13.61%) 3.3 Incorrect Verification

From Why Do Multi-Agent LLM Systems Fail?

The Bitter Lesson for Agent

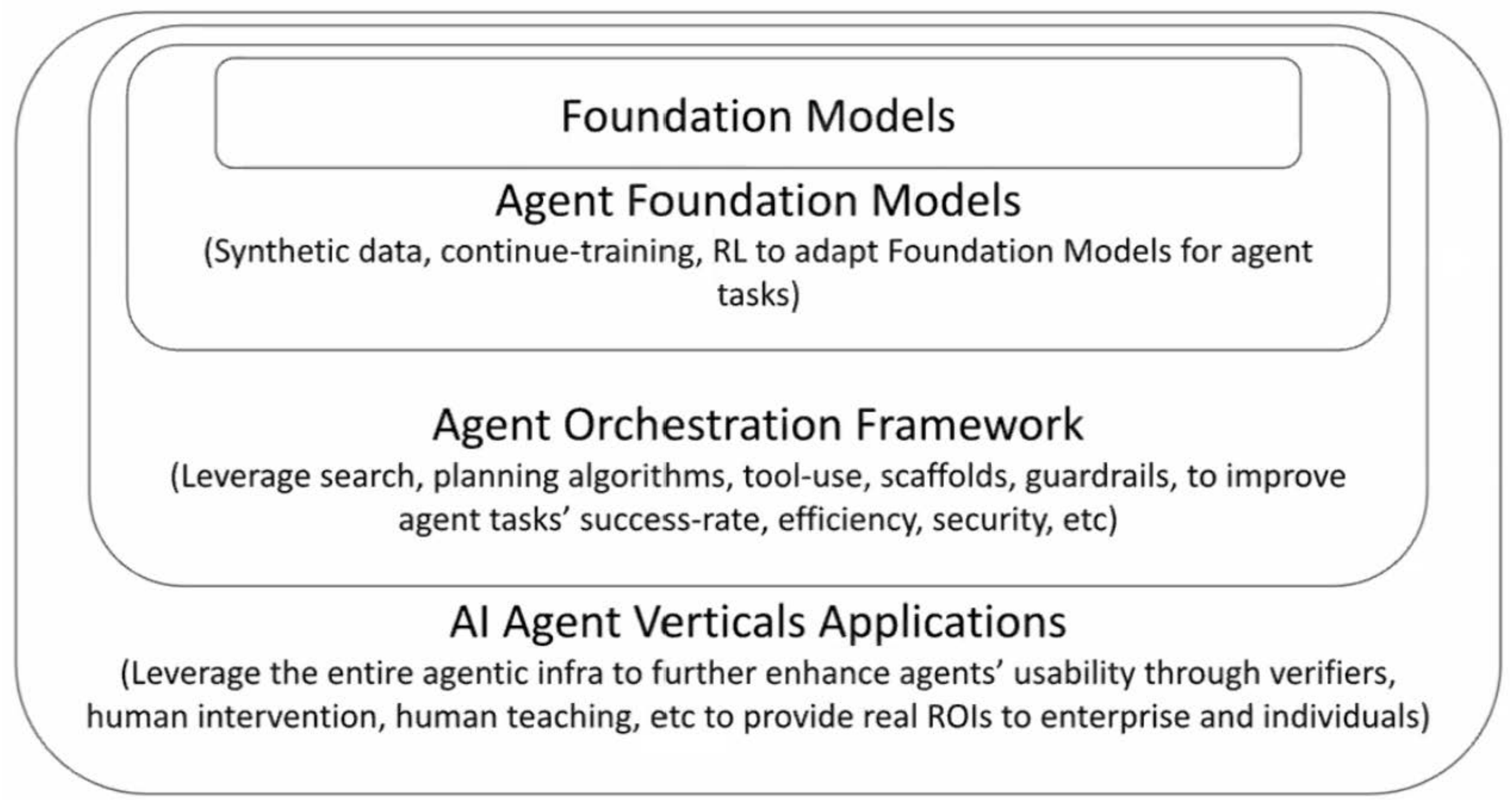
- Less structure, more intelligence



Sutton 这样总结道：

> 「我们必须学会苦涩的教训：人为地去预设我们思考的方式，长期来看并不奏效。AI 研究的历史已经反复验证：
>
> 1）研究者经常试图将知识提前写入智能体；
>
> 2）这种做法短期内效果明显，也让研究者本人很有成就感；
>
> 3）但长期来看，性能很快达到上限，甚至阻碍后续发展；
>
> 4）最终的突破反而来自完全相反的方法，即通过大量计算资源进行搜索和学习。最终的成功让人有些苦涩，因为它否定了人们偏爱的、以人为中心的方法。」

Current Agent Research Ecosystem



Foundation Models

Agent Foundation Models

(Synthetic data, continue-training, RL to adapt Foundation Models for agent tasks)

Agent Orchestration Framework

(Leverage search, planning algorithms, tool-use, scaffolds, guardrails, to improve agent tasks' success-rate, efficiency, security, etc)

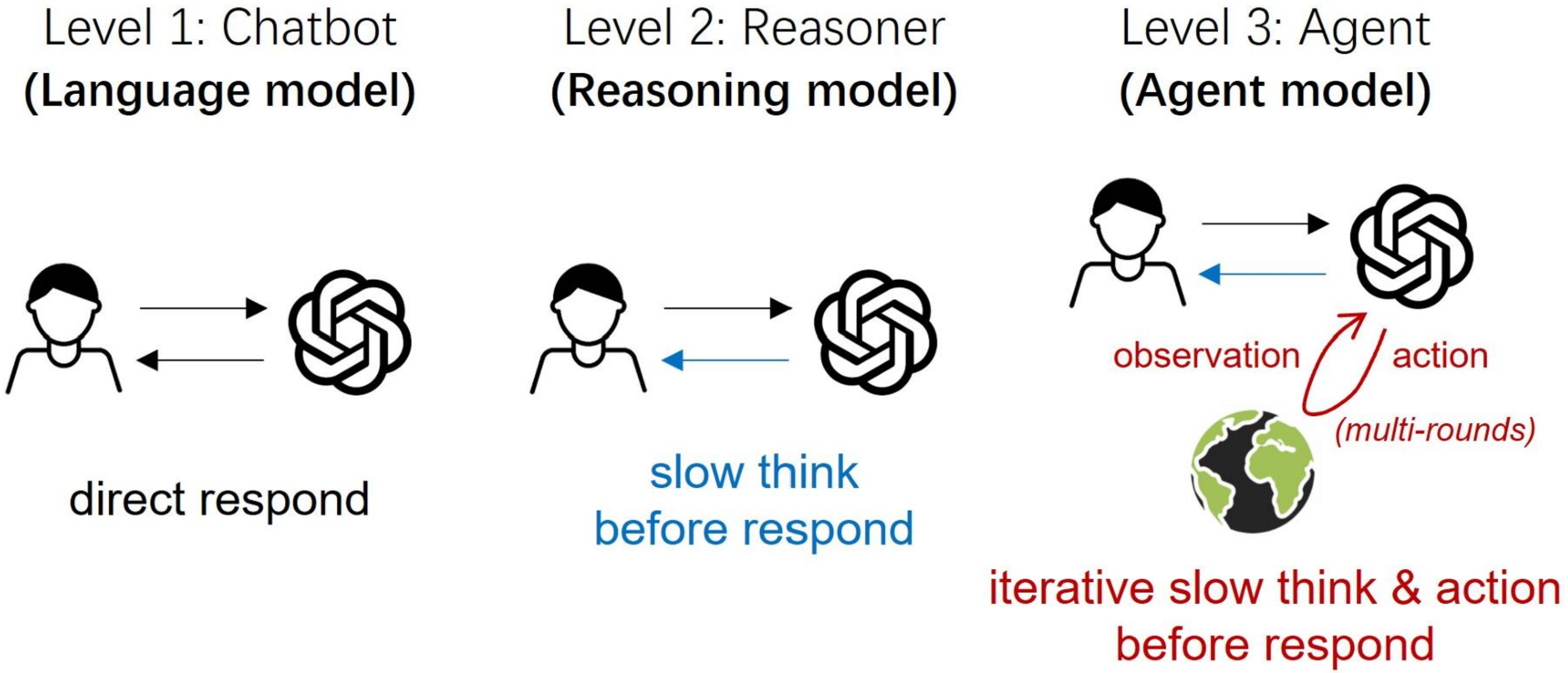
AI Agent Verticals Applications

(Leverage the entire agentic infra to further enhance agents' usability through verifiers, human intervention, human teaching, etc to provide real ROIs to enterprise and individuals)

Internalize Reasoning/Action into LLMs

- How to internalize
  - Environment
  - Data
  - Feedback loop
  - Reward model

AGI Roadmap: From Chatbot to Reasoner to Agent



Level 1: Chatbot
(Language model)

direct respond

Level 2: Reasoner
(Reasoning model)

slow think
before respond

Level 3: Agent
(Agent model)

observation

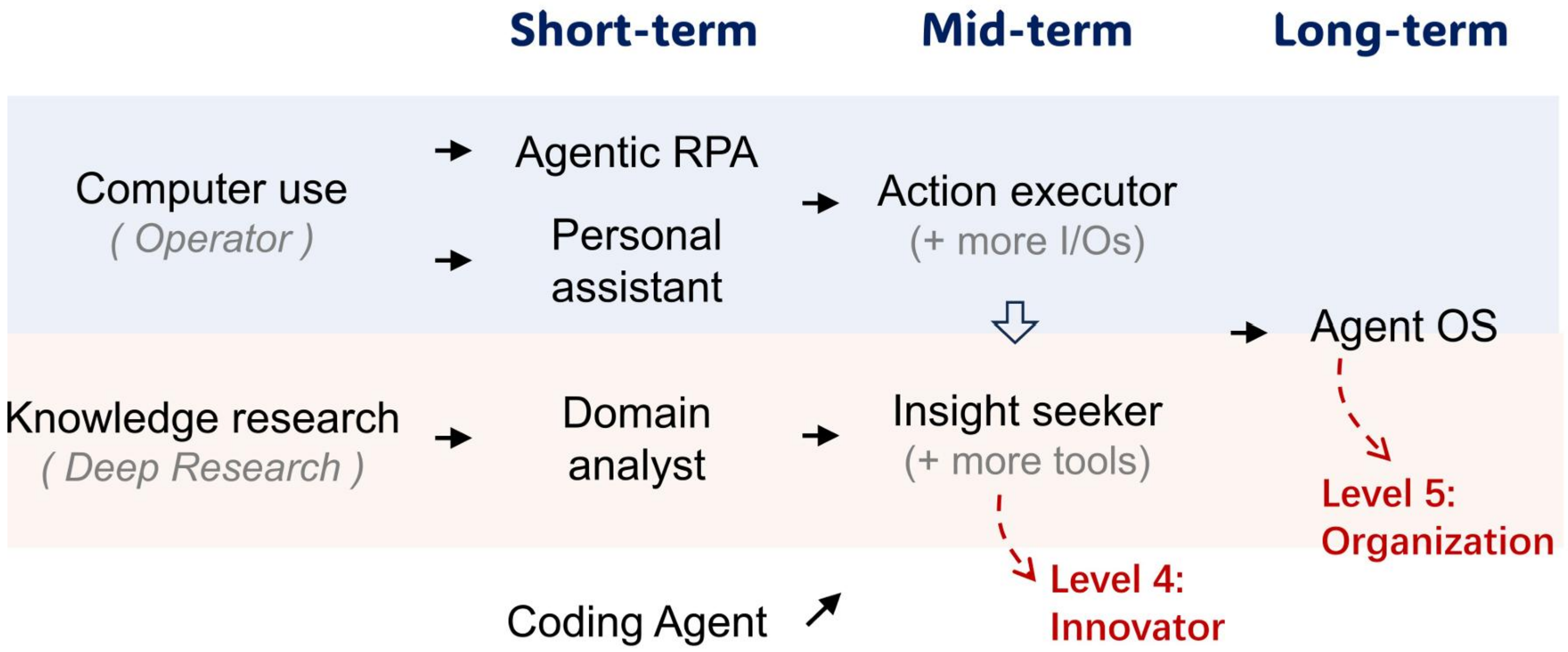
action

(multi-rounds)

iterative slow think & action
before respond

From 2503.06580

AGI Roadmap: From Agent to Innovator to Organization



Short-term

Mid-term

Long-term

Computer use
(Operator)

Agentic RPA

Personal assistant

Action executor
(+ more I/Os)

Agent OS

Knowledge research
(Deep Research)

Domain analyst

Insight seeker
(+ more tools)

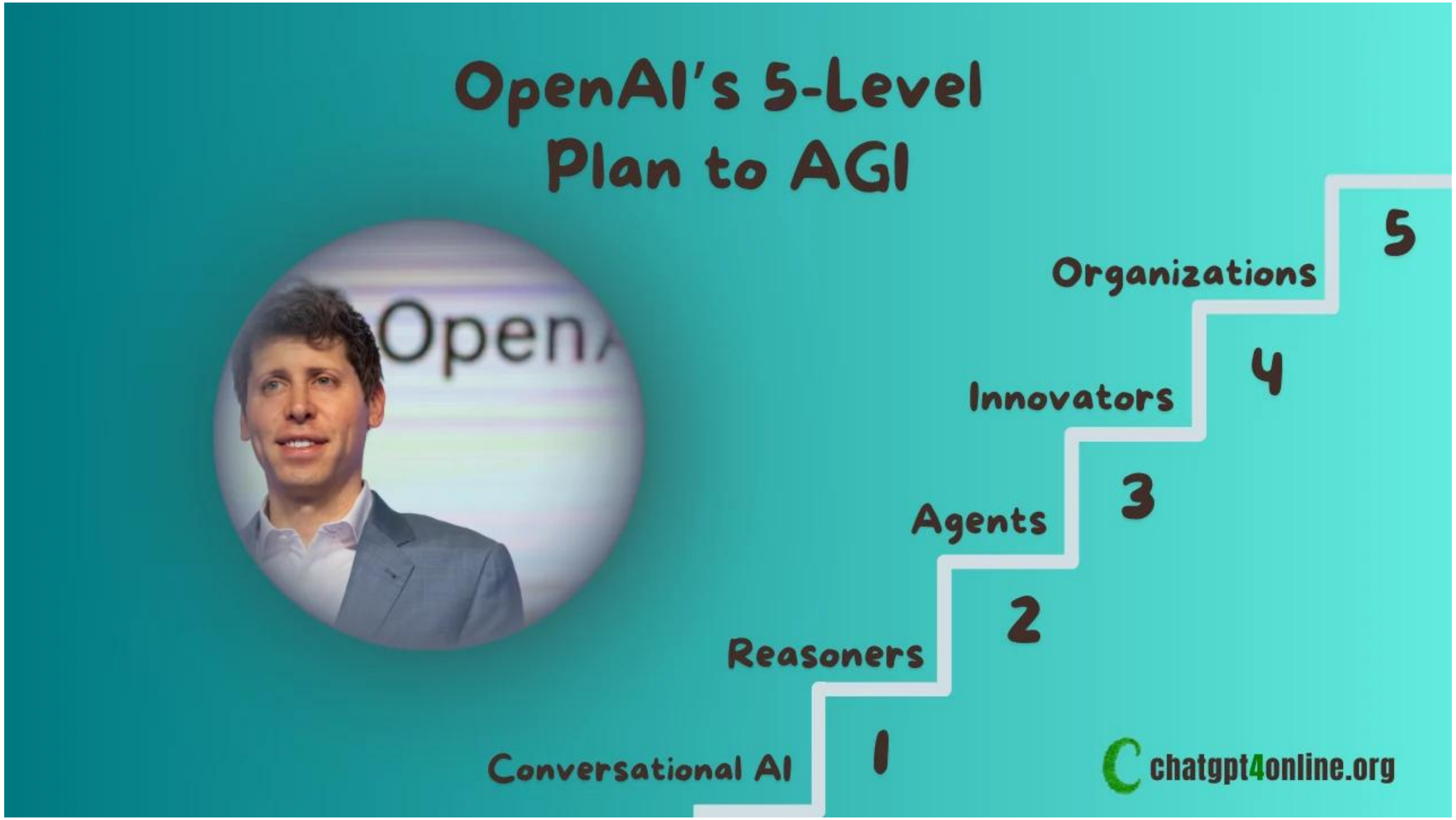
Coding Agent

Level 4: Innovator

Level 5: Organization

From 2503.06580

AGI Roadmap


OpenAI's 5-Level Plan to AGI
OpenAI
5
Organizations
4
Innovators
3
Agents
2
Reasoners
1
Conversational AI
chatgpt4online.org

Thanks

Xu Tan/谭旭
tanxu2012@gmail.com