

---

# **CS 294/194-196: Large Language Model Agents**

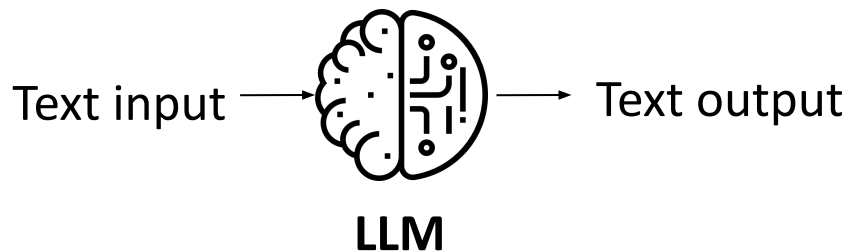
# Teaching Staff

---

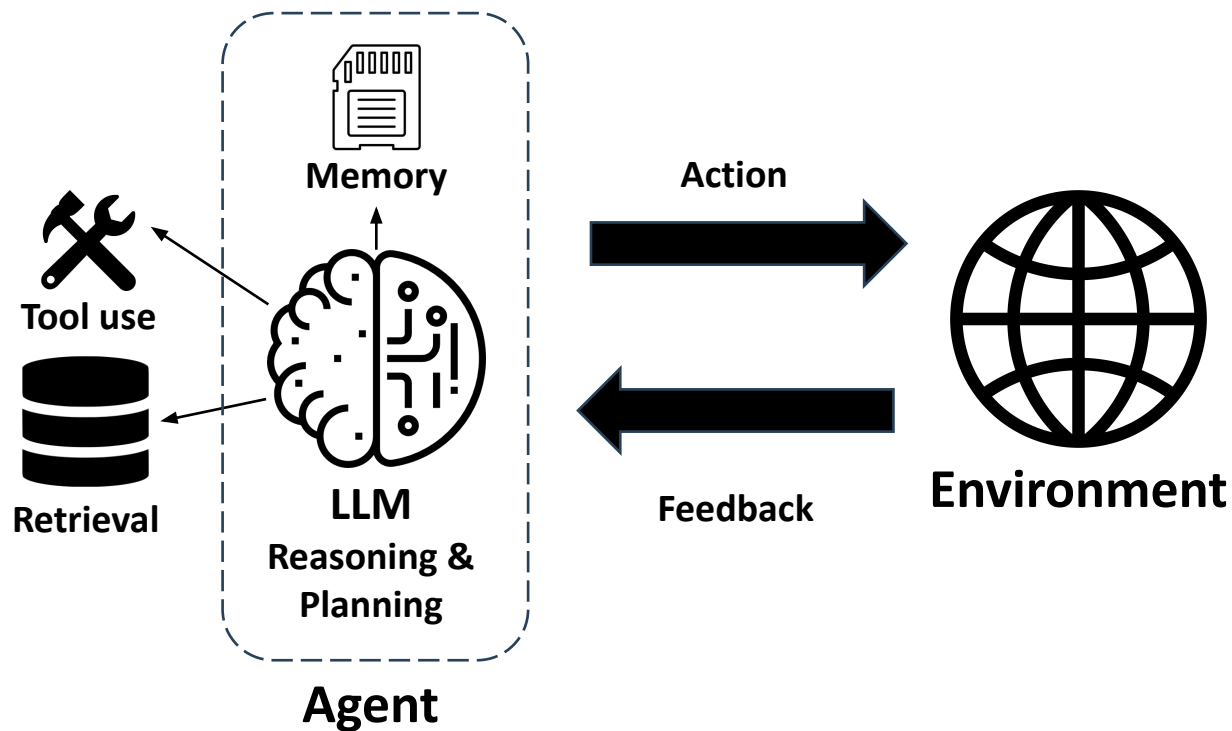
- **Instructor: Prof. Dawn Song**
- **(guest) Co-instructor: Dr. Xinyun Chen**
- **GSIs: Alex Pan & Sehoon Kim**
- **Readers: Tara Pande & Ashwin Dara**

---

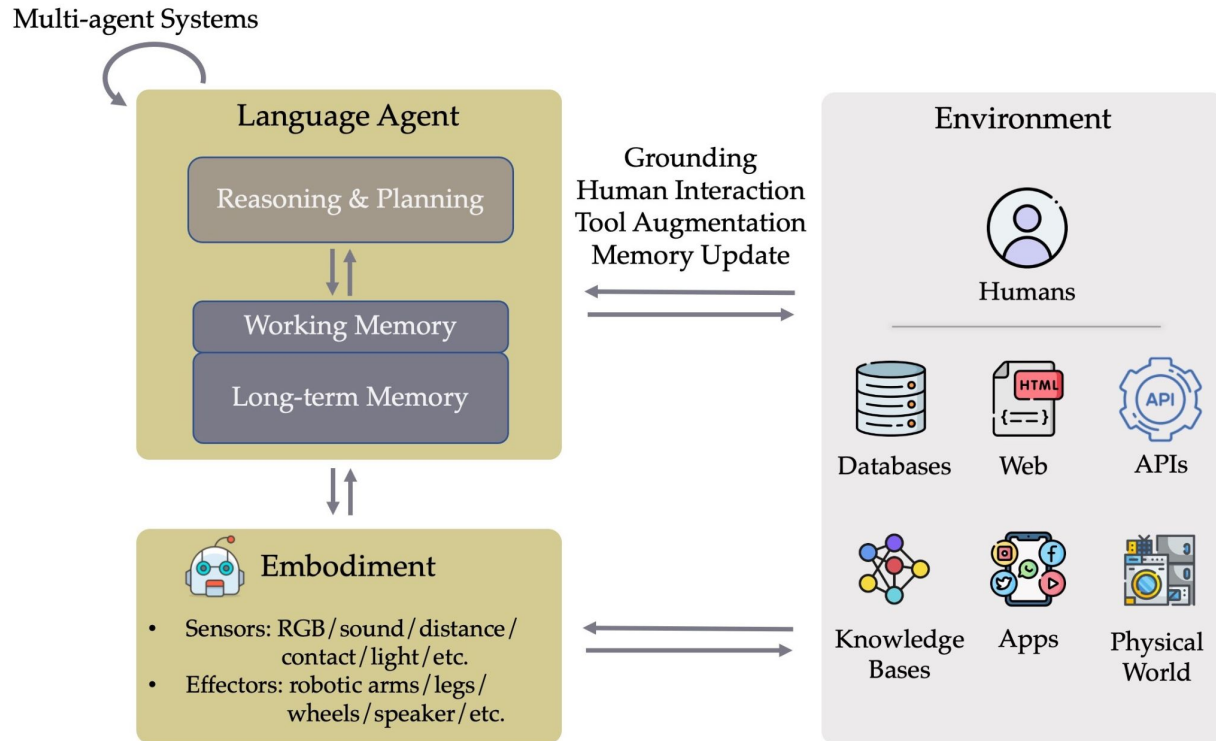
● Amazon-owned ● Anthropic ● Apple ● Chinese ● Google ● Meta / Facebook ● Microsoft ● OpenAI ● Other



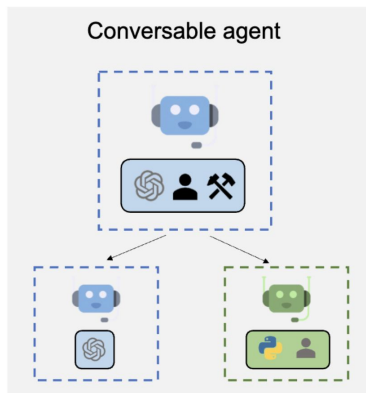
# LLM agents: enabling LLMs to interact with the environment



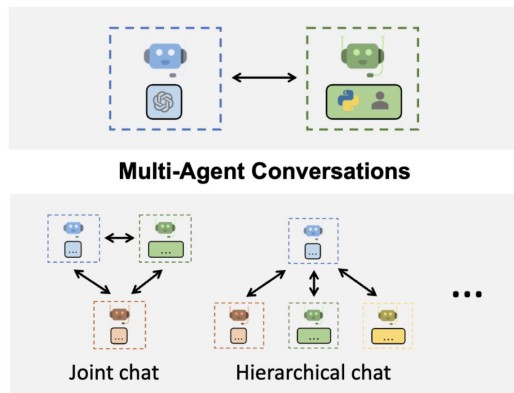
# LLM Agents in Diverse Environments



# Multi-agent collaboration: division of labor for complex tasks



Agent Customization



Flexible Conversation Patterns

## Specialized agents for different subtasks

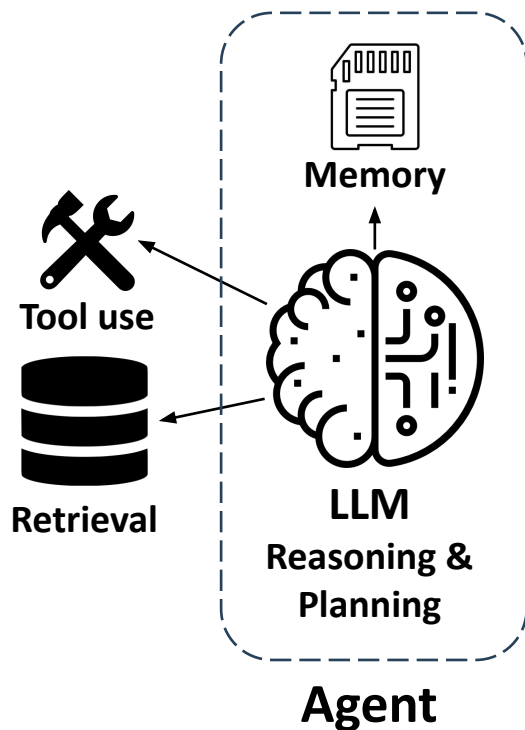
Autogen, CrewAI, CAMEL, Mixture-of-Agents,...



## Emergence of social behaviors with role-play LLMs

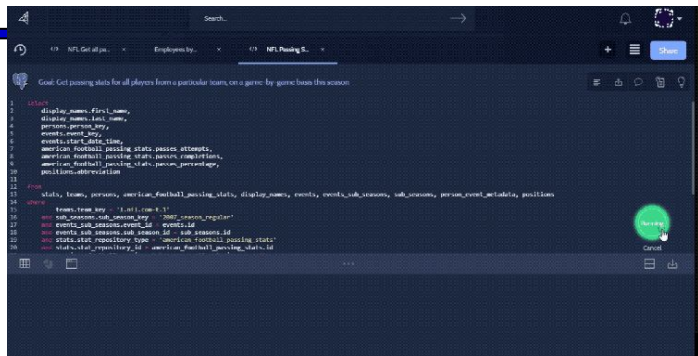
Generative agents, Project Sid,...

# Why empowering LLMs with the agent framework



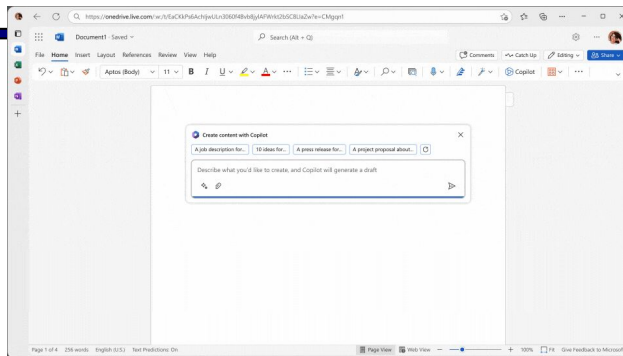
- Solving real-world tasks typically involves a trial-and-error process
- Leveraging external tools and retrieving from external knowledge expand LLM's capabilities
- Agent workflow facilitates complex tasks
  - Task decomposition
  - Allocation of subtasks to specialized modules
  - Division of labor for project collaboration
  - Multi-agent generation inspires better responses

# LLM agents transformed various applications



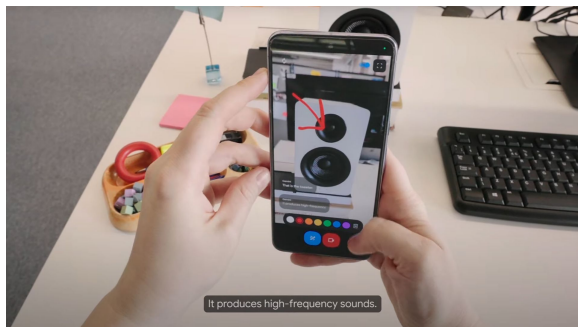
## Code generation

Cursor, GitHub Copilot, Devin, Replit,...



## Workflow automation

Microsoft Copilot, Multi-On,...



## Personal assistant

Google Astra, OpenAI GPT-4o,...



## Robotics

Figure AI, Tesla Optimus,...

- Education
- Law
- Finance
- Healthcare
- Cybersecurity

...



# LLM agents are improving

## Leaderboard

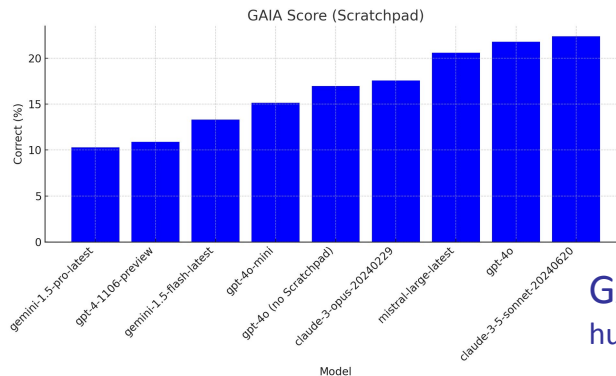
Lite	Verified	Full			
Model	% Resolved	Date	Logs	Trajs	Site
🦋 Gru(2024-08-24)	45.20	2024-08-24	🔗	🔗	🔗
🦋 Honeycomb	40.60	2024-08-20	🔗	🔗	🔗
🦋 Amazon Q Developer Agent (v20240719-dev)	38.80	2024-07-21	🔗	🔗	🔗
AutoCodeRover (v20240620) + GPT 4o (2024-05-13)	38.40	2024-06-28	🔗	-	🔗
Factory Code Droid	37.00	2024-06-17	🔗	-	🔗
🦋✅ SWE-agent + Claude 3.5 Sonnet	33.60	2024-06-20	🔗	🔗	-
🦋✅ AppMap Navie + GPT 4o (2024-05-13)	26.20	2024-06-15	🔗	-	🔗
Amazon Q Developer Agent (v20240430-dev)	25.60	2024-05-09	🔗	-	🔗
EPAM AI/Run Developer Agent + GPT4o	24.00	2024-08-20	🔗	🔗	🔗
🦋✅ SWE-agent + GPT 4o (2024-05-13)	23.20	2024-07-28	🔗	🔗	🔗
🦋✅ SWE-agent + GPT 4 (1106)	22.40	2024-04-02	🔗	🔗	🔗
🦋✅ SWE-agent + Claude 3 Opus	18.20	2024-04-02	🔗	🔗	-
🦋✅ RAG + Claude 3 Opus	7.00	2024-04-02	🔗	-	🔗
🦋✅ RAG + Claude 2	4.40	2023-10-10	🔗	-	-
🦋✅ RAG + GPT 4 (1106)	2.80	2024-04-02	🔗	-	-
🦋✅ RAG + SWE-Llama 7B	1.40	2023-10-10	🔗	-	-
🦋✅ RAG + SWE-Llama 13B	1.20	2023-10-10	🔗	-	-
🦋✅ RAG + ChatGPT 3.5	0.40	2023-10-10	🔗	-	-

SWE-bench **Lite** is a subset of SWE-bench that's been curated to make evaluation less costly and more accessible [Post].

SWE-bench **Verified** is a human annotator filtered subset that has been deemed to have a ceiling of 100% resolution rate [Post].

- The **% Resolved** metric refers to the percentage of SWE-bench instances (2294 for test, 500 for verified, 300 for lite) that were resolved by the model.
- **✅ Checked** indicates that we, the SWE-bench team, received access to the system and were able to reproduce the patch generations.
- **🔗 Open** refers to submissions that have open-source code. This does not necessarily mean the underlying model is open-source.
- The leaderboard is updated once a week on **Monday**.
- If you would like to submit your model to the leaderboard, please check the [submission](#) page.
- All submissions are Pass@1, do not use [hints](#), [text](#), and are in the unassisted setting.

SWE-Bench (Jimenez\*, Yang\*, et al.)  
swebench.com



GAIA (Mialon et al.)  
[huggingface.co/gaia-benchmark](https://huggingface.co/gaia-benchmark)

X-WebArena-Leaderboard									
Menu									
Comment only									
AI									
Release Date									
1	Release Date	Model Size (billion)	Model	Success Rate (%)	Result Source	Work	Traj	Open?	Note
2	08/2024	Unknown	Jaco AI	57.1	Reported by <a href="#">ezalabai</a>	<a href="https://www.jaco.ai/">https://www.jaco.ai/</a>		X	Note from the developer of the work, see the
3	08/2024	Unknown	WebPlot	37.2	<a href="#">WebPlot</a>				
4	04/2024	Unknown	Step	33.5	<a href="#">Step</a>				
5	04/2024	Unknown	BrowsersGym + GPT-4	23.5	<a href="#">WorkArena</a>	<a href="#">BrowsersGym</a>			
6	04/2024	Unknown	GPT-4 + Auto Eval	20.2	<a href="#">Auto Eval &amp; Refine</a>	<a href="#">Auto Eval &amp; Refine</a>			
7	06/2024	Unknown	GPT-4o + Tree Search	19.2	<a href="#">Tree Search for LLM Agents</a>	<a href="#">Tree Search for LLM Agents</a>			
8	04/2024	7	AutoWebArena	19.2	<a href="#">AutoWebArena</a>				
9	06/2023	Unknown	gpt-4o-0813	14.9	<a href="#">WebArena</a>	<a href="#">AutoWebArena</a>			
10	05/2024	Unknown	gpt-4o-2024-05-13	13.1	<a href="#">WebArena Team</a>	<a href="#">GPT</a>			
11	06/2023	Unknown	gpt-4o-0813	11.7	<a href="#">WebArena</a>	<a href="#">GPT</a>			
12	05/2024	72b	Patel et al + 2024	9.36	<a href="#">Patel et al + 2024</a>	<a href="#">Patel et al + 2024</a>			
13	03/2023	Unknown	gpt-3.5-turbo-16k-0813	8.87	<a href="#">WebArena</a>	<a href="#">GPT</a>			
14	06/2023	72b	Queen-1.5-chat-72b	7.14	<a href="#">Patel et al + 2024</a>	<a href="#">Queen</a>			
15	12/2023	Unknown	Gemini Pro	7.12	<a href="#">WebArena</a>	<a href="#">Gemini Pro</a>			
16	04/2024	70	Llama3-chat-70b	7.02	<a href="#">WebArena Team</a>	<a href="#">Llama3</a>			
17	10/2023	70	Lemur-chat-70b	5.3	<a href="#">Lemur</a>	<a href="#">Lemur</a>			
18	03/2024	7	Agent Plan	4.68	<a href="#">Agent Plan</a>	<a href="#">Agent Plan</a>			
19	06/2023	34	Codellama-instruct-34b	4.06	<a href="#">Lemur</a>	<a href="#">Llama2</a>			
20	10/2023	70	AgentLM-70b	3.81	<a href="#">Agent Tuning</a>	<a href="#">Agent Tuning</a>			
21	04/2024	8	Llama3-chat-8b	3.32	<a href="#">WebArena Team</a>	<a href="#">Llama3</a>			
22	02/2024	7	CodeAct Agent	2.3	<a href="#">WebArena Team</a>	<a href="#">CodeAct</a>			
23	10/2023	13	AgentLM-13b	1.6	<a href="#">Agent Tuning</a>	<a href="#">Agent Tuning</a>			
24	01/2024	6.7	Mistral	1.39	<a href="#">Gemini In-depth look</a>	<a href="#">Mistral</a>			
25	10/2023	7	AgentLM-7b	0.74	<a href="#">Agent Tuning</a>	<a href="#">Agent Tuning</a>			
26	10/2023	7	FireAct	0.25	<a href="#">Agent Plan</a>	<a href="#">FireAct</a>			
27	06/2023	7	Codellama-instruct-7b	0	<a href="#">WebArena Team</a>	<a href="#">CodeLlama</a>			
Comment here or email <a href="mailto:shuyanzhou@cs.cmu.edu">shuyanzhou@cs.cmu.edu</a> to submit your work!									
28	03/2024	-	Human	78.24	<a href="#">WebArena</a>				Selected tasks by templates
29			AutoGuide	43.7	<a href="#">AutoGuide</a>	<a href="#">AutoGuide</a>			Reddit subset

WebArena  
(Zhou et al.)  
webarena.dev

# Challenges for LLM agent deployment in the wild

---

- Reasoning and planning
  - LLM agents tend to make mistakes when performing complex tasks end-to-end
- Embodiment and learning from environment feedback
  - LLM agents are not yet efficient at recovering from mistakes for long-horizon tasks
  - Continuous learning, self-improvement
  - Multimodal understanding, grounding and world models
- Multi-agent learning, theory of mind
- Safety and privacy
  - LLMs are susceptible to adversarial attacks, can emit harmful messages and leak private data
- Human-agent interaction, ethics
  - How to effectively control the LLM agent behavior, and design the interaction mode between humans and LLM agents

# Topics covered in this course

---

- Model core capabilities
  - Reasoning
  - Planning
  - Multimodal understanding
- LLM agent frameworks
  - Workflow design
  - Tool use
  - Retrieval-augmented generation
  - Multi-agent systems
- Applications
  - Software development
  - Workflow automation
  - Multimodal applications
  - Enterprise applications
- Safety and ethics

# Large Language Model Agents MOOC



**Dawn Song**



**Xinyun Chen**



**Denny Zhou**



**Shunyu Yao**



**Chi Wang**



**Jerry Liu**



**Burak Gokturk**



**Omar Khattab**



**Graham Neubig**



**Nicolas Chapados**



**Yuandong Tian**



**Jim Fan**



**Percy Liang**



**Ben Mann**



# Course Work

---

- Weekly Reading Assignment
  - Due midnight PT Sunday before the next Monday's lecture
- 1 hands-on Lab
- Semester-long course project

# Grading

---

lecture attendance & weekly reading assignment

+

- 1 unit: article about the topic of a lecture (at least 2 pages)
- 2 units: lab + project (implementation not required)
- 3 units: lab + project with implementation
- 4 units: lab + project with significant implementation and end-to-end demo

# Grading

	1 unit	2 units	3/4 units
Participation	45%	20%	10%
Reading Summaries & Q/A	10%	4%	2%
Article	45%		
Lab		16%	8%
Project			
<i>Proposal</i>		10%	10%
<i>Milestone 1</i>		10%	10%
<i>Milestone 2</i>		10%	10%
<i>Presentation</i>		15%	15%
<i>Report</i>		15%	15%
<i>Implementation</i>			20%

# Class Project

---

- 5 students per group; can be part of a hackathon (more details later)

## Applications Track

- Build LLM agent applications in novel domains

## Benchmarks Track

- Create and improve benchmarks for LLM agents

## Fundamentals Track

- Enhance core agent capabilities (memory, planning, tool use)

## Safety Track

- Address safety concerns in deployment (misuse, privacy, etc.)

## Decentralized and Multi-agent Track

- Enhance decentralized multi-agent systems



# Timeline

---

	<b>Released</b>	<b>Due</b>
Project group formation	9/9	9/16
Project proposal	9/16	9/30
Lab	9/23	10/7
Project milestone #1	10/8	10/21
Project milestone #2	10/29	11/18
Project final presentation	11/19	12/12
Project final report	11/19	12/12