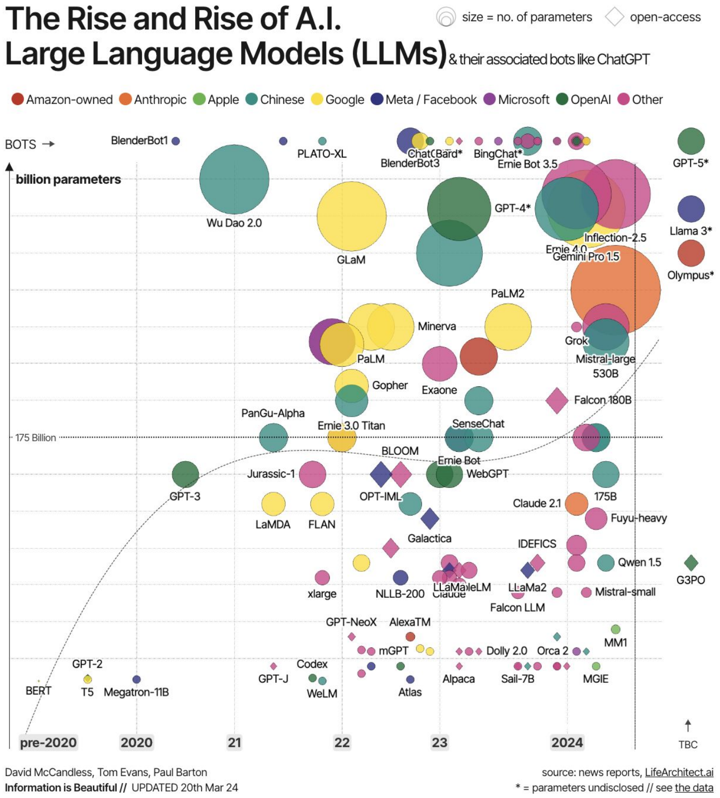
CS 294/194-196:
Large Language Model Agents

Teaching Staff

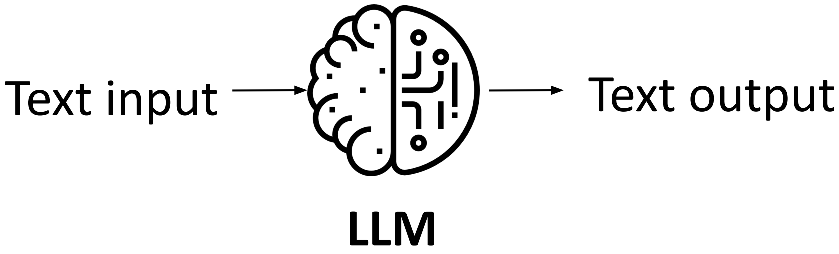
- **Instructor: Prof. Dawn Song**
- **(guest) Co-instructor: Dr. Xinyun Chen**
- GSIs: Alex Pan & Sehoon Kim
- Readers: Tara Pande & Ashwin Dara

Accelerated development of large language models (LLMs)



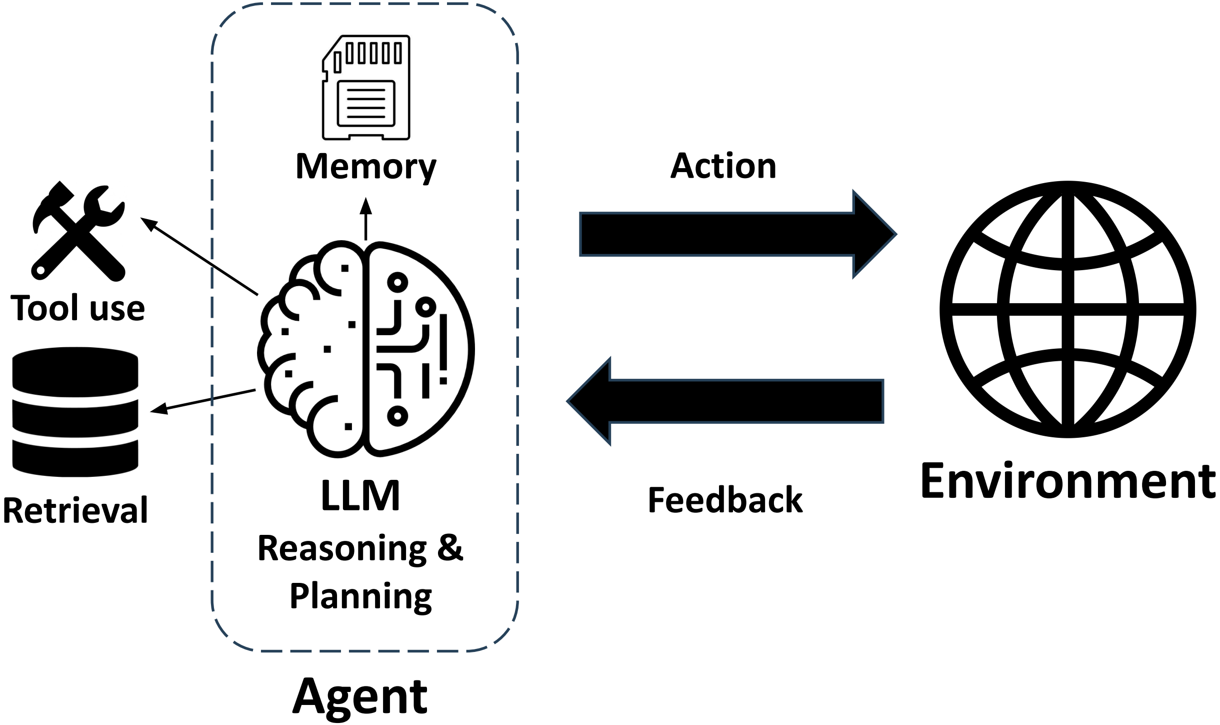
David McCandless, Tom Evans, Paul Barton
Information is Beautiful // UPDATED 20th Mar 24

source: news reports, LifeArchitect.ai
* = parameters undisclosed // see the data

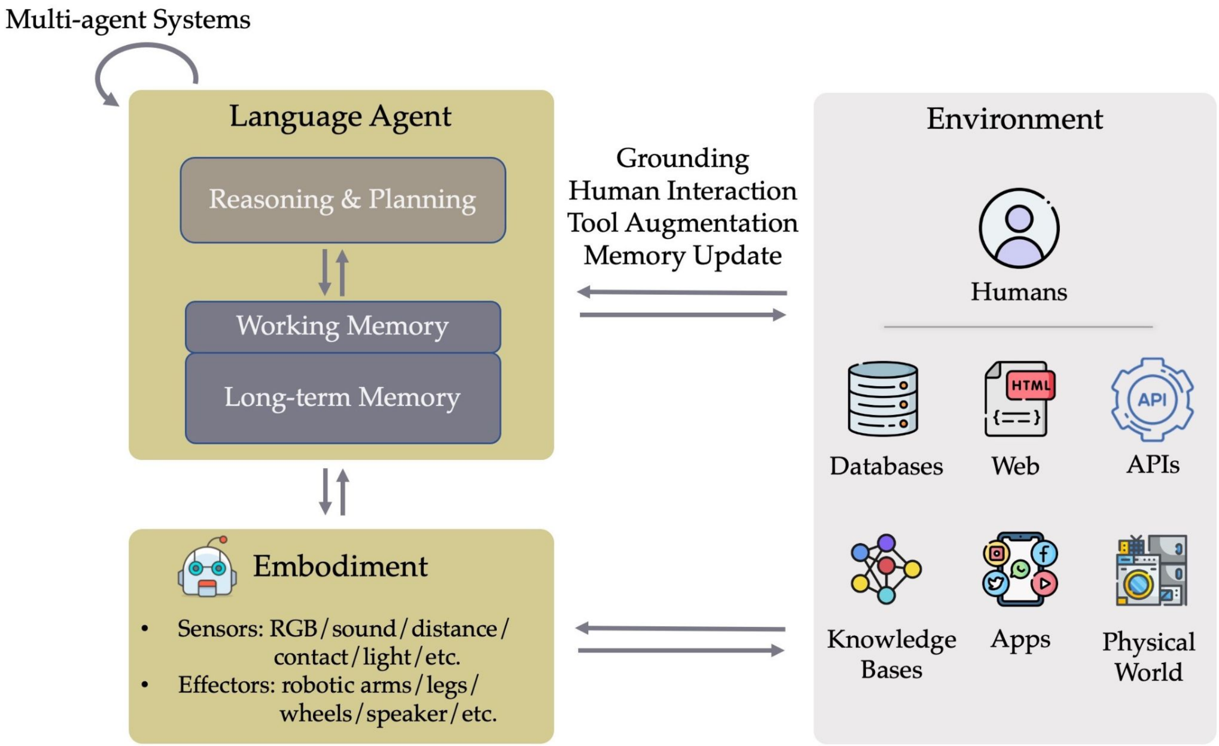


Text input → LLM → Text output

LLM agents: enabling LLMs to interact with the environment



LLM Agents in Diverse Environments



Multi-agent Systems

Language Agent

Reasoning & Planning

Working Memory

Long-term Memory

Embodiment

- Sensors: RGB / sound / distance / contact / light / etc.
- Effectors: robotic arms / legs / wheels / speaker / etc.

Grounding
Human Interaction
Tool Augmentation
Memory Update

Environment

Humans

Databases

Web

APIs

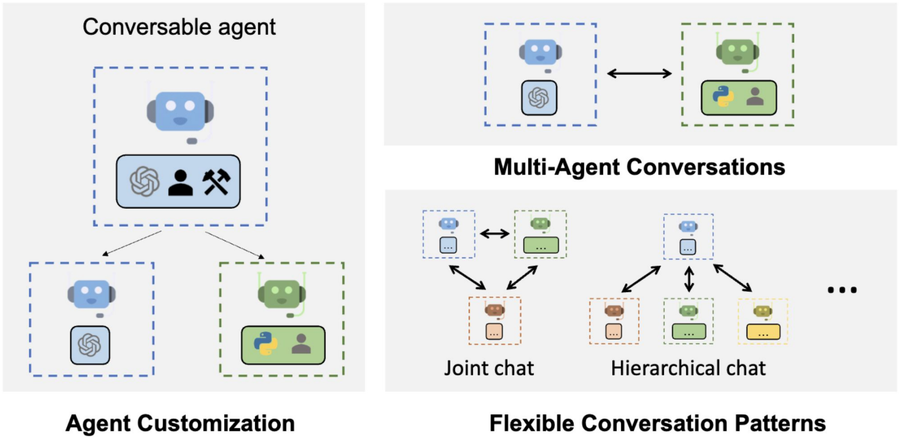
Knowledge Bases

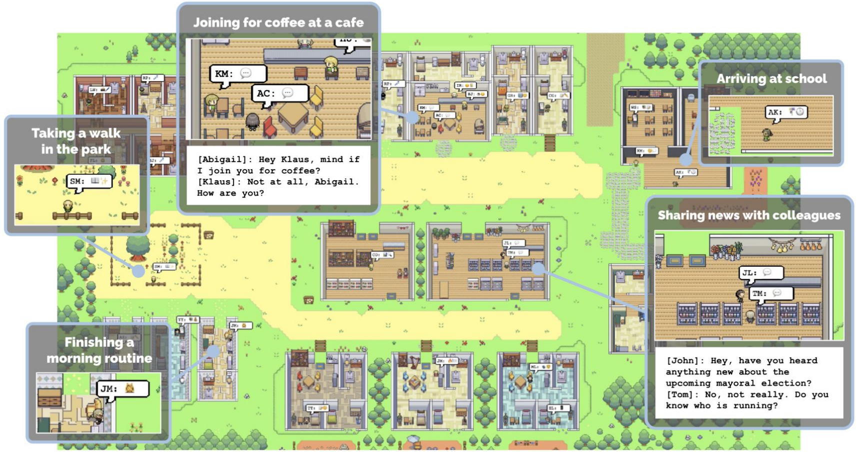
Apps

Physical World

https://yusu.substack.com/p/language-agents

Multi-agent collaboration: division of labor for complex tasks




Joining for coffee at a cafe
KM:
AC:
Taking a walk in the park
SM:
Arriving at school
AK:
Sharing news with colleagues
JL:
TM:
Finishing a morning routine
JM:

```
[Abigail]: Hey Klaus, mind if
I join you for coffee?
[Klaus]: Not at all, Abigail.
How are you?
```

```
[John]: Hey, have you heard
anything new about the
upcoming mayoral election?
[Tom]: No, not really. Do you
know who is running?
```

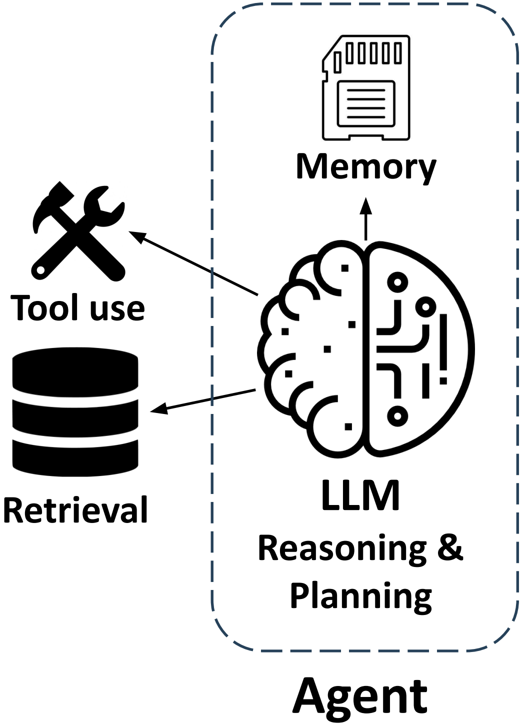
Specialized agents for different subtasks

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

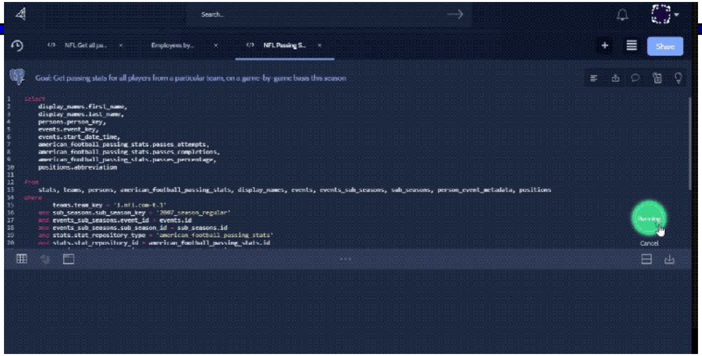


- Solving real-world tasks typically involves a trial-and-error process

- 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



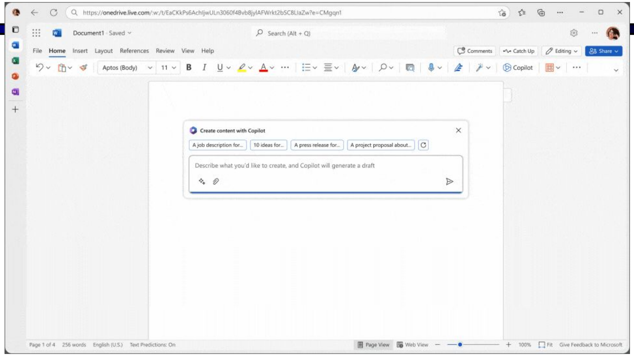
Goal: Get passing stats for all players from a particular team, on a game-by-game basis this season

```
select
    display_names.first_name,
    display_names.last_name,
    persons.person_key,
    events.event_key,
    events.start_date_time,
    american_football_passing_stats.passes_attempts,
    american_football_passing_stats.passes_completions,
    american_football_passing_stats.passes_percentage,
    positions.abbreviation

from
    stats, teams, persons, american_football_passing_stats, display_names, events, events_sub_seasons, sub_seasons, person_event_metadata, positions
where
        teams.team_key = '1.nfl.com-t.1'
    and sub_seasons.sub_season_key = '2007_season_regular'
    and events_sub_seasons.event_id = events.id
    and events_sub_seasons.sub_season_id = sub_seasons.id
    and stats.stat_repository_type = 'american_football_passing_stats'
    and stats.stat_repository_id = american_football_passing_stats.id
```

Code generation

Cursor, GitHub Copilot, Devin, Replit,…



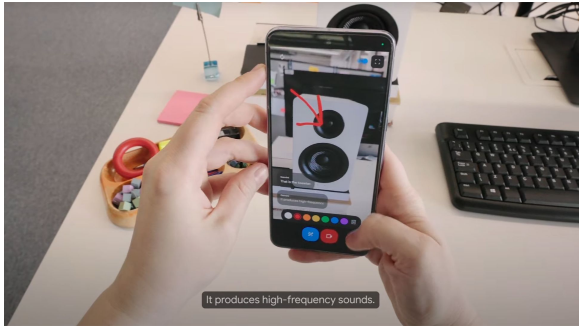
Describe what you'd like to create, and Copilot will generate a draft

Workflow automation

Microsoft Copilot, Multi-On,…

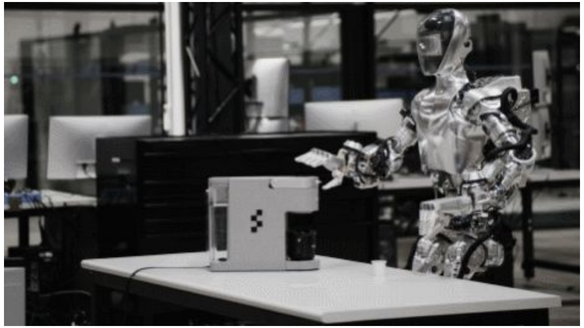
- Education
- Law
- Finance
- Healthcare
- Cybersecurity

…



Personal assistant

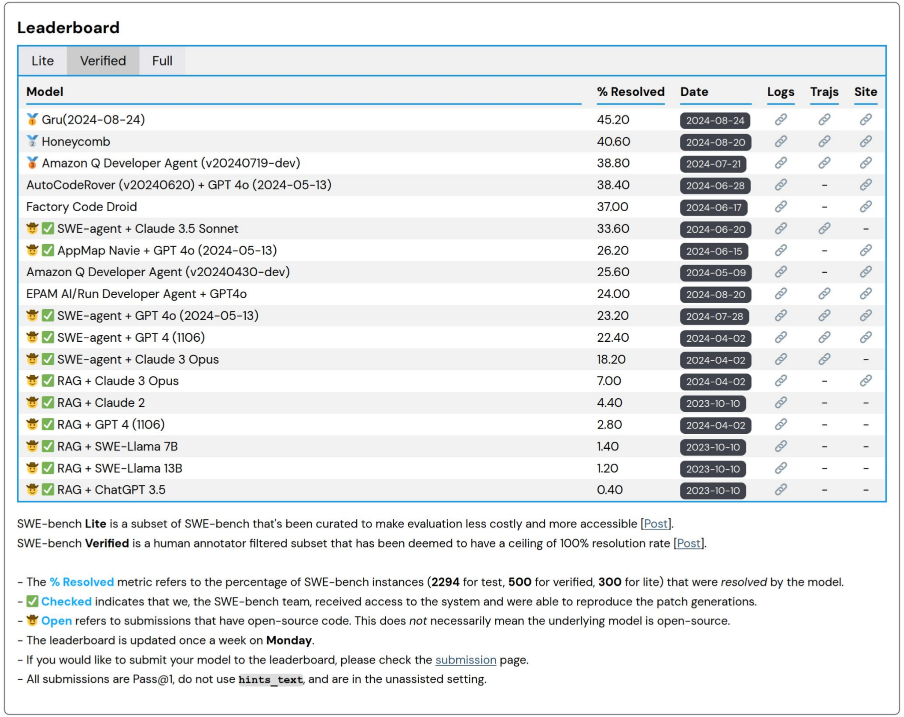
Google Astra, OpenAI GPT-4o,…



Robotics

Figure AI, Tesla Optimus,…

LLM agents are improving



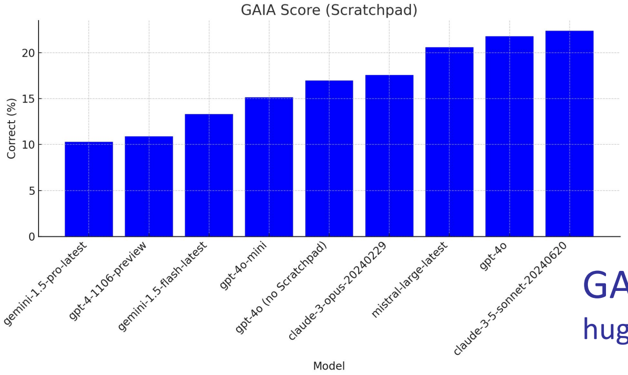
Leaderboard

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	🔗	-	-

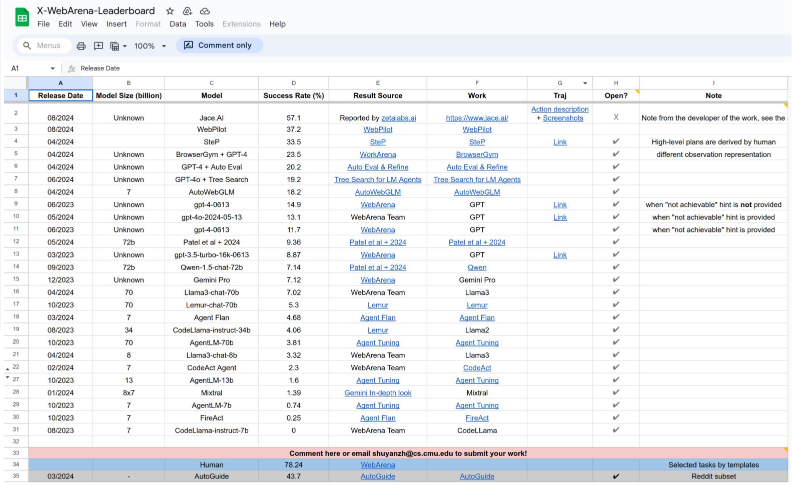
SWE-bench **Lite** is a subset of SWE-bench that's been curated to make evaluation less costly and more accessible [Post].
SWE-bench **Verified** is a human annotator filtered subset that has been deemed to have a ceiling of 100% resolution rate [Post].

- 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



Release Date	Model Size (billion)	Model	Success Rate (%)	Result Source	Work	Traj	Open?	Note
08/2024	Unknown	Jace.AI	57.1	Reported by zetalabs.ai	https://www.jace.ai/	Action description + Screenshots	X	Note from the developer of the work, see the
08/2024		WebPilot	37.2	WebPilot	WebPilot			
04/2024		SteP	33.5	SteP	SteP	Link	✓	High-level plans are derived by human
04/2024	Unknown	BrowserGym + GPT-4	23.5	WorkArena	BrowserGym		✓	different observation representation
04/2024	Unknown	GPT-4 + Auto Eval	20.2	Auto Eval & Refine	Auto Eval & Refine		✓	
06/2024	Unknown	GPT-4o + Tree Search	19.2	Tree Search for LM Agents	Tree Search for LM Agents		✓	
04/2024	7	AutoWebGLM	18.2	AutoWebGLM	AutoWebGLM		✓	
06/2023	Unknown	gpt-4-0613	14.9	WebArena	GPT	Link	✓	when "not achievable" hint is **not** provided
05/2024	Unknown	gpt-4o-2024-05-13	13.1	WebArena Team	GPT	Link	✓	when "not achievable" hint is provided
06/2023	Unknown	gpt-4-0613	11.7	WebArena	GPT		✓	when "not achievable" hint is provided
05/2024	72b	Patel et al + 2024	9.36	Patel et al + 2024	Patel et al + 2024		✓	
03/2023	Unknown	gpt-3.5-turbo-16k-0613	8.87	WebArena	GPT	Link	✓	
09/2023	72b	Qwen-1.5-chat-72b	7.14	Patel et al + 2024	Qwen		✓	
12/2023	Unknown	Gemini Pro	7.12	WebArena	Gemini Pro		✓	
04/2024	70	Llama3-chat-70b	7.02	WebArena Team	Llama3		✓	
10/2023	70	Lemur-chat-70b	5.3	Lemur	Lemur		✓	
03/2024	7	Agent Flan	4.68	Agent Flan	Agent Flan		✓	
08/2023	34	CodeLlama-instruct-34b	4.06	Lemur	Llama2		✓	
10/2023	70	AgentLM-70b	3.81	Agent Tuning	Agent Tuning		✓	
04/2024	8	Llama3-chat-8b	3.32	WebArena Team	Llama3		✓	
02/2024	7	CodeAct Agent	2.3	WebArena Team	CodeAct		✓	
10/2023	13	AgentLM-13b	1.6	Agent Tuning	Agent Tuning		✓	
01/2024	8x7	Mixtral	1.39	Gemini In-depth look	Mixtral		✓	
10/2023	7	AgentLM-7b	0.74	Agent Tuning	Agent Tuning		✓	
10/2023	7	FireAct	0.25	Agent Flan	FireAct		✓	
08/2023	7	CodeLlama-instruct-7b	0	WebArena Team	CodeLLama		✓	
Comment here or email shuyanzh@cs.cmu.edu to submit your work!								
		Human	78.24	WebArena				Selected tasks by templates
03/2024	-	AutoGuide	43.7	AutoGuide	AutoGuide		✓	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

Berkeley UNIVERSITY OF CALIFORNIA

Xinyun Chen

DeepMind

Denny Zhou

DeepMind

Shunyu Yao

OpenAI

Chi Wang

DeepMind

Jerry Liu

LlamaIndex

Burak Gokturk

Google

Omar Khattab

databricks

Graham Neubig

Carnegie Mellon University

Nicolas Chapados

servicenow

Yuandong Tian

Meta AI

Jim Fan

NVIDIA

Percy Liang

Stanford University

Ben Mann

ANTHROP\C

Berkeley RDI

Course Work

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

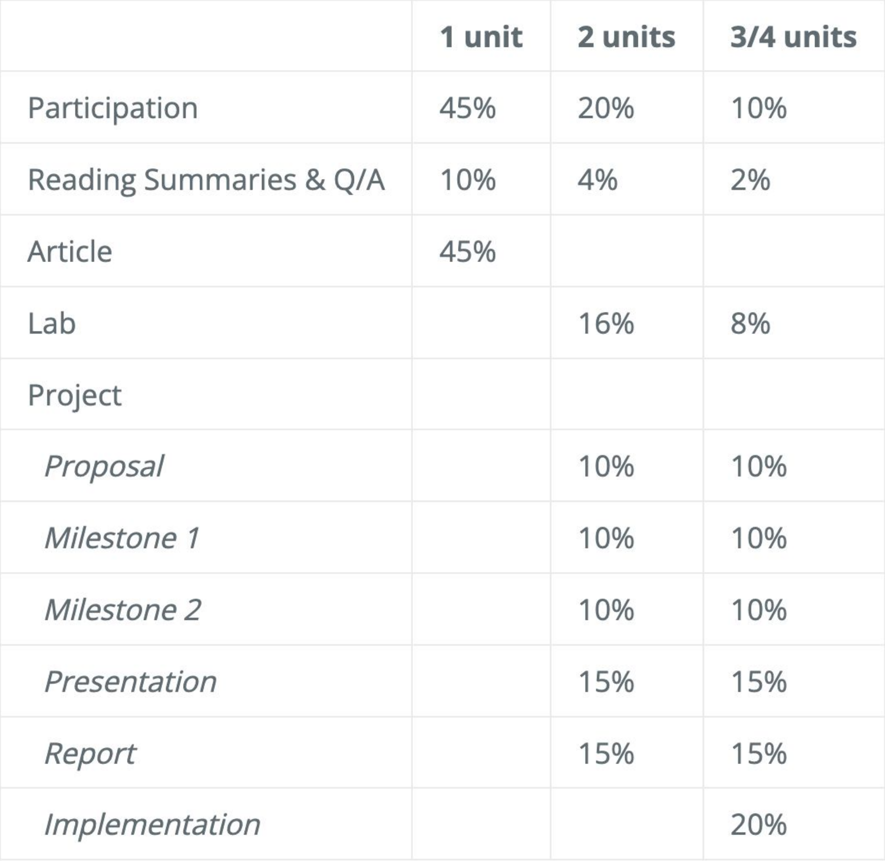
Grading

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			
Proposal		10%	10%
Milestone 1		10%	10%
Milestone 2		10%	10%
Presentation		15%	15%
Report		15%	15%
Implementation			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)

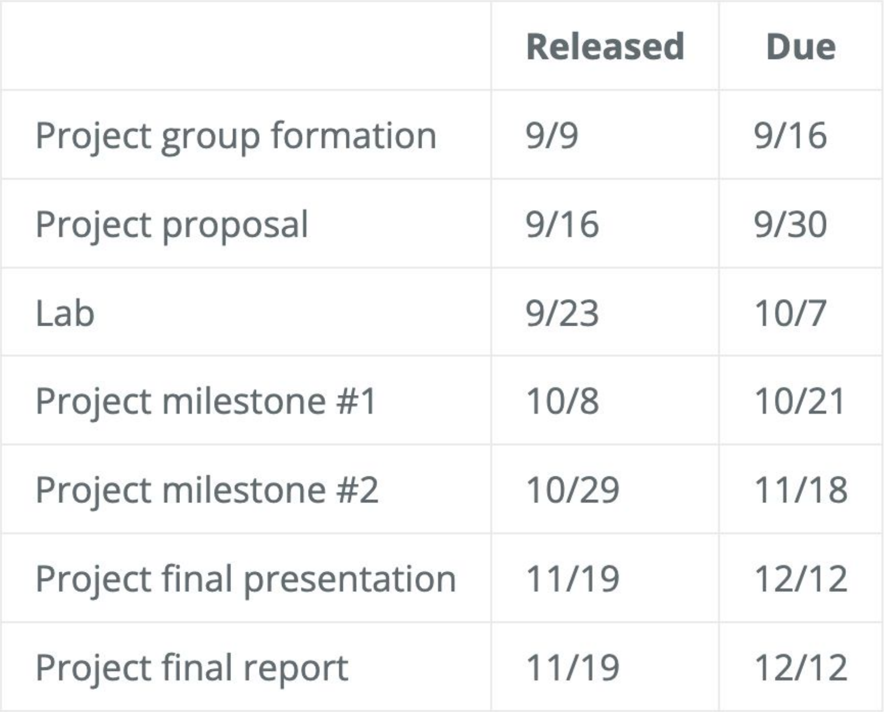
Safety Track

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

Decentralized and Multi-agent Track

- Enhance decentralized multi-agent systems

Timeline



	Released	Due
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