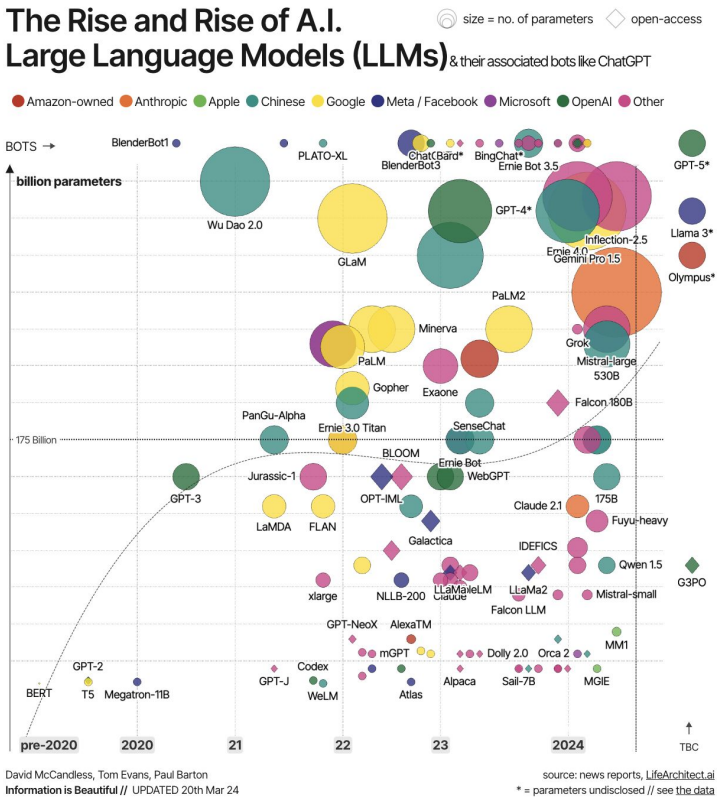
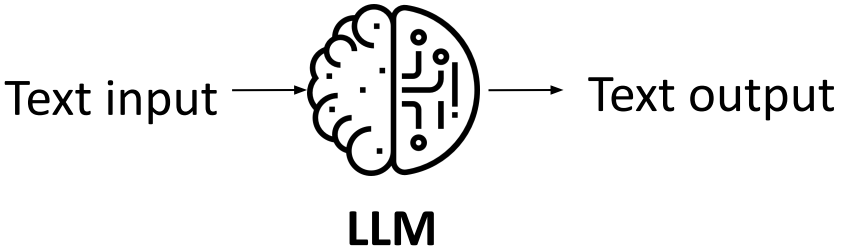
CS 294/194-196:
Large Language Model Agents

Teaching Staff

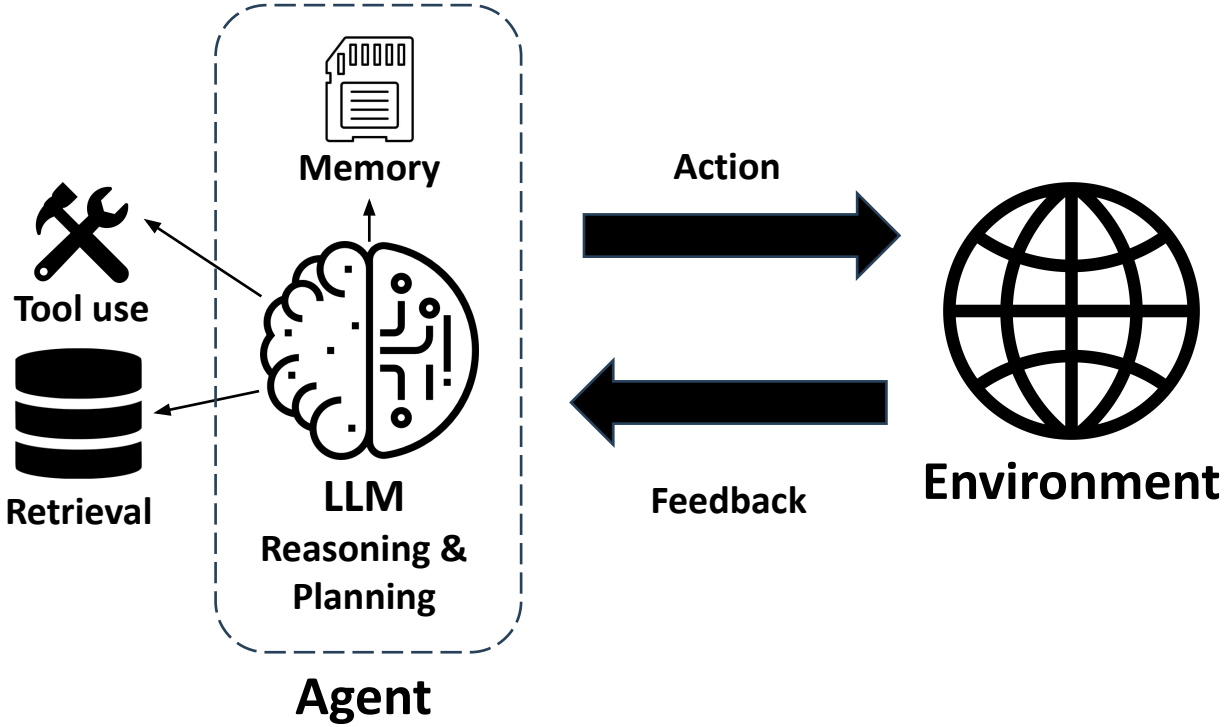
- **Instructor: Prof. Dawn Song**
- **(guest) Co-instructor: Dr. Xinyun Chen**
- GSIs: Alex Pan & Sehoon Kim
- Readers: Tara Pande & Ashwin Dara

Accelerated development of large language models (LLMs)

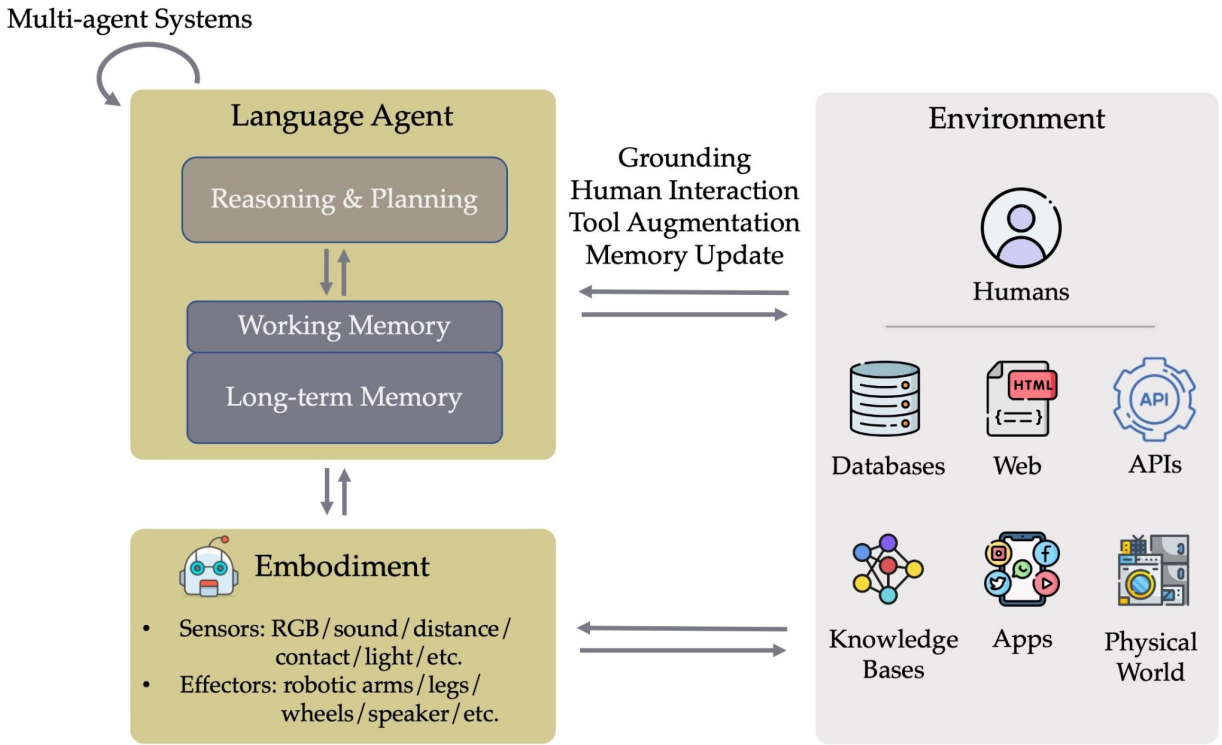




LLM agents: enabling LLMs to interact with the environment

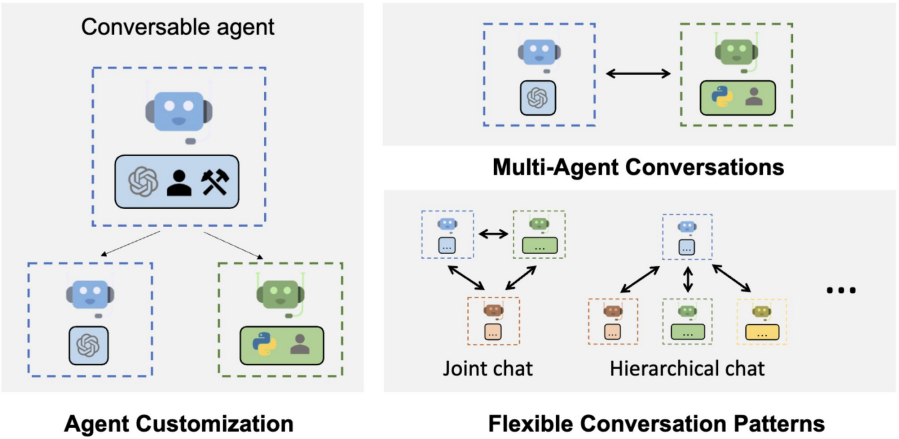


LLM Agents in Diverse Environments



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

Multi-agent collaboration: division of labor for complex tasks



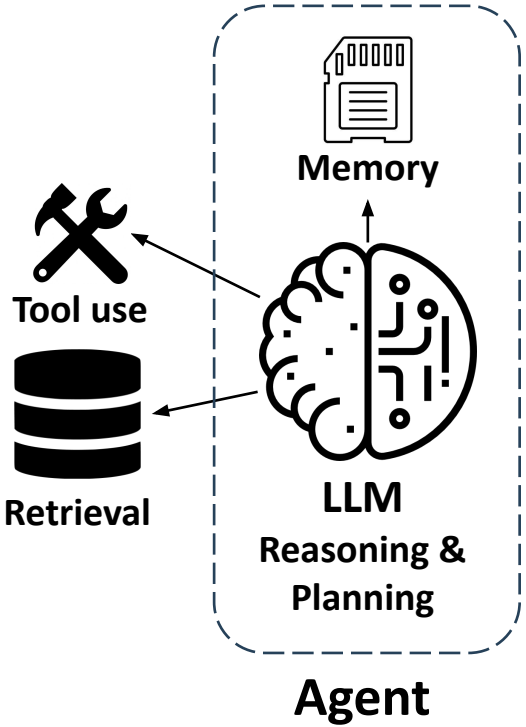
Specialized agents for different subtasks

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

Emergence of social behaviors with role-play LLMs

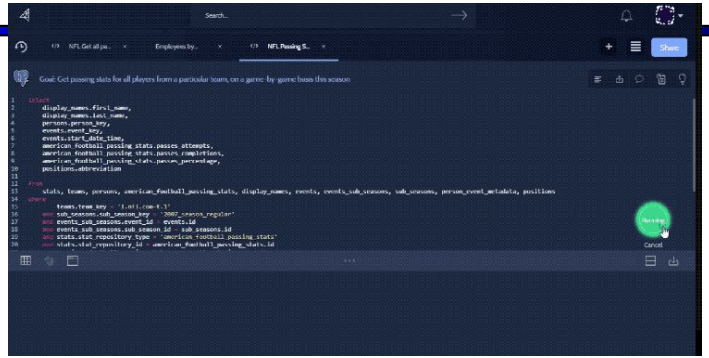
Generative agents, Project Sid,...

Why empowering LLMs with the agent framework



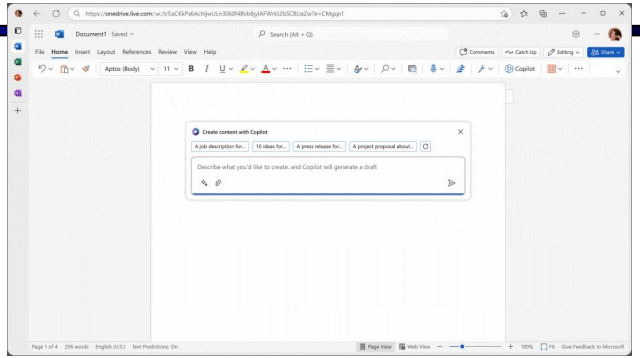
- Solving real-world tasks typically involves a trial-and-error process
- Leveraging external tools and retrieving from external knowledge expand LLM's capabilities
- Agent workflow facilitates complex tasks
  - Task decomposition
  - Allocation of subtasks to specialized modules
  - Division of labor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

LLM agents transformed various applications



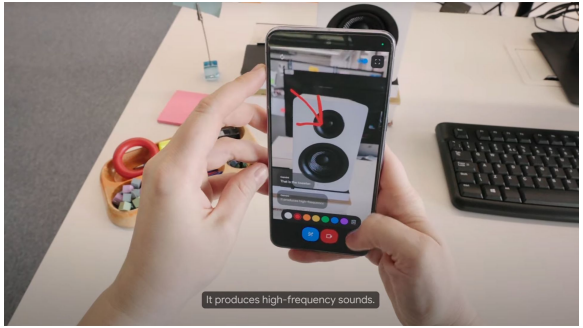
Code generation

Cursor, GitHub Copilot, Devin, Replit,…



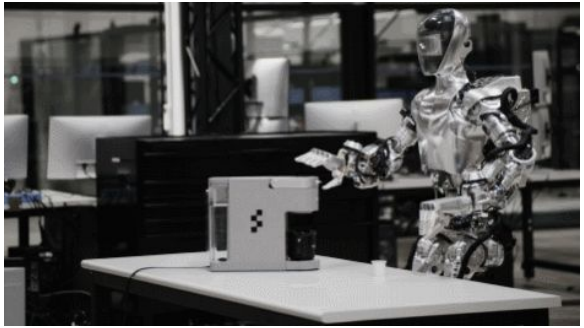
Workflow automation

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

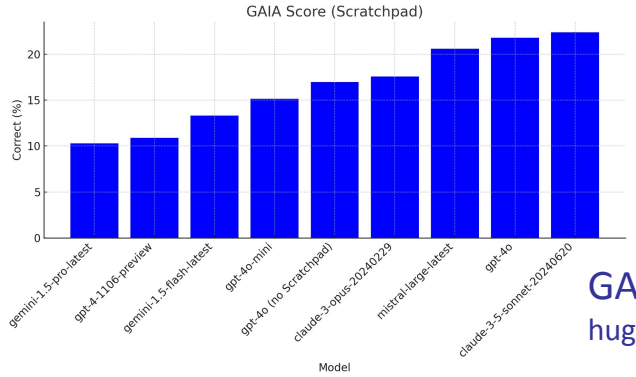
Lite | Verified | Full

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	🔗	-	-

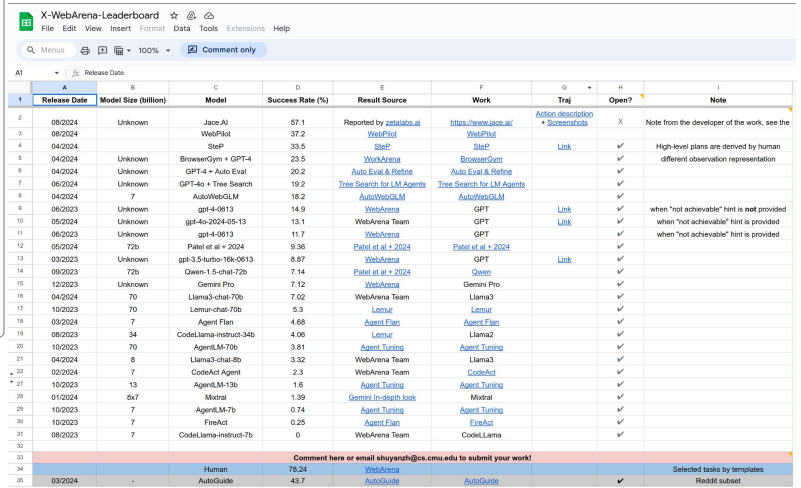
SWE-bench **Lite** is a subset of SWE-bench that's been curated to make evaluation less costly and more accessible [Post].
SWE-bench **Verified** is a human annotator filtered subset that has been deemed to have a ceiling of 100% resolution rate [Post].

- 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



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

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

ANTHROPIC

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