AgentBoard: An Analytical Evaluation Board of Multi-Turn LLM Agents


Chang Ma*,
HKU


Junlei Zhang*,
Westlake Univ


Zhihao Zhu*,
SJTU


Cheng Yang*,
Tsinghua Univ


Yujiu Yang,
Tsinghua Univ


Yaohui Jin,
Tsinghua Univ


Zhenzhong Lan,
Westlake Univ


Lingpeng Kong,
HKU


Junxian He,
HKUST

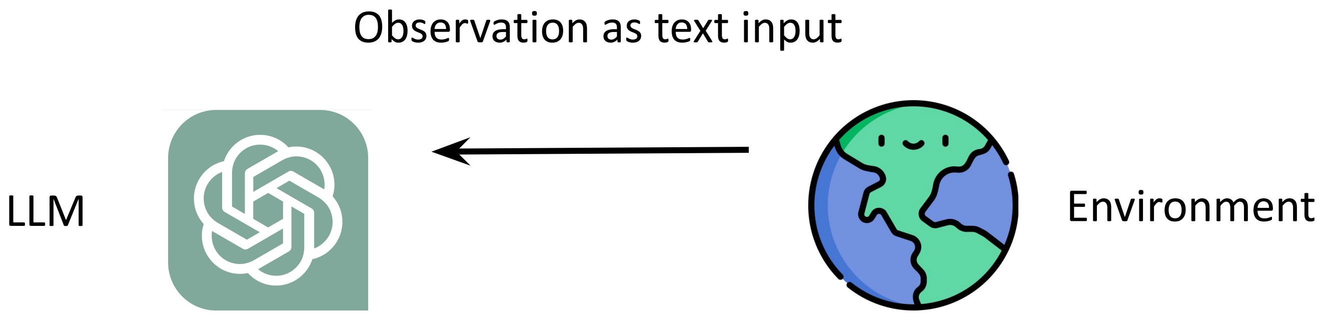
Background

LLM Powered Autonomous Agents

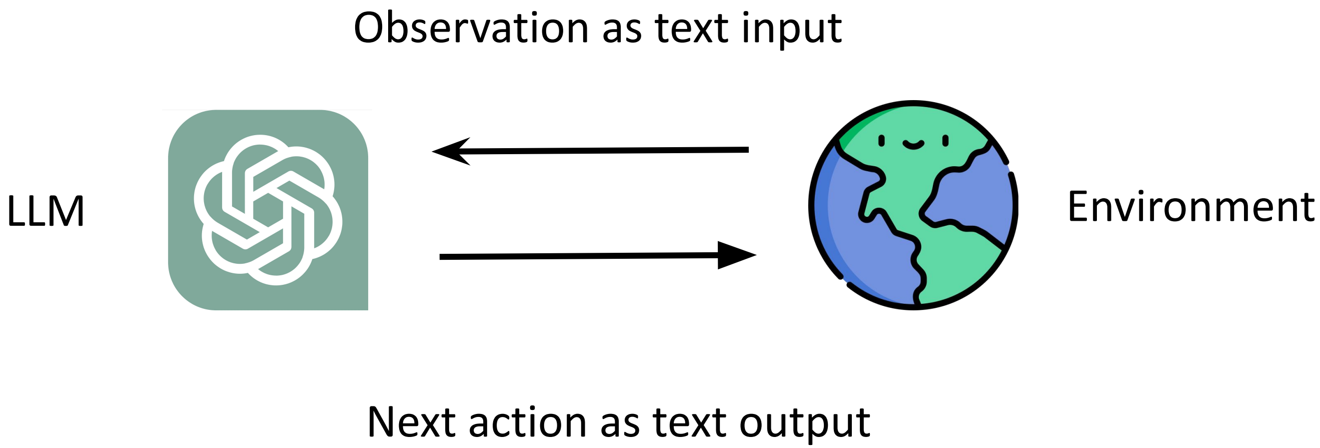
LLM

Environment

LLM Powered Autonomous Agents



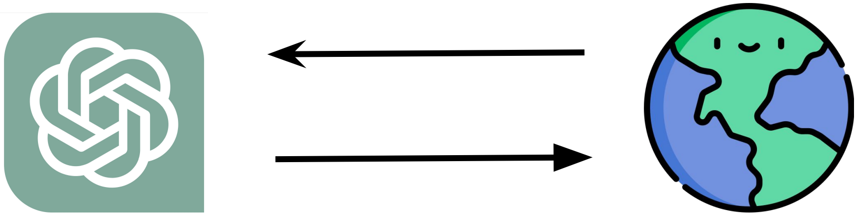
LLM Powered Autonomous Agents



LLM Powered Autonomous Agents

Observation as text input

LLM

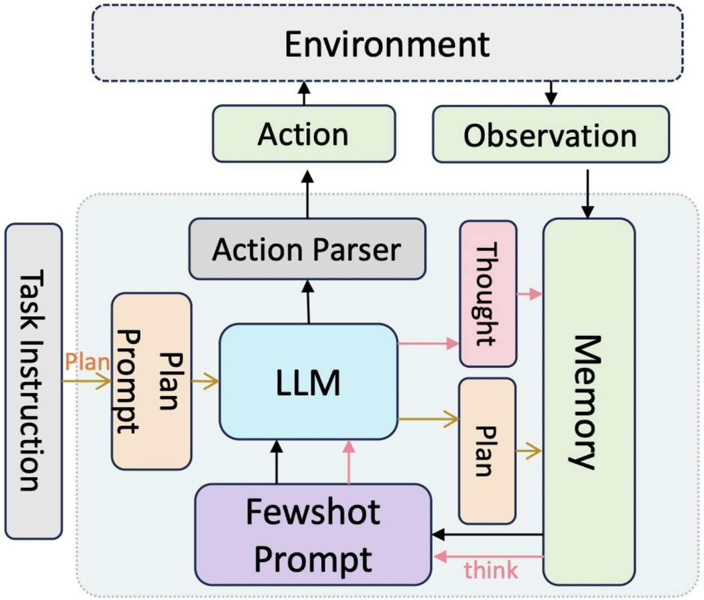


Environment

Next action as text output

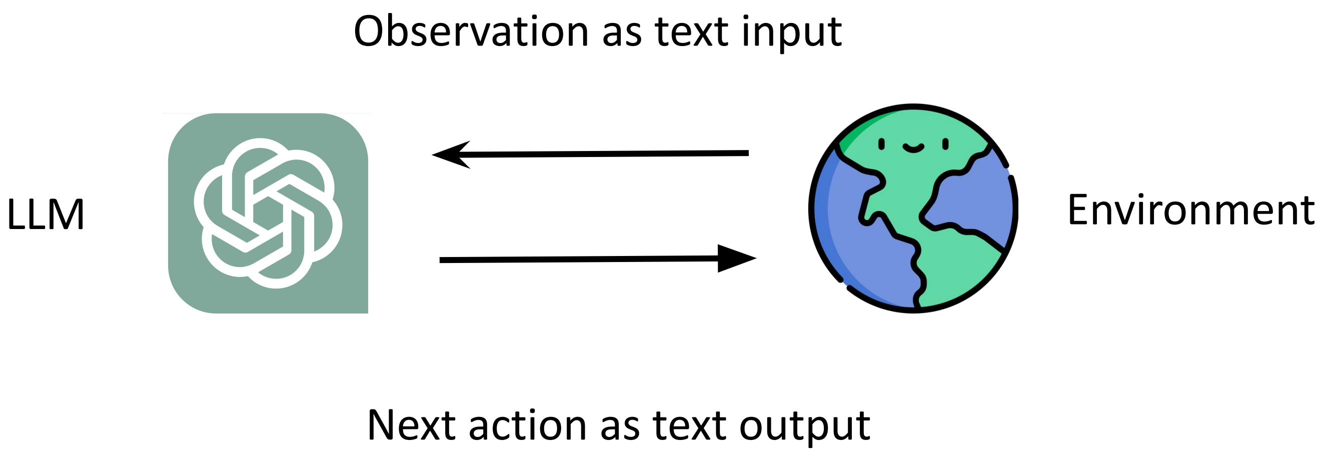
Autoregressive LLMs can reason and plan. They could interact with environments as agents.

Evaluating LLM Agents

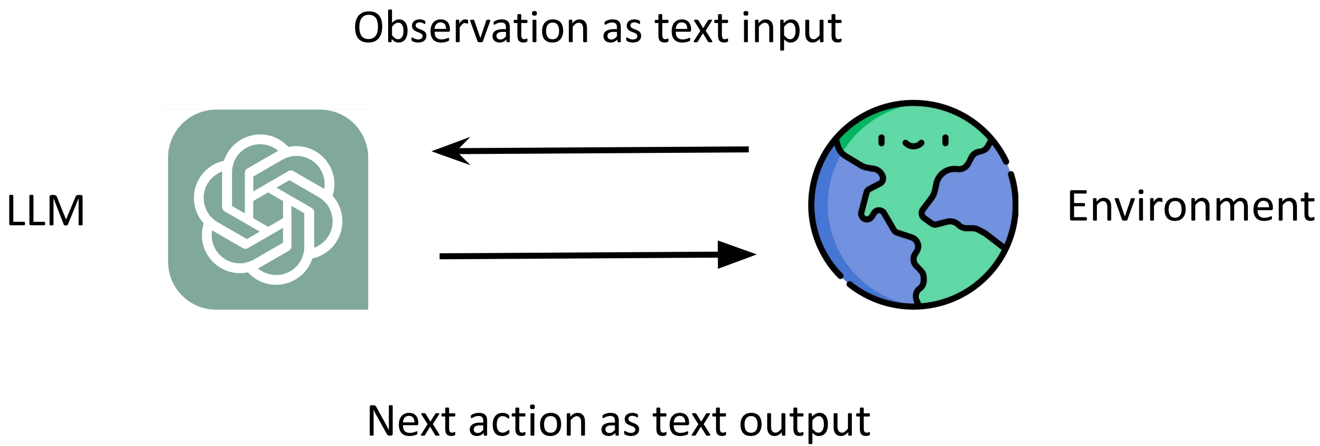


BOLAA, Liu et al 2023

Evaluating LLM as Agents



Evaluating LLM as Agents



Use simple, unified agent design to understand the varying agentic abilities of different LLM.

How to Comprehensively benchmark LLM as Agents ?

Motivation - LLM Agent Benchmark

Goal:

Motivation - LLM Agent Benchmark

Goal:

Compare key agentic abilities of LLM through benchmarking.

Motivation - LLM Agent Benchmark

Goal:

Compare key agentic abilities of LLM through benchmarking.

Our Work: AgentBoard

Motivation - LLM Agent Benchmark

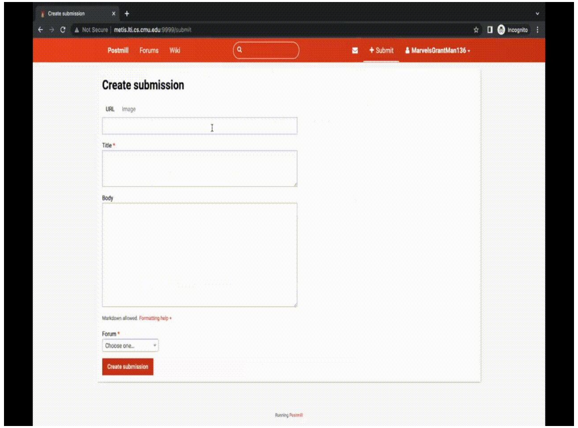
Goal:

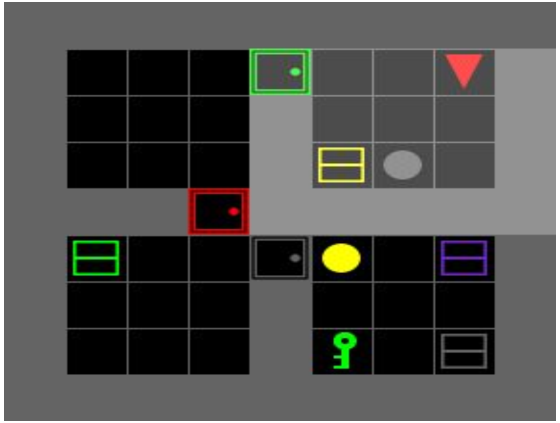
Compare key agentic abilities of LLM through benchmarking.

Our Work: AgentBoard

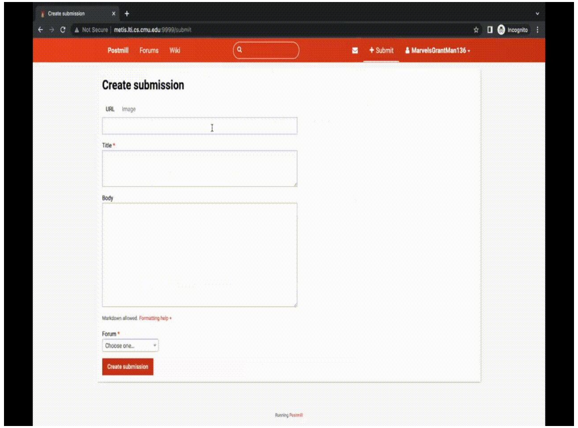
- Unified and Diverse Tasks

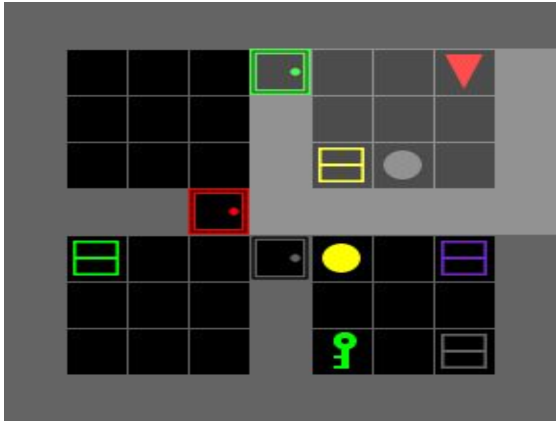
Evaluating LLM as Generalist





Evaluating LLM as Generalist





LLM Agents possess generalist ability. It’s essential to evaluate LLM as Agents on a diverse set of tasks.

Motivation - LLM Agent Benchmark

Goal:

Compare key agentic abilities of LLM through benchmarking.

Our Work: AgentBoard

- Unified and Diverse Tasks - Multi-turn

Motivation - LLM Agent Benchmark

Goal:

Compare key agentic abilities of LLM through benchmarking.

Our Work: AgentBoard

- Unified and Diverse Tasks - Multi-turn, Partially-observable

Motivation - LLM Agent Benchmark

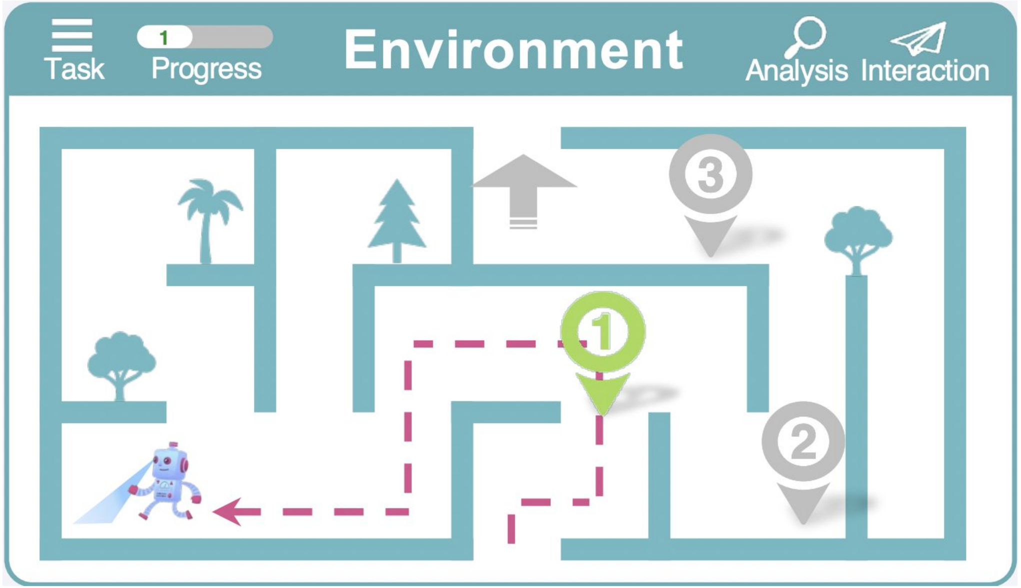
Goal:

Compare key agentic abilities of LLM through benchmarking.

Our Work: AgentBoard

- Unified and Diverse Tasks - Multi-turn, Partially-observable

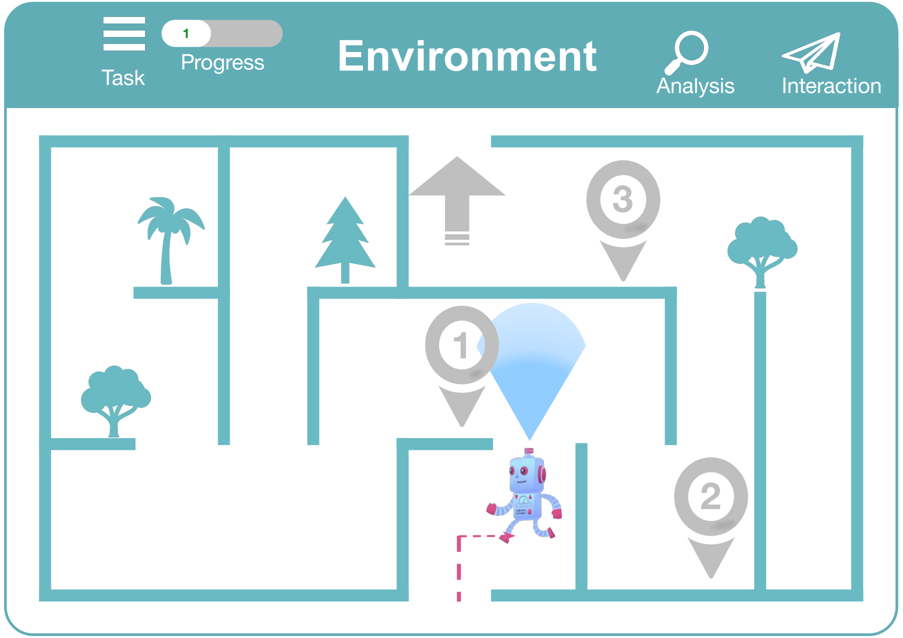
Important Features for Agent Evaluation



Multi-Turn

Partially Observable

Important Features for Agent Evaluation



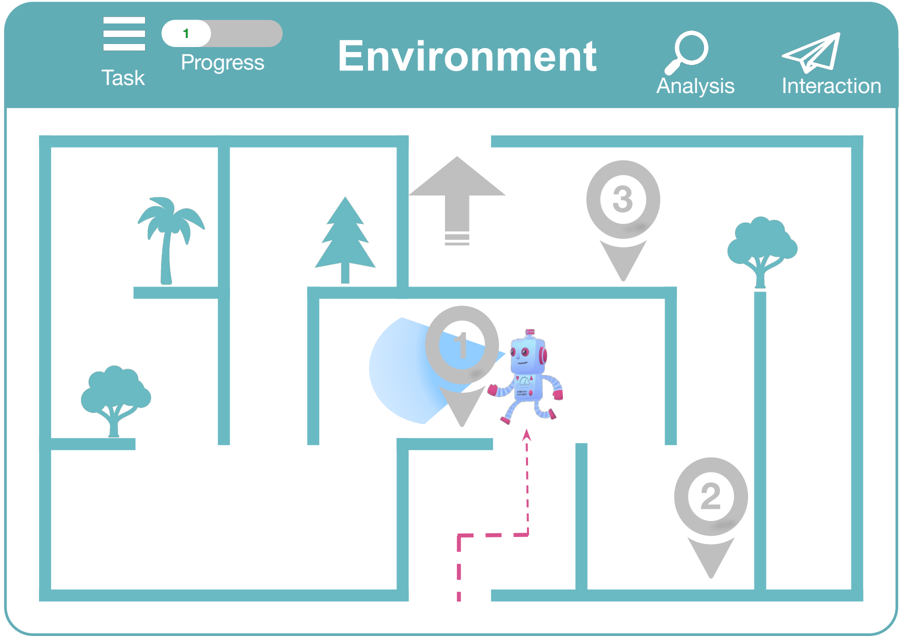
1. **Multi-Turn**

Step 1:

: Action 1

: Observation 1

Important Features for Agent Evaluation



1. **Multi-Turn**

Step 1:

: Action 1

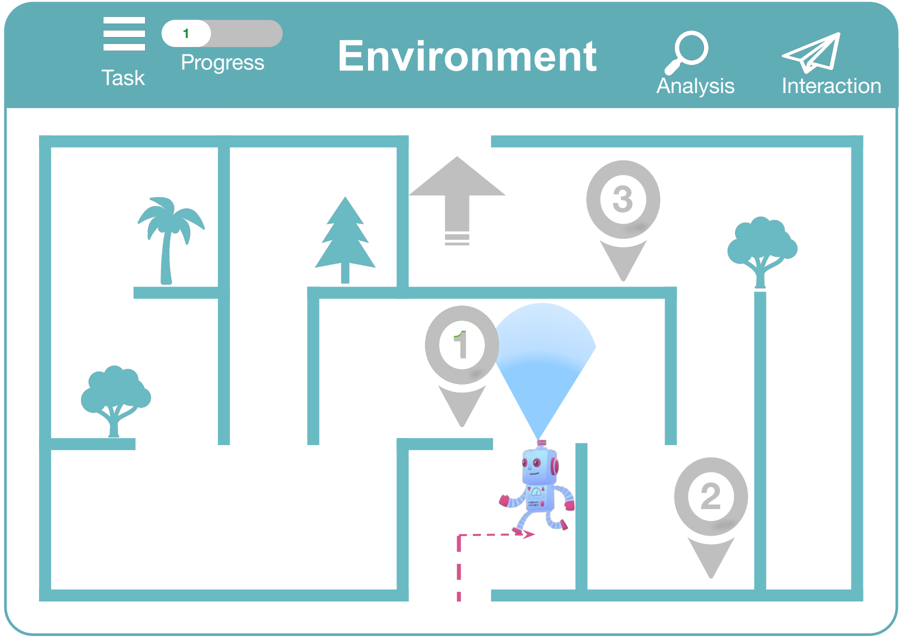
: Observation 1

Step 2:

: Action 2

: Observation 2

Important Features for Agent Evaluation



1. **Multi-Turn**

Step 1:

: Action 1

: Observation 1

Step 2:

: Action 2

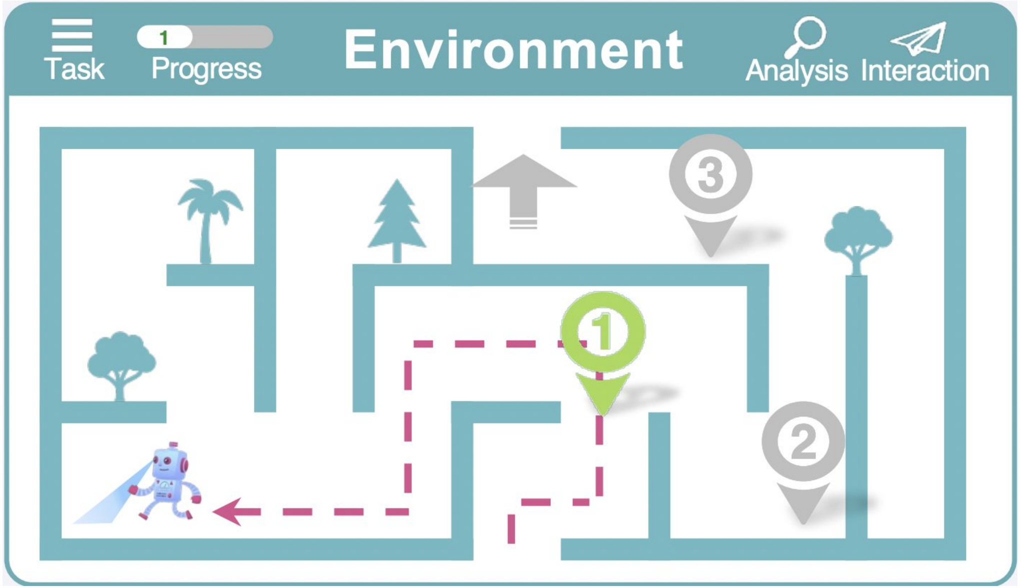
: Observation 2

Step 3:

: Action 3

: Observation 3

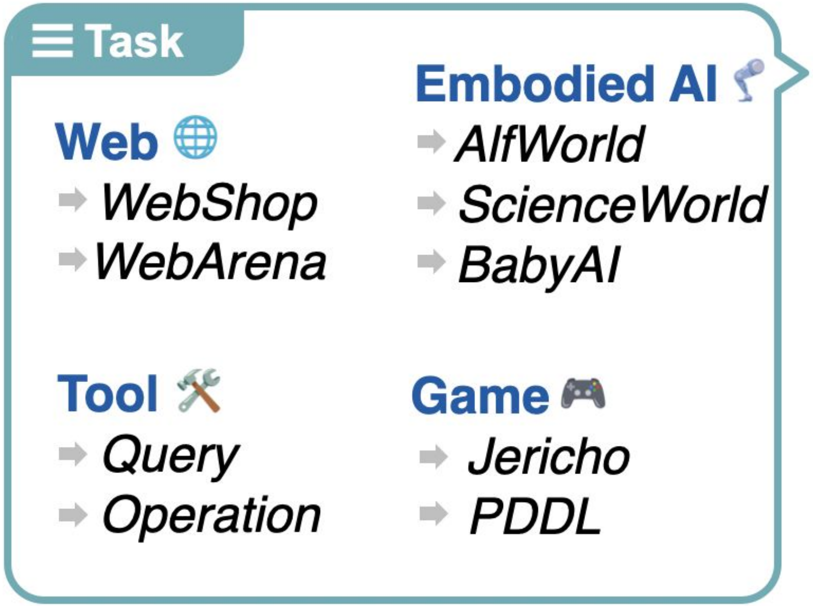
Important Features for Agent Evaluation



Multi-Turn

Partially Observable

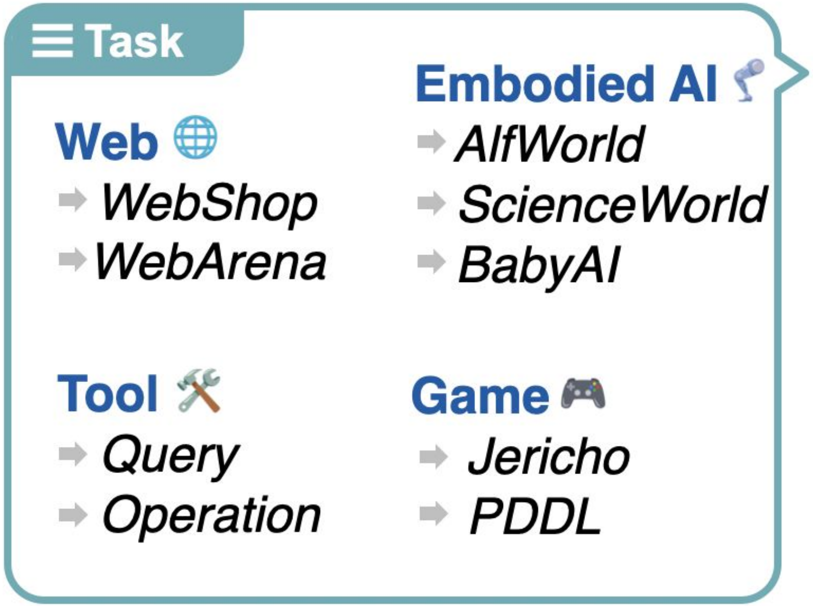
Unified and Diverse Tasks



Diverse testbeds:

- **9** Tasks
- **1012** Environments
- **6-20 Turns** Interaction
- **Diverse Action Space**

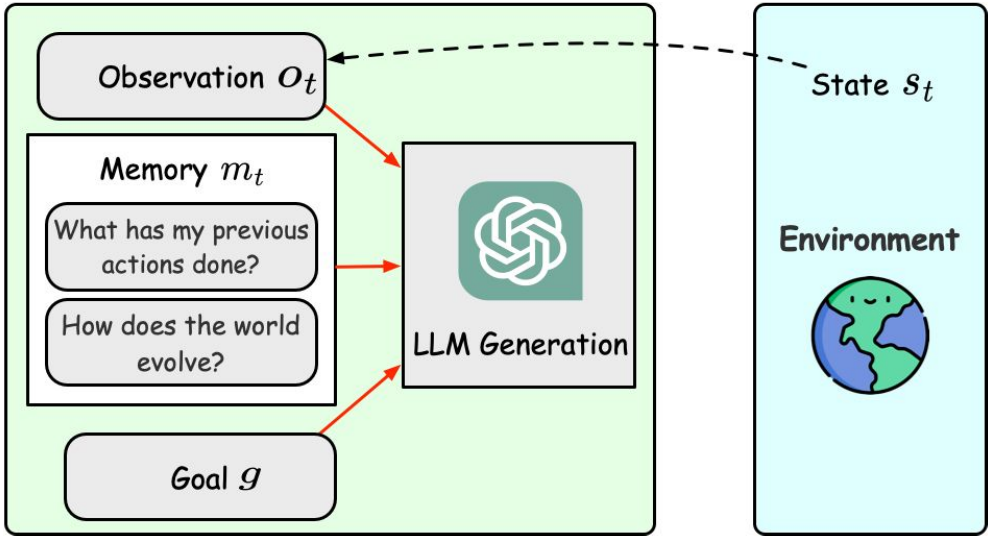
Unified and Diverse Tasks



Unified Formatting:

- **Multi-turn** interactions.
- **Natural language interface**.
- Unified observations and actions format.

Unified Framework for Evaluating LLM Agents



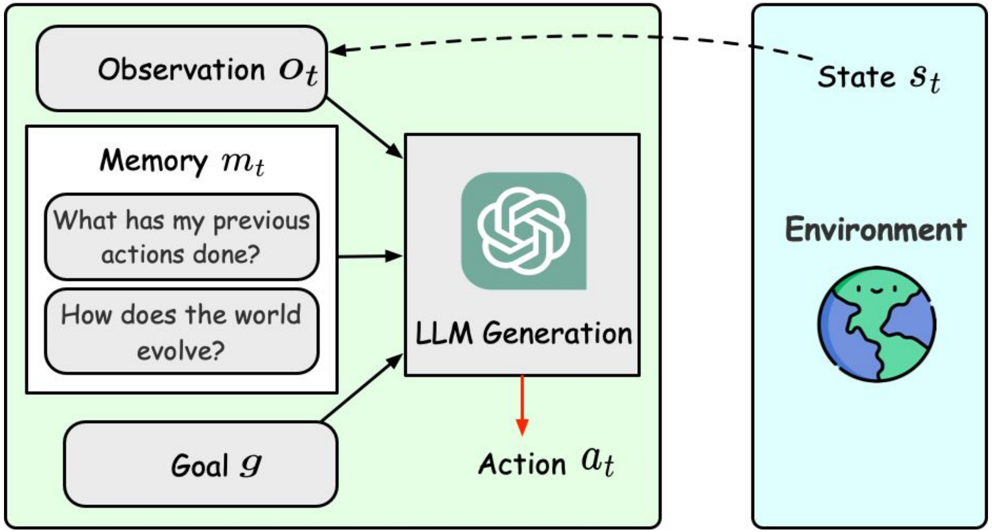
>**[Instruction]:** You are an agent in a virtual science school environment, tasked to interact with various elements. Here are commands that you can use: open, close, look around ...

>[Goal]: You should perform actions to accomplish the goal: **boil some water.**

>[Memory]:
Observation: This room is called the workshop. In it, you see: the agent, a table, a door to the hallway…
Action: go to kitchen
Observation: You move to the kitchen.
Action: open cupboard
Observation: The cupboard is open. There is a mug, a thermometer, and a cloth.

LLM is prompted with current task goal, observation, as well as previous **memory**.

Unified Framework for Evaluating LLM Agents



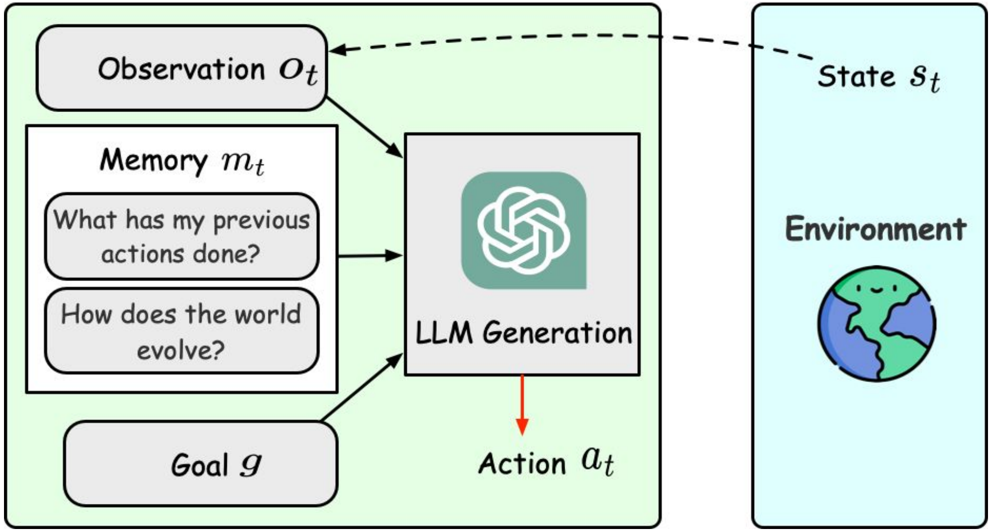
>**[Instruction]:** You are an agent in a virtual science school environment, tasked to interact with various elements. Here are commands that you can use: open, close, look around ...

>[**Goal**]: You should perform actions to accomplish the goal: **boil some water.**

>[**Memory**]:
Observation: This room is called the workshop. In it, you see: the agent, a table, a door to the hallway…
Action: go to kitchen
Observation: You move to the kitchen.
Action: open cupboard
Observation: The cupboard is open. There is a mug, a thermometer, and a cloth.

Action:

Unified Framework for Evaluating LLM Agents



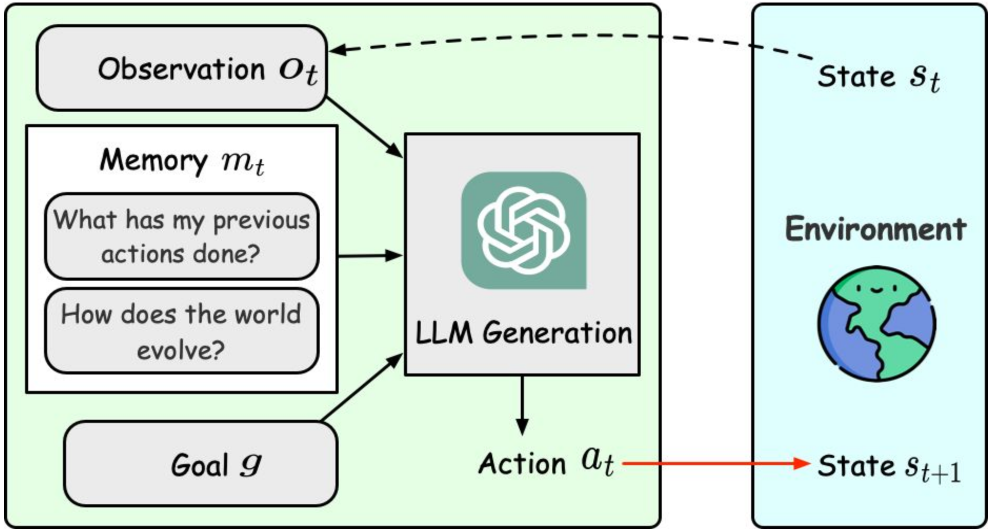
>**[Instruction]:** You are an agent in a virtual science school environment, tasked to interact with various elements. Here are commands that you can use: open, close, look around ...

>[**Goal**]: You should perform actions to accomplish the goal: **boil some water.**

>[**Memory**]:
Observation: This room is called the workshop. In it, you see: the agent, a table, a door to the hallway…
Action: go to kitchen
Observation: You move to the kitchen.
Action: open cupboard
Observation: The cupboard is open. There is a mug, a thermometer, and a cloth.

Action: pickup mug from the cupboard

Unified Framework for Evaluating LLM Agents



>[**Instruction**]: You are an agent in a virtual science school environment, tasked to interact with various elements. Here are commands that you can use: open, close, look around ...

>[**Goal**]: You should perform actions to accomplish the goal: **boil some water.**

>[**Memory**]:
Observation: This room is called the workshop. In it, you see: the agent, a table, a door to the hallway…
Action: go to kitchen
Observation: You move to the kitchen.
Action: open cupboard
Observation: The cupboard is open. There is a mug, a thermometer, and a cloth.

Action: pickup mug from the cupboard

Observation: You move the mug to the inventory.

Motivation - LLM Agent Benchmark

Goal:

Compare key agentic abilities of LLM through benchmarking.

Our Work: AgentBoard

- Unified and Diverse Tasks - Multi-turn, Partially-observable
-

Motivation - LLM Agent Benchmark

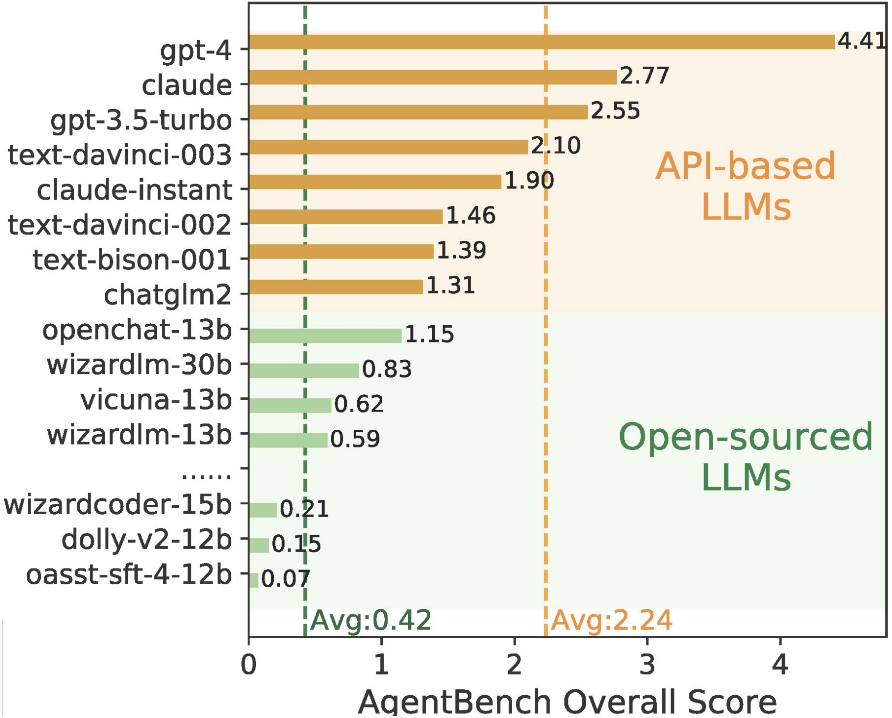
Goal:

Compare key agentic abilities of LLM through benchmarking.

Our Work: AgentBoard

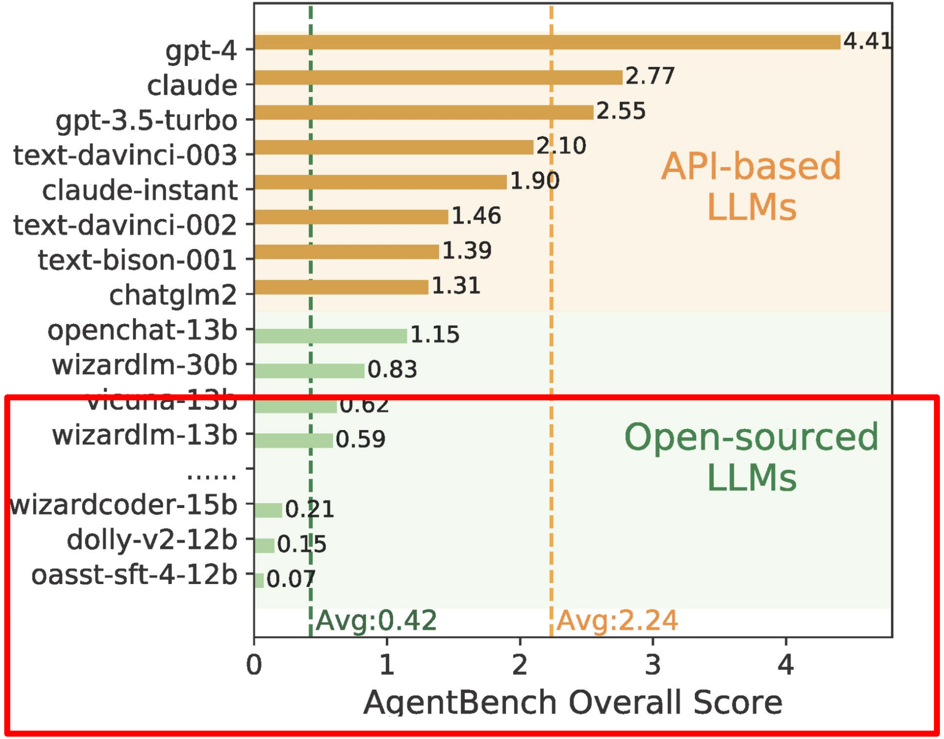
- Unified and Diverse Tasks - Multi-turn, Partially-observable
- Fine-grained Evaluation Metrics

Why do we need Fine-grained Evaluation Metrics?



Liu, Xiao, et al. "Agentbench: Evaluating llms as agents."(2023).

Why do we need Fine-grained Evaluation Metrics?



Success rate is not discriminative enough for opensource models.

Liu, Xiao, et al. "Agentbench: Evaluating llms as agents."(2023).

Fine-grained Evaluation Metrics

Task: put a clean bowl in the fridge

go to countertop 1	pickup bowl 1	go to sinkbasin 1	clean bowl 1 in sinkbasin	put bowl 1 in fridge 1
Success rate: 0 Progress rate: 0.25	Success rate: 0 Progress rate: 0.5	Success rate: 0 Progress rate: 0.5	Success rate: 0 Progress rate: 0.75	Success rate: 1 Progress rate: 1

Progress rate metric accurately reflects LM agents' goal attainment at various stages.

Fine-grained Progress Rate Calculation

f(goal state, current state)

Match current state against goal state.

Fine-grained Progress Rate Calculation

f(goal state, current state)

Task: Insert "Nelson 99 75 80 79" and "Robert 63 75 92 72" into the "Sheet9" and sort this table by "Name" in ascending order.

Nelson	Robert
99	63
80	75
79	92
75	72

Progres Rate: 0.6

Progres-Rate-Match: Directly calculate state similarity.

Fine-grained Progress Rate Calculation

f(goal state, current state)

go to countertop 1 | pickup bowl 1 | go to sinkbasin 1 | clean bowl 1 in sinkbasin | put bowl 1 in fridge 1

explore and find bowl
Progress rate: 0.25

pickup and carry bowl
Progress rate: 0.5

clean the bowl
Progress rate: 0.75

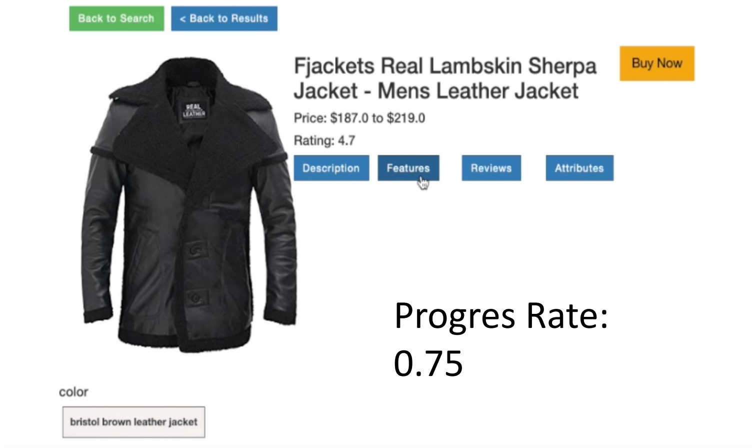
put the bowl in fridge
Progress rate: 1.0

Progres-Rate-Subgoal: Human annotate subgoal decomposition.
Calculate percentage of subgoals attained.

Fine-grained Progress Rate Calculation

f(goal state, current state)

Task: buy women fur leather jacket



Progres Rate: 0.75

women fur leather jacket

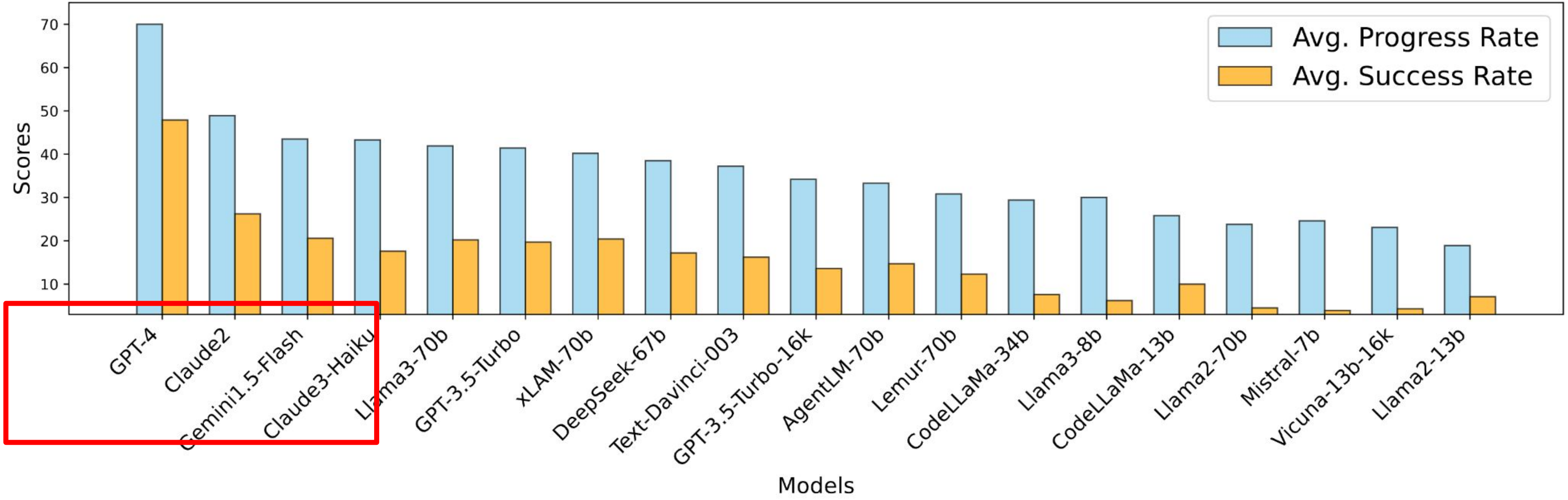
Task: Insert "Nelson 99 75 80 79" and "Robert 63 75 92 72" into the "Sheet9" and sort this table by "Name" in ascending order.

Nelson	Robert
99	63
80	75
79	92
75	72

Progres Rate: 0.6

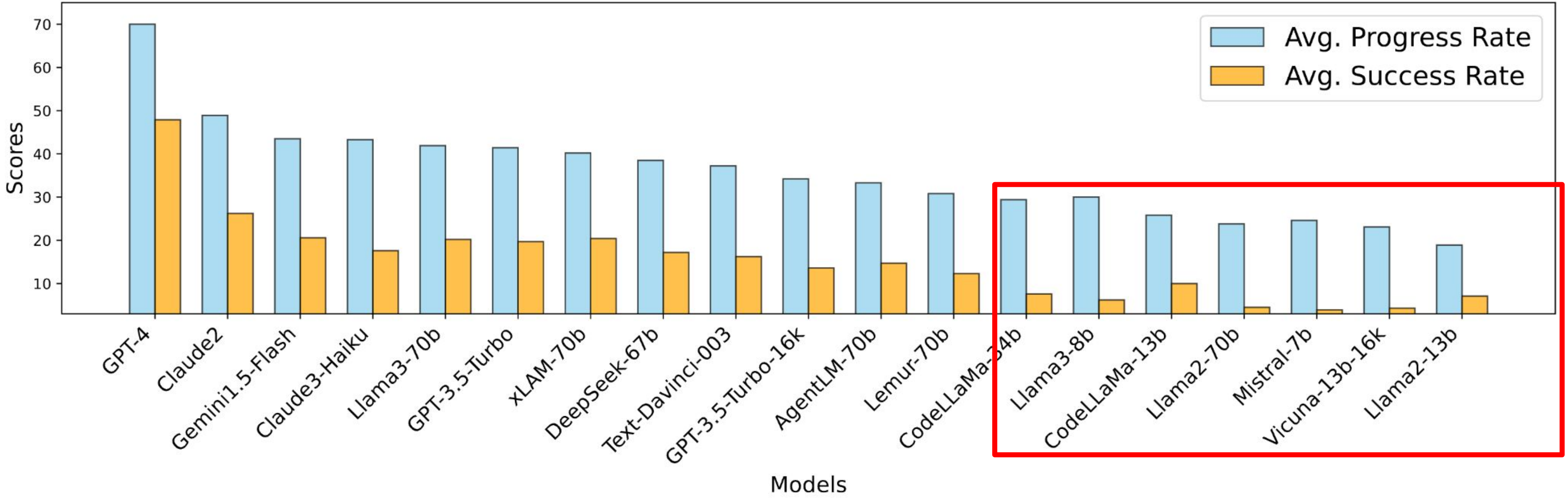
Progres-Rate-Match: Directly calculate state similarity.

Main Results



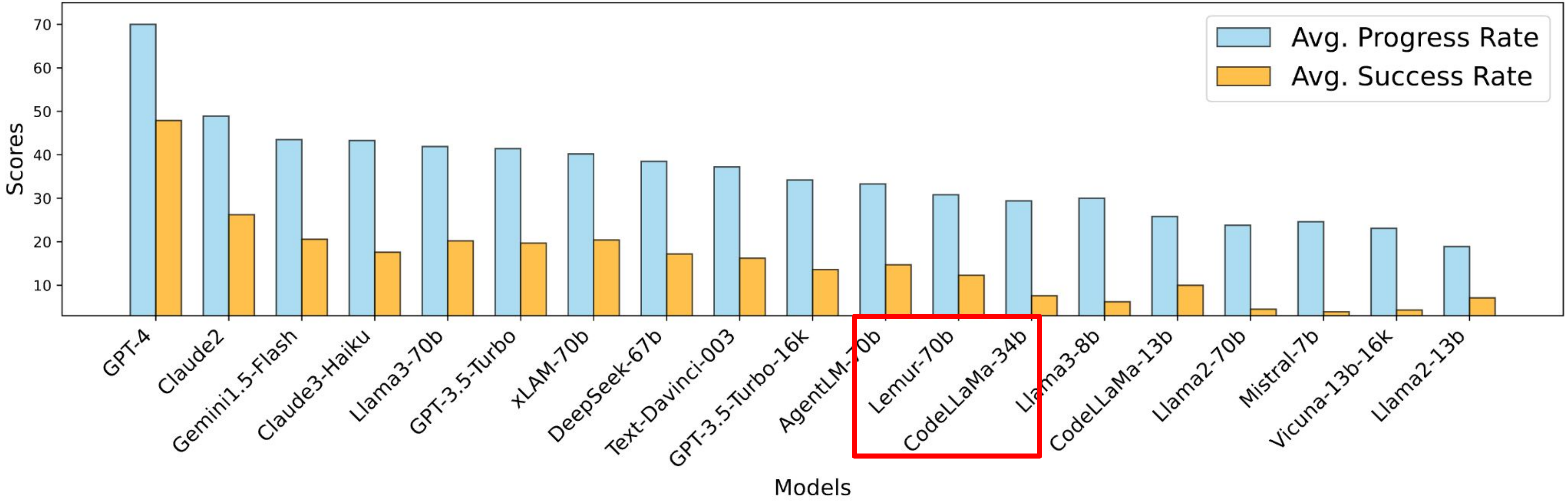
Proprietary models outperform the open-weight ones.

Main Results



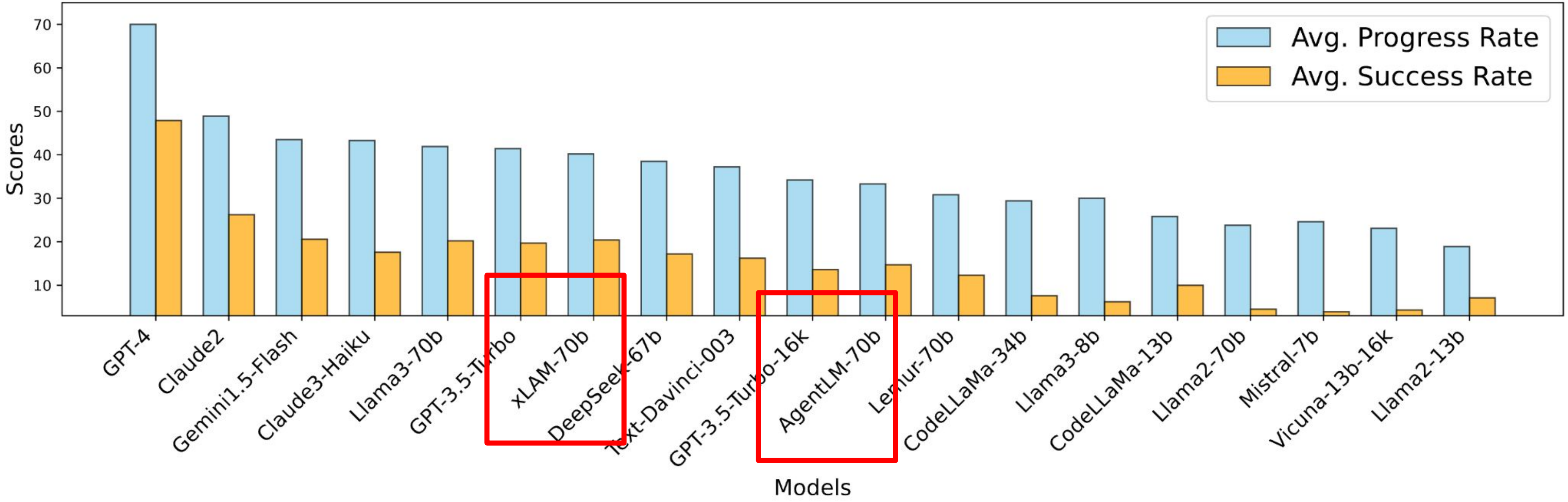
Progress Rate is more informative and discriminative than success rate.

Main Results



Strong coding skills help agent tasks.

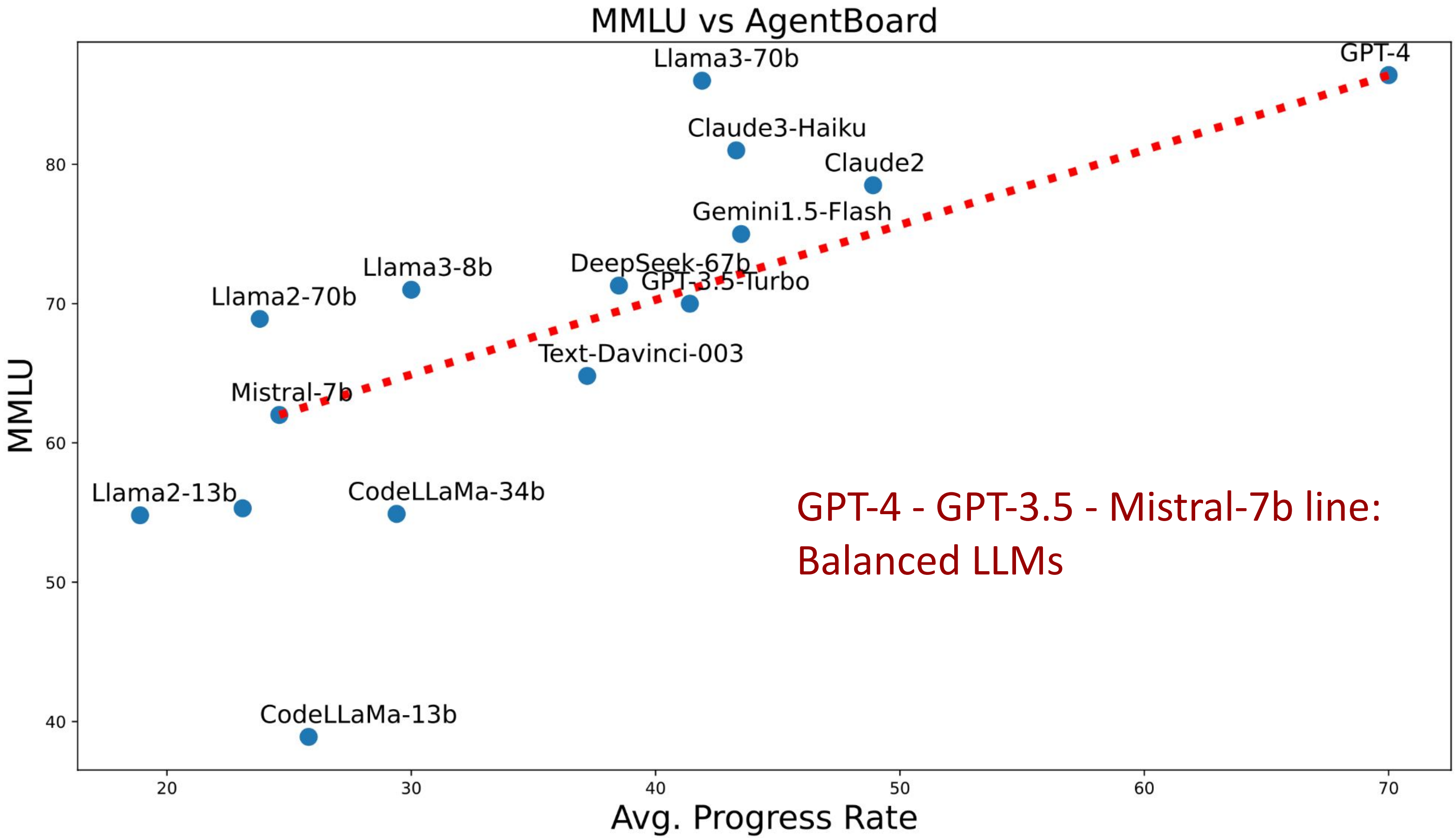
Main Results



Agent tuning improves general agentic abilities of LLM.

Analytical Benchmarking: What makes a LLM better as agents?

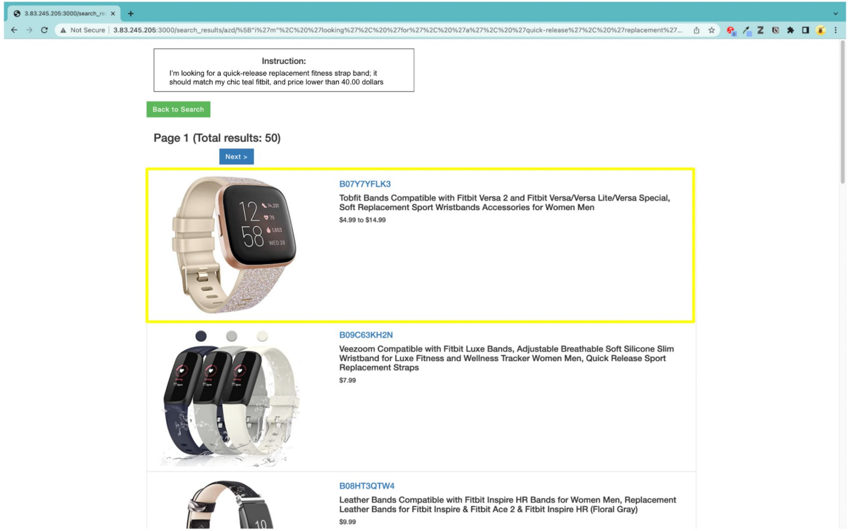
Better LLMs may not be better agent models



What makes a LLM a better agent ?

Understanding why some LLMs are better agents require independent evaluation of **Each Agent Ability**.

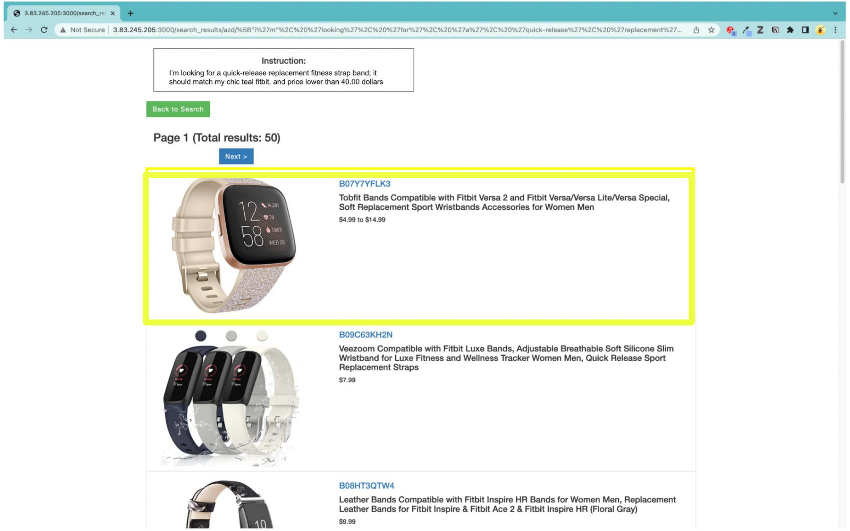
LLM Grounding Ability



Available Actions:

Click [back to search]
Click [Next >]
Click [Tobfit Bands…]
Click [Veezoom Compatible …]
Click [Leather Bands …]

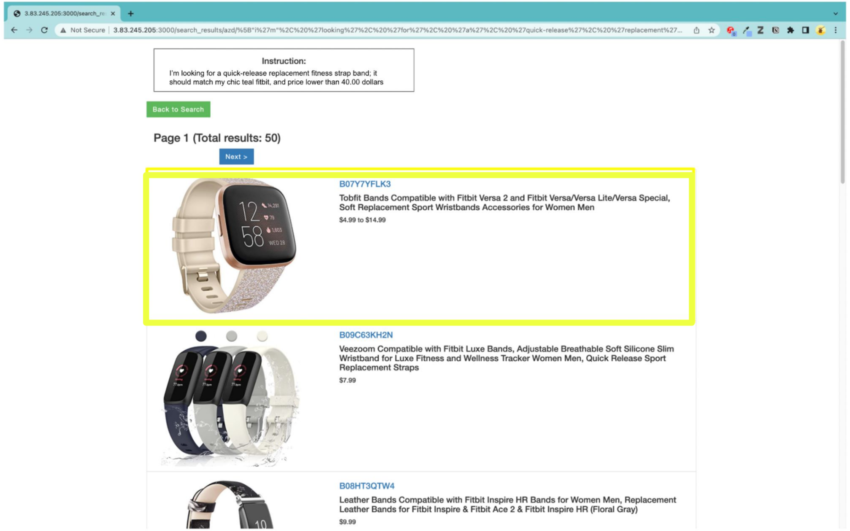
LLM Grounding Ability



Available Actions:

Click [back to search]
Click [Next >]
Click [Tobfit Bands…]
Click [Veezoom Compatible …]
Click [Leather Bands …]

LLM Grounding Ability



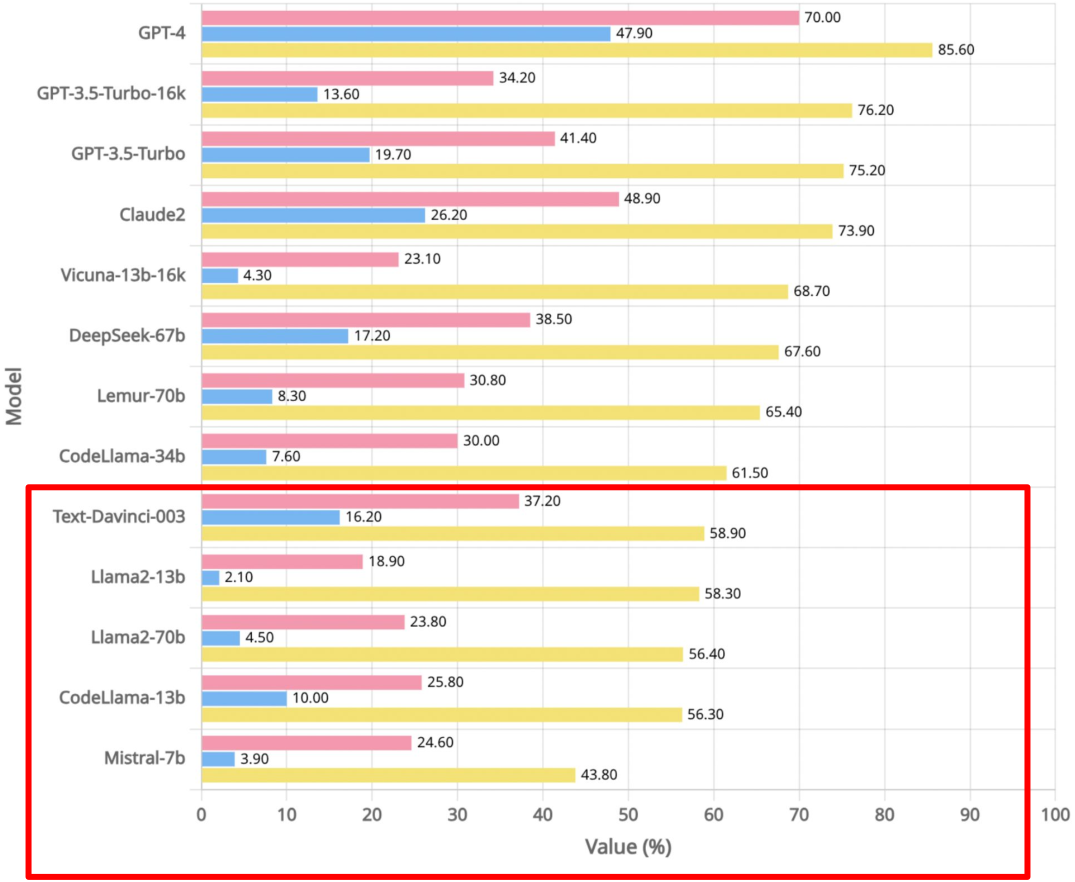
Available Actions:

Click [back to search]
Click [Next >]
Click [Tobfit Bands…]
Click [Veezoom Compatible …]
Click [Leather Bands …]

Click [Buy Now]

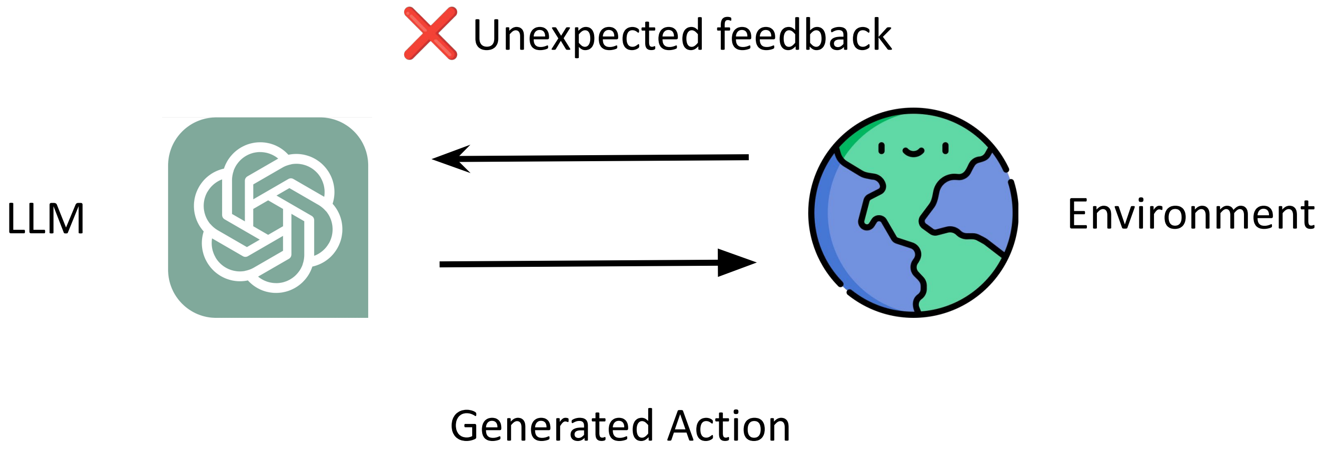
Grounding investigates whether LLM could map high-level plans to executable steps

Can LLM Perform Grounding Well ?

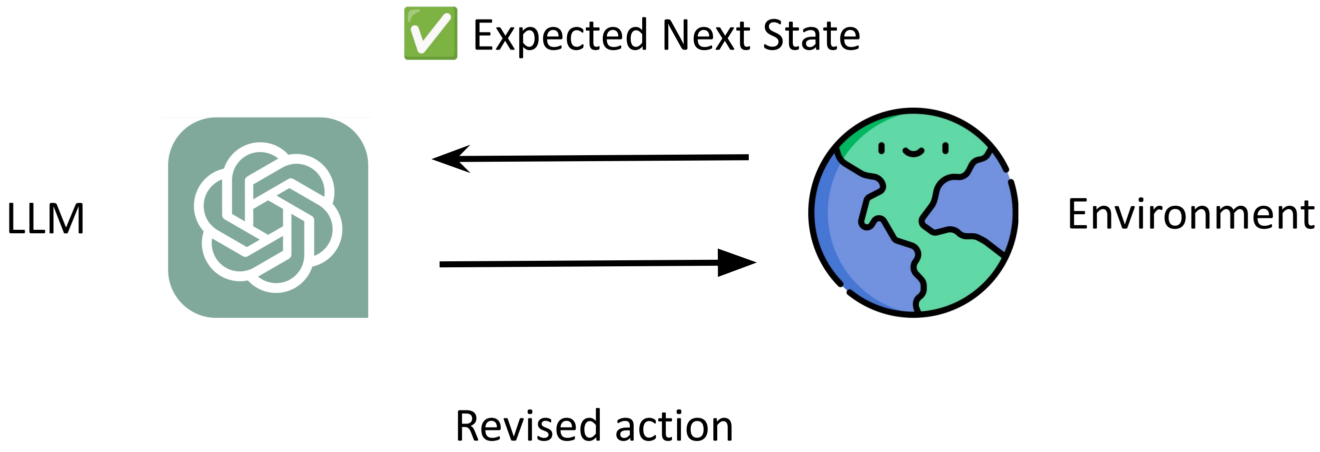


Grounding is crucial to the performance of LLM as agents.

LLM Reflection Ability



LLM Reflection Ability

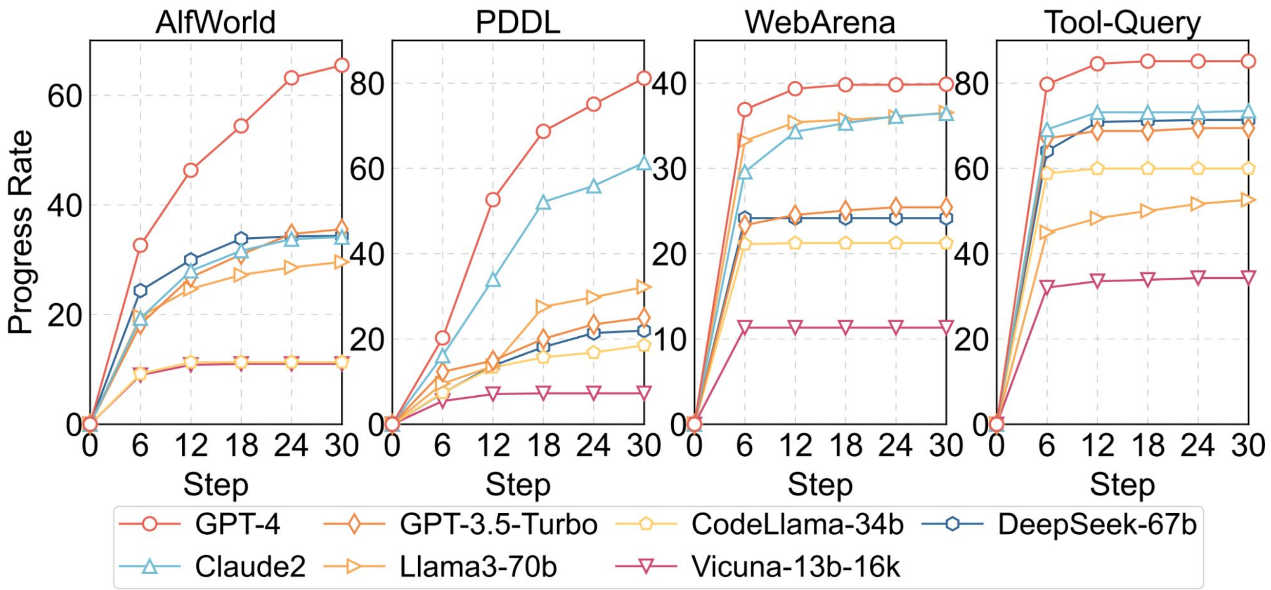


Reflection enables LLM to correct and improve its actions.

Long-Range Interaction - Reflection Challenge

Strong LLM as Agent consistently improves throughout long-term reflection.

Long-Range Interaction - Reflection Challenge



Most open-source models performance saturate after around 6 steps, while strong models like GPT-4 improves consistently through 30 steps.

LLM Planning Ability

Task: put a clean bowl in the fridge

explore and find bowl

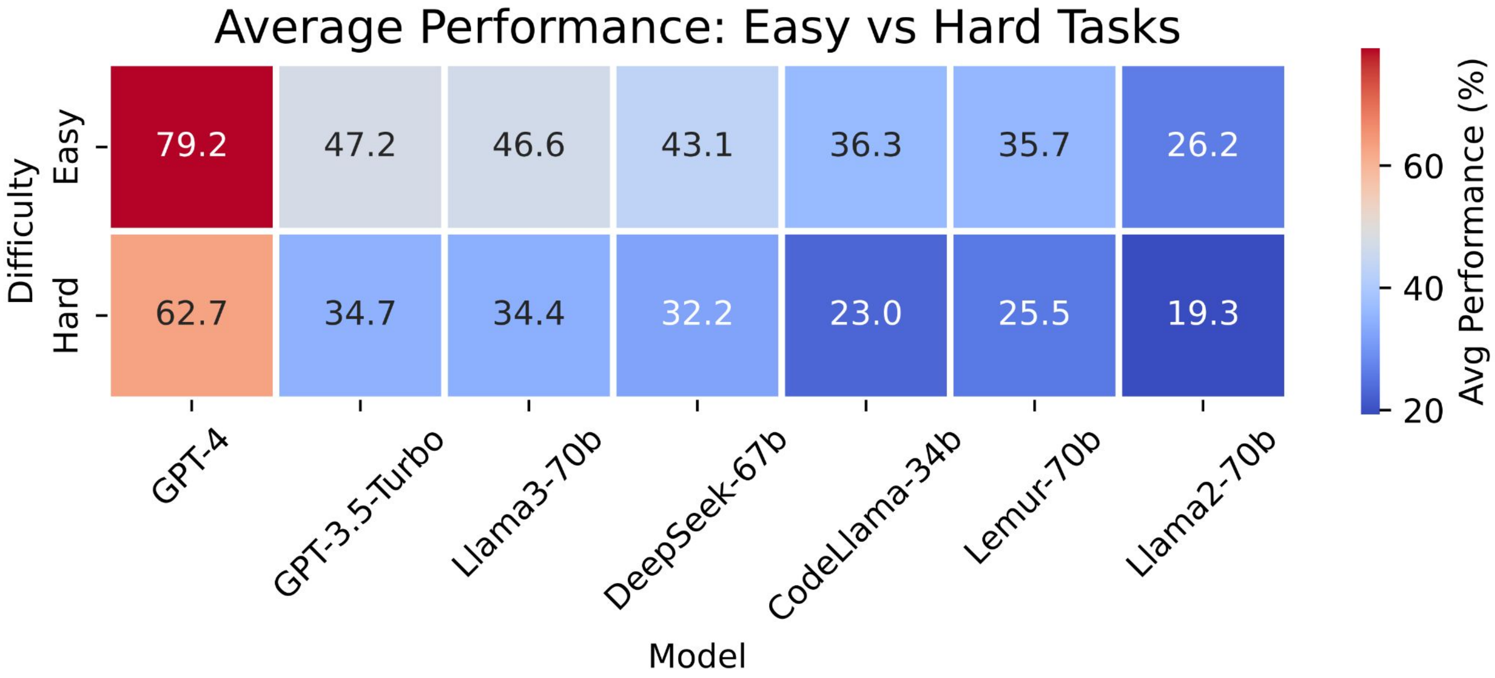
pickup and carry bowl

clean the bowl

place the bowl in fridge

Decompose a complex goal into several manageable subgoals.

Is LLM planning sensitive to task complexity？

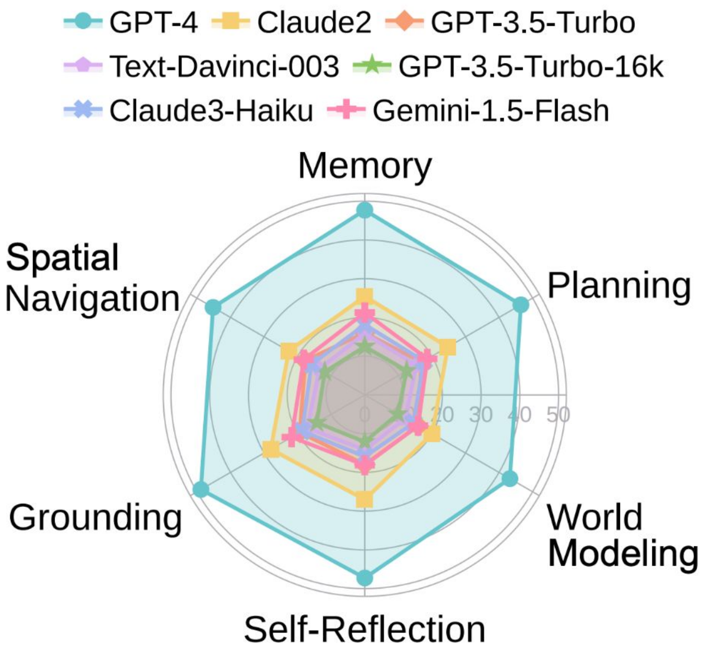


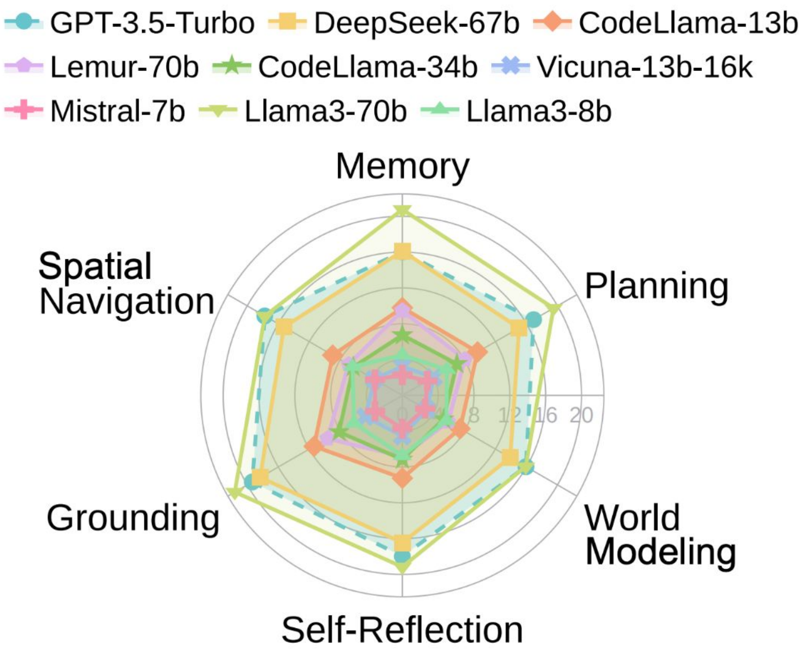
Easy - 1-3 subgoals

Hard: >=4 subgoals

All LLMs perform badly when task complexity scales, showing deficiency in long planning.

Agent Abilities are Multi-fold





Effective agent models exhibit balanced and robust capabilities in all areas.

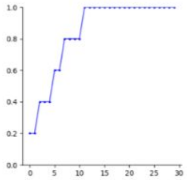
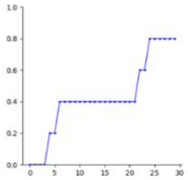
Convenient Analysis and Visualization

Analytical Visualization Board

Run AgentBoard benchmarking and visualize all analysis on WandB with a single command.

summary 4

runs.summary["scienceworld/predictions"]

	id	is_done	env.difficulty	env.goal	env.task_name	reward	grounding_accuracy	reward_wrt_step	trajectory
29	28	True	hard	Your task is to find a dove egg in the outside. First, focus on the thing. Then, move it to the red box in the bathroom.	find-living-thing	1	0.3333		**Step 0 Goal:** Your task is to find a dove egg in the outside. First, focus on the thing. Then, move it to the red box in the bathroom. **Step 0 Observation:** This room is called the greenhouse. In it, you see: the agent a substance called air a bee hive. The bee hive door is closed. a flower pot 1 (containing a apple tree in the adult stage, soil) ... **Step 0 Action:** go to outside **Step 0 Observation:** You move to the outside. **Step 0 Reward:** 0.2 **Step 1 Action:** look around
30	29	False	hard	Your task is to find a butterfly egg in the outside. First, focus on the thing. Then, move it to the green box in the bathroom.	find-living-thing	0.8	0.2333		**Step 0 Goal:** Your task is to find a butterfly egg in the outside. First, focus on the thing. Then, move it to the green box in the bathroom. **Step 0 Observation:** This room is called the hallway. In it, you see: the agent a substance called air a drawing You also see: ... **Step 0 Action:** go to outside **Step 0 Observation:** No known action matches that input. **Step 0 Reward:** 0.0 **Step 1 Action:** open door to outside **Step 1 Observation:** No known action matches that input.

Details of Each Problem Trajectory

29 - 30 of 90

Export as CSV Columns... Reset table

Takeaways: Fine-grained Evaluation + Analysis

- LLM Agents are complex systems that involve multiple abilities. Improving each ability is necessary for building good LLMs for agents.
- We need to perform analytical benchmarking of its various abilities to interpret whether the agent is good or why it works badly.
- Evaluating the process is as important as evaluating the final results !

Homepage

Code and Data