NYU

LLM Agent

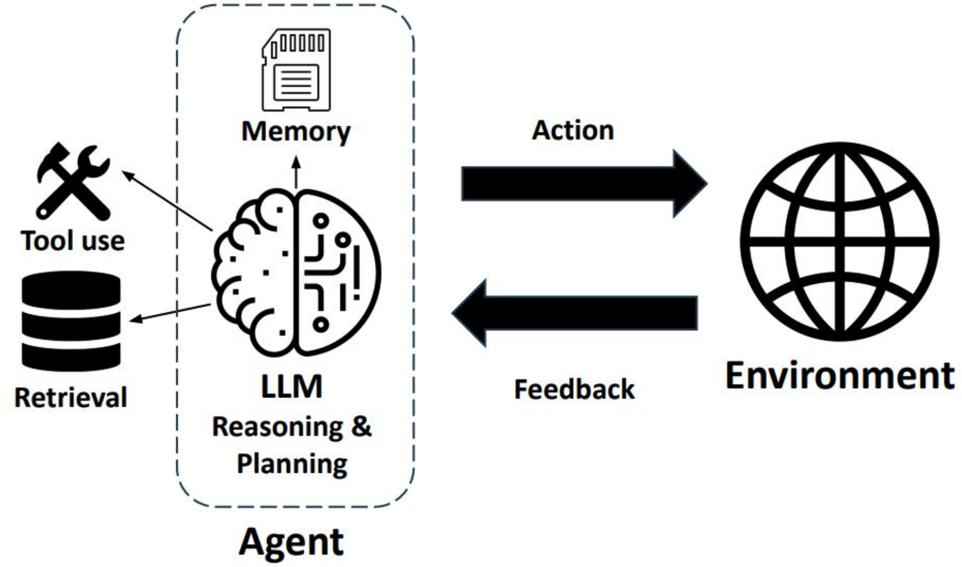
Advanced Topics in Embodied Learning and Vision

Ying Wang

2025.02.27

LLM Agents

"An agent is anything that can be viewed as perceiving its environment through sensors and acting upon that environment through actuators." — Russell & Norvig, AI: A Modern Approach (2020)



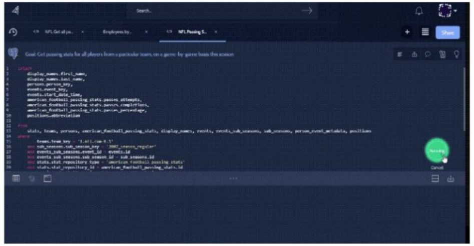
cr: https://rdi.berkeley.edu/adv-llm-agents/sp25

Why LLM Agents?

- 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

cr: https://rdi.berkeley.edu/adv-llm-agents/sp25

Applications

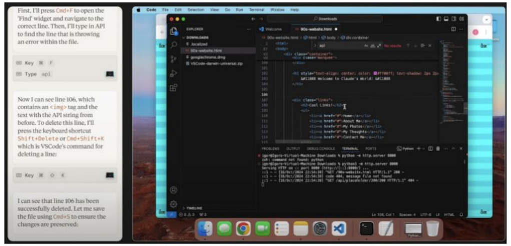


Goal: Get passing stats for all players from a particular team, on a game-by-game basis this season

```
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 = 'l.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, Google Jules…



First, I'll press Cmd+F to open the 'Find' widget and navigate to the correct line. Then, I'll type in API to find the line that is throwing an error within the file.

Now I can see line 106, which contains an <img> tag and the text with the API string from before. To delete this line, I'll press the keyboard shortcut Shift+Delete or Cmd+Shift+K which is VSCode's command for deleting a line:

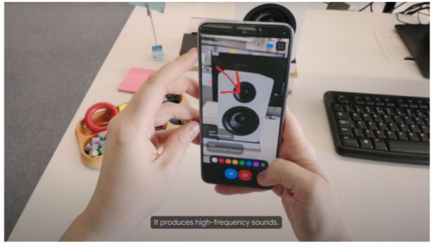
I can see that line 106 has been successfully deleted. Let me save the file using Cmd+S to ensure the changes are preserved:

Computer use

Anthropic Claude, Google Jarvis, OpenAI Operator

- Education
- Law
- Finance
- Healthcare
- Cybersecurity

  …



Personal assistant

Google Astra, OpenAI GPT-4o,…

Robotics

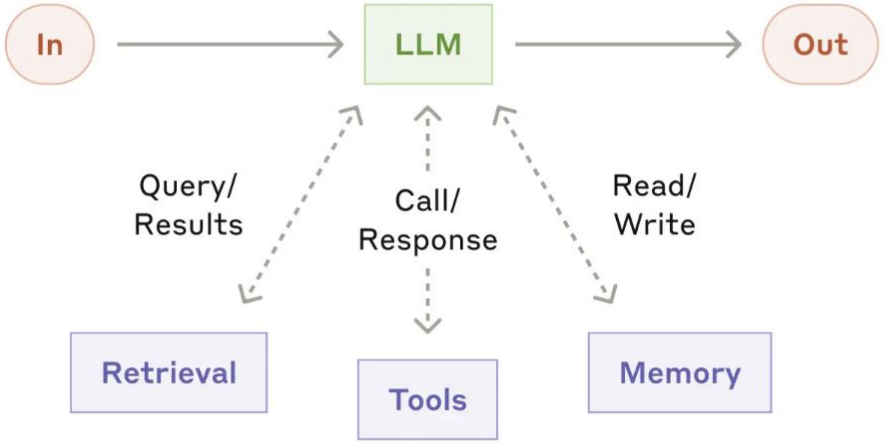
Figure AI, Tesla Optimus, NVIDIA GR00T…

cr: https://rdi.berkeley.edu/adv-llm-agents/sp25

Agenda

- Introduction
- Building block, workflows, agent
- Cognitive architectures for language agents
- LLM Agent Environments
  - Embodied Agent Interface
  - AgentBoard

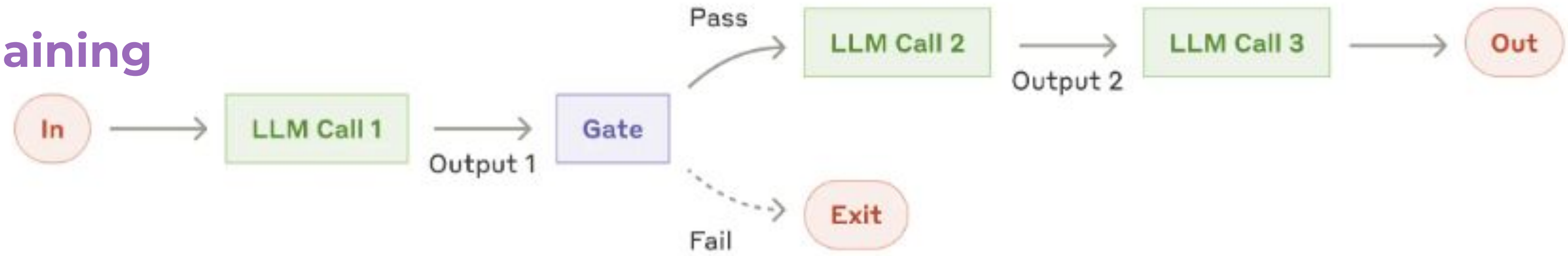
Building Block



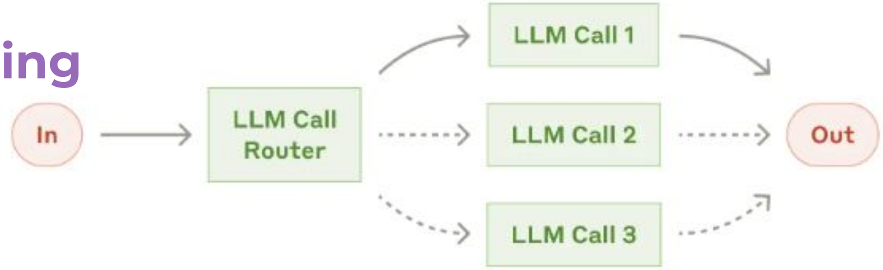
cr: https://www.anthropic.com/research/building-effective-agents

Workflows

chaining

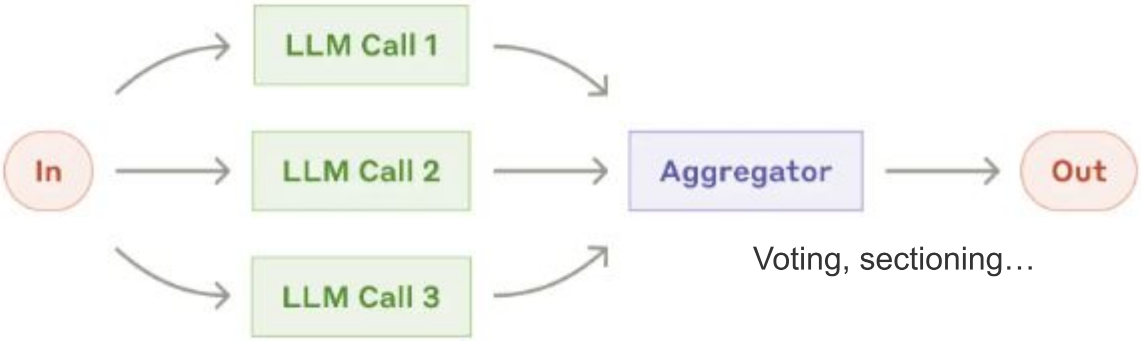


routing



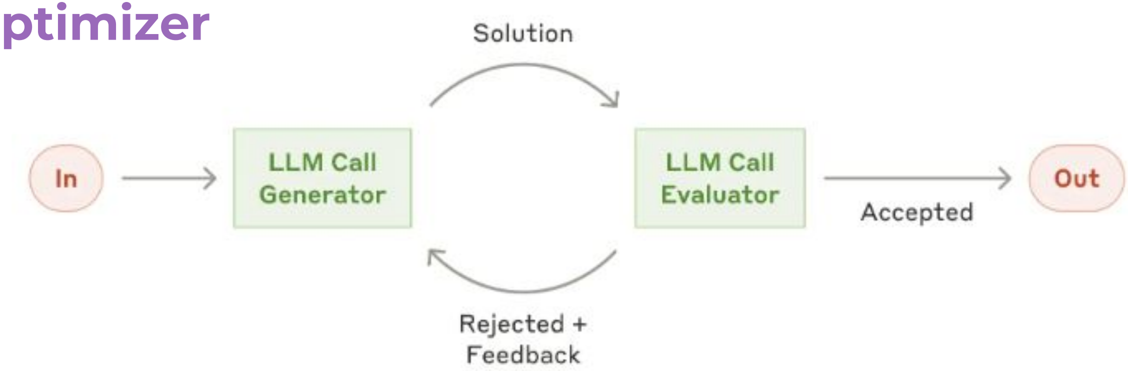
cr: https://www.anthropic.com/research/building-effective-agents

Parallelization



Voting, sectioning…

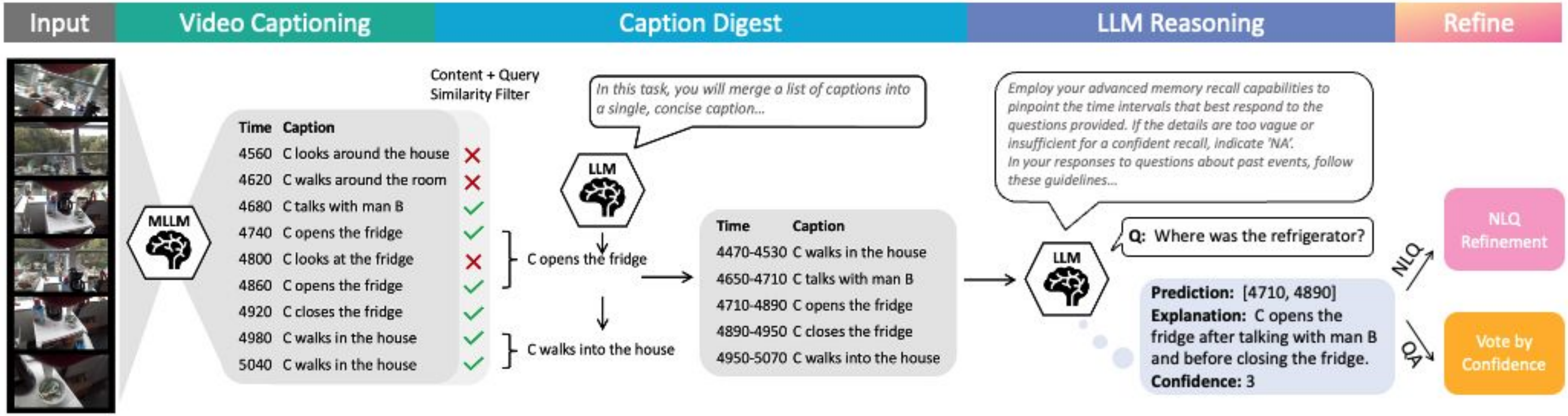
Evaluator-optimizer



cr: https://www.anthropic.com/research/building-effective-agents

Example: LifelongMemory

- You may need to combine different workflows to solve a challenging task!



Time	Caption	
4560	C looks around the house	✗
4620	C walks around the room	✗
4680	C talks with man B	✓
4740	C opens the fridge	✓
4800	C looks at the fridge	✗
4860	C opens the fridge	✓
4920	C closes the fridge	✓
4980	C walks in the house	✓
5040	C walks in the house	✓

C opens the fridge

C walks into the house

In this task, you will merge a list of captions into a single, concise caption...

Time	Caption
4470-4530	C walks in the house
4650-4710	C talks with man B
4710-4890	C opens the fridge
4890-4950	C closes the fridge
4950-5070	C walks into the house

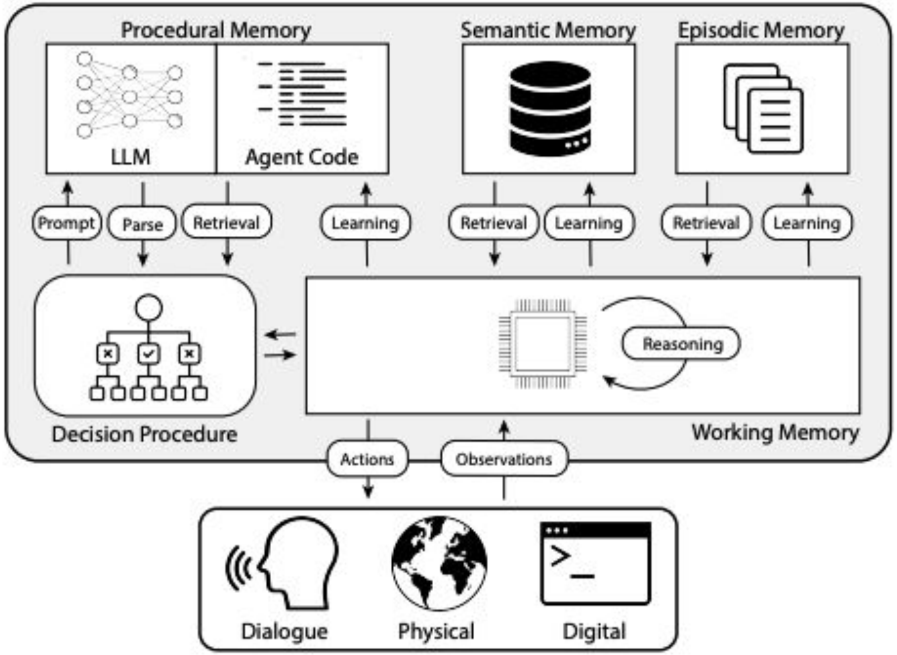
Employ your advanced memory recall capabilities to pinpoint the time intervals that best respond to the questions provided. If the details are too vague or insufficient for a confident recall, indicate 'NA'. In your responses to questions about past events, follow these guidelines...

Q: Where was the refrigerator?

Prediction: [4710, 4890]
Explanation: C opens the fridge after talking with man B and before closing the fridge.
Confidence: 3

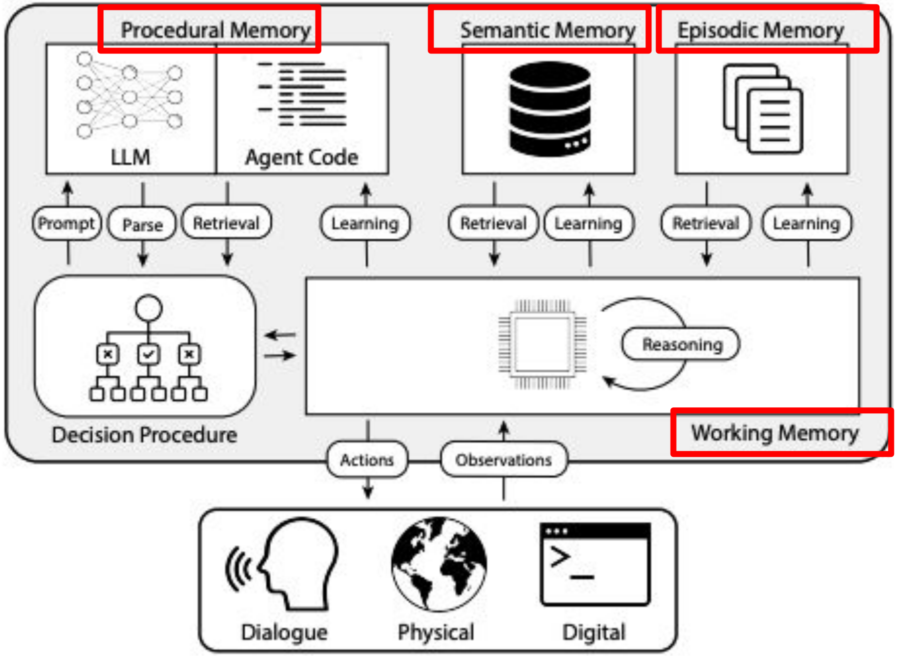
cr: https://agenticlearning.ai/lifelong-memory/

Cognitive architectures for language agents



cr: Cognitive Architectures for Language Agents (Sumers et al.)

CoALA: Memory



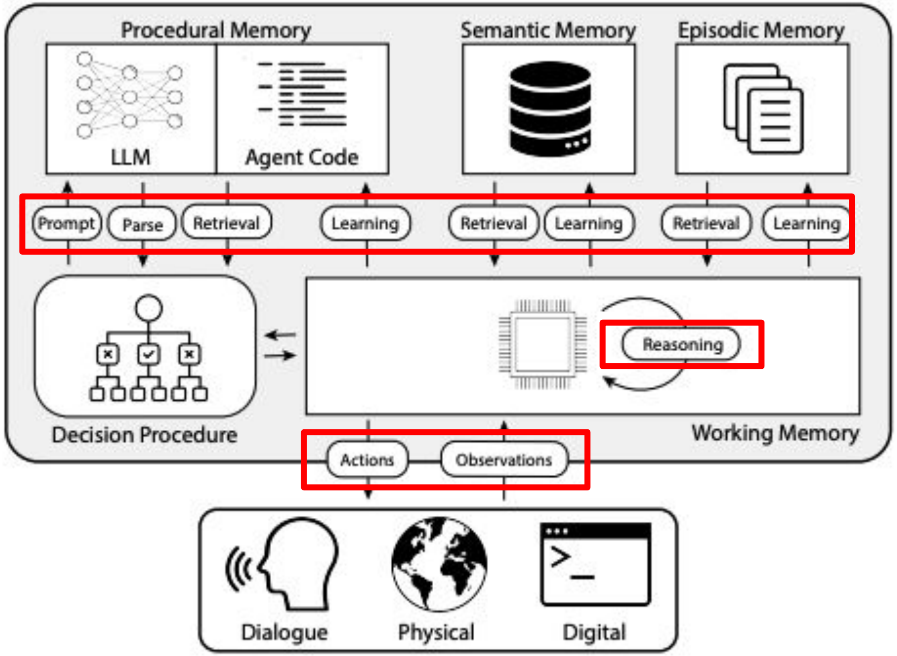
Episodic memory stores experience from earlier decision cycles.

Semantic memory stores an agent's knowledge about the world and itself.

Procedural memory: (i) implicit knowledge stored in the LLM weights; (ii) explicit knowledge written in the agent's code.

Working memory maintains active and readily available information as symbolic variables for the current decision cycle

CoALA: Action

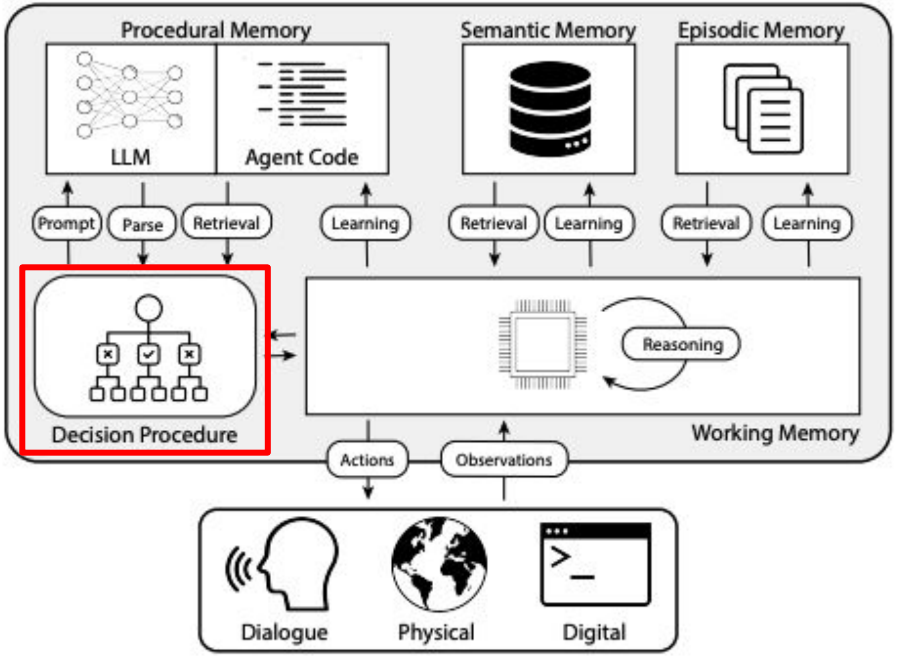


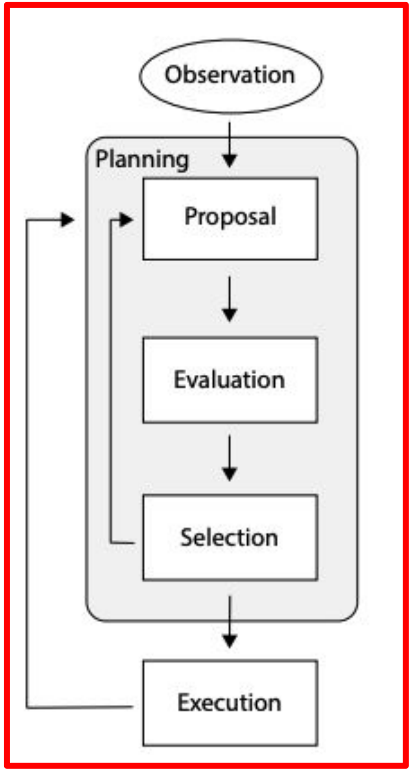
External actions interact with external environments. E.g., control a robot, communicate with a human, navigate a website

Internal actions interact with internal memories.

- retrieval (read from long-term memory)
- learning (write to long-term memory)
- reasoning (update the short-term working memory with LLM)

CoALA: Decision making





Agenda

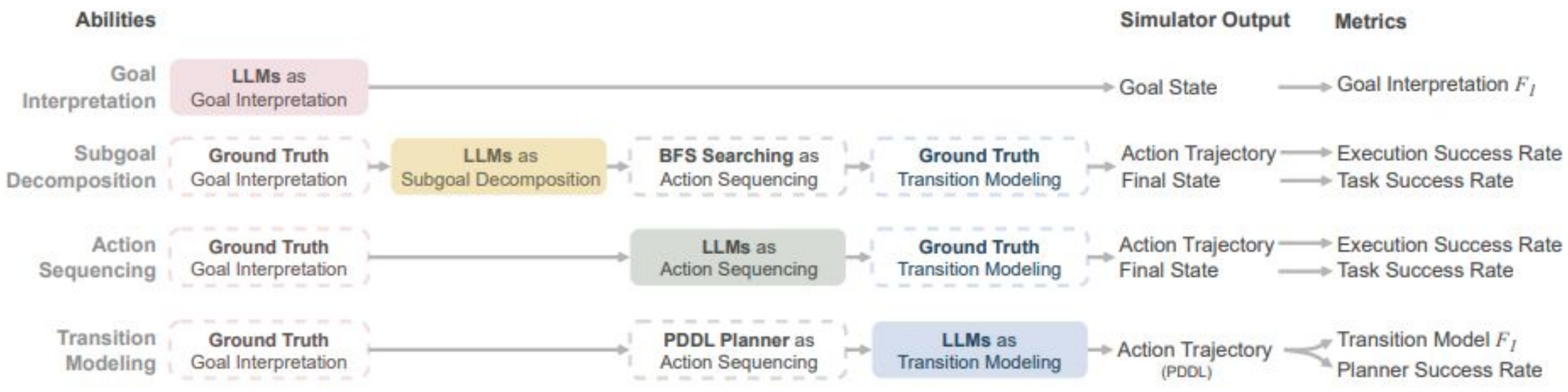
- Introduction
- Building block, workflows, agent
- Cognitive architectures for language agents
- LLM Agent Environments
  - Embodied Agent Interface: Benchmarking LLMs for Embodied Decision Making (NeurIPS 2024) https://embodied-agent-interface.github.io/
  - AgentBoard: An Analytical Evaluation Board of Multi-turn LLM Agents (NeurIPS 2024) https://hkust-nlp.github.io/agentboard/

Embodied Agent Interface: Benchmarking LLMs for Embodied Decision Making

Unlike existing evaluations rely solely on a final success rate, EAI breaks down the evaluation into

- **four modules** for decision making: goal interpretation, subgoal decomposition, action sequencing, and transition modeling
- a collection of **fine-grained metrics**, such as hallucination errors, affordance errors, various types of planning errors, etc.

Embodied Agent Interface


Abilities
Simulator Output
Metrics
Goal Interpretation
LLMs as Goal Interpretation
Goal State
Goal Interpretation F_1
Subgoal Decomposition
Ground Truth Goal Interpretation
LLMs as Subgoal Decomposition
BFS Searching as Action Sequencing
Ground Truth Transition Modeling
Action Trajectory
Final State
Execution Success Rate
Task Success Rate
Action Sequencing
Ground Truth Goal Interpretation
LLMs as Action Sequencing
Ground Truth Transition Modeling
Action Trajectory
Final State
Execution Success Rate
Task Success Rate
Transition Modeling
Ground Truth Goal Interpretation
PDDL Planner as Action Sequencing
LLMs as Transition Modeling
Action Trajectory (PDDL)
Transition Model F_1
Planner Success Rate

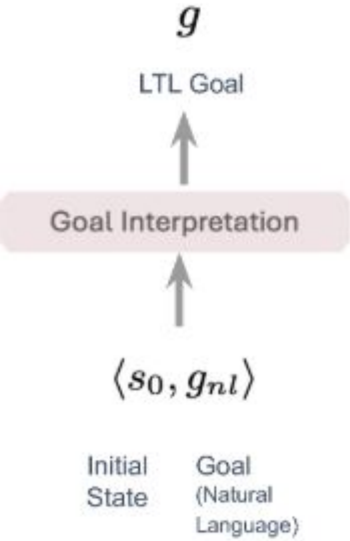
For each ability module, to provide a comprehensive evaluation for it, we isolate this single module to be handled by the LLMs while using existing data or tools for the other modules.

M1. Goal Interpretation

Ground the natural language instruction to the environment representations of objects, states, relations, and actions.

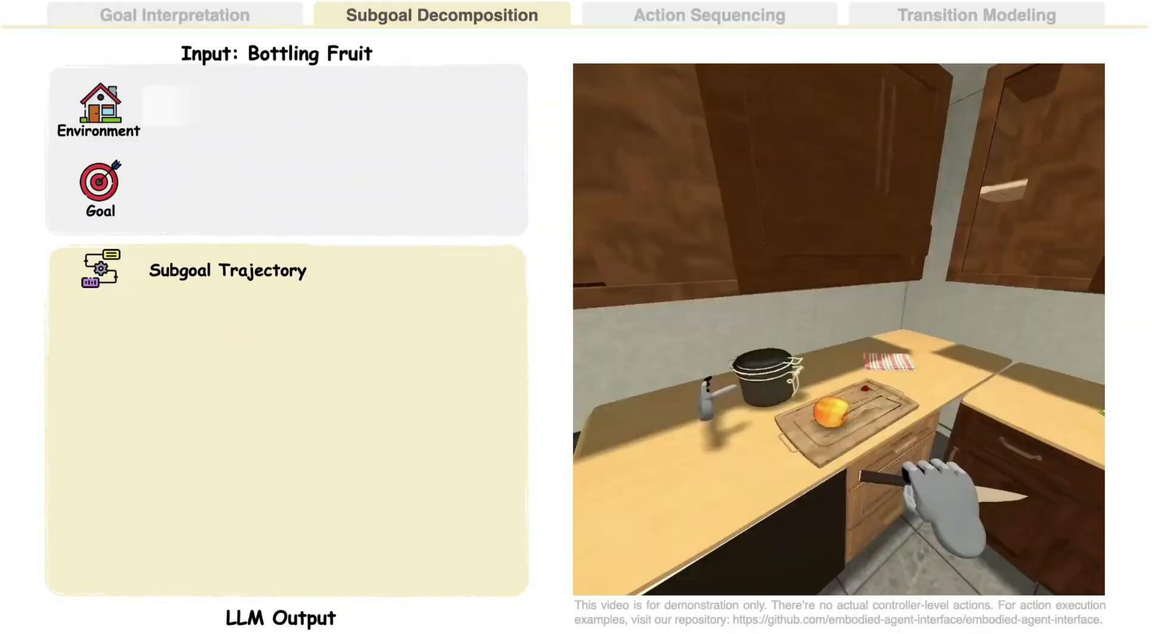
Goal Interpretation | Subgoal Decomposition | Action Sequencing | Transition Modeling

This video is for demonstration only. There're no actual controller-level actions. For action execution examples, visit our repository: https://github.com/embodied-agent-interface/embodied-agent-interface.



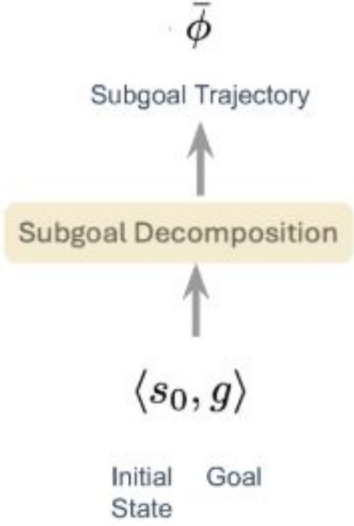
M2. Subgoal Decomposition

Subgoal Decomposition generates a sequence of states, where each state can be a set of objects and their states.



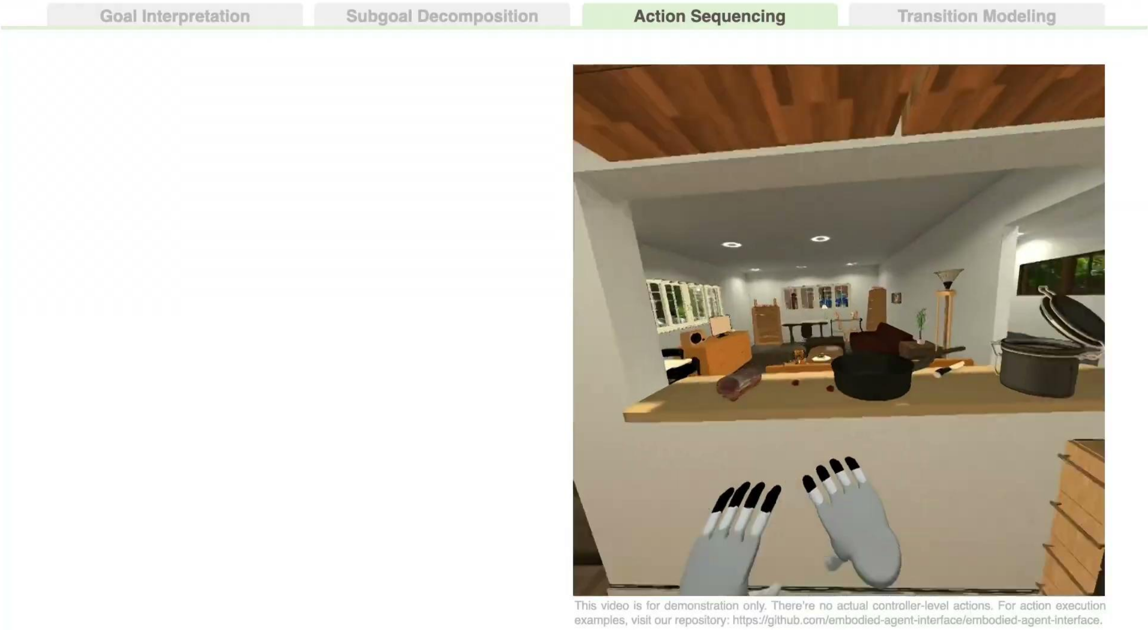
Input: Bottling Fruit

This video is for demonstration only. There're no actual controller-level actions. For action execution examples, visit our repository: https://github.com/embodied-agent-interface/embodied-agent-interface.

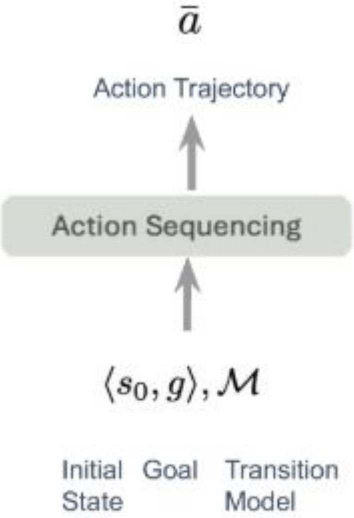


M3. Action Sequencing

Action Sequences are essential to achieve the state transitions identified in Subgoal Decomposition.



This video is for demonstration only. There're no actual controller-level actions. For action execution examples, visit our repository: https://github.com/embodied-agent-interface/embodied-agent-interface.

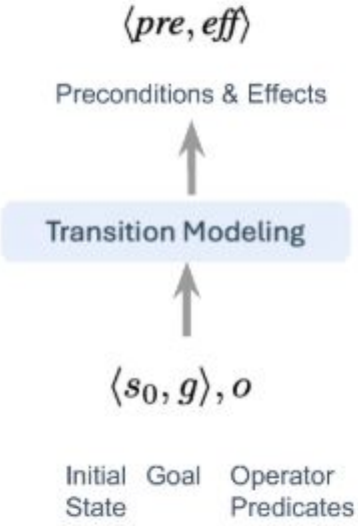


M4. Transition Modeling

Transition Modeling serves as the low-level controller to guide the simulator in performing state transitions from preconditions to post-effects.

Goal Interpretation | Subgoal Decomposition | Action Sequencing | **Transition Modeling**

This video is for demonstration only. There're no actual controller-level actions. For action execution examples, visit our repository: https://github.com/embodied-agent-interface/embodied-agent-interface.



$\langle pre, eff \rangle$

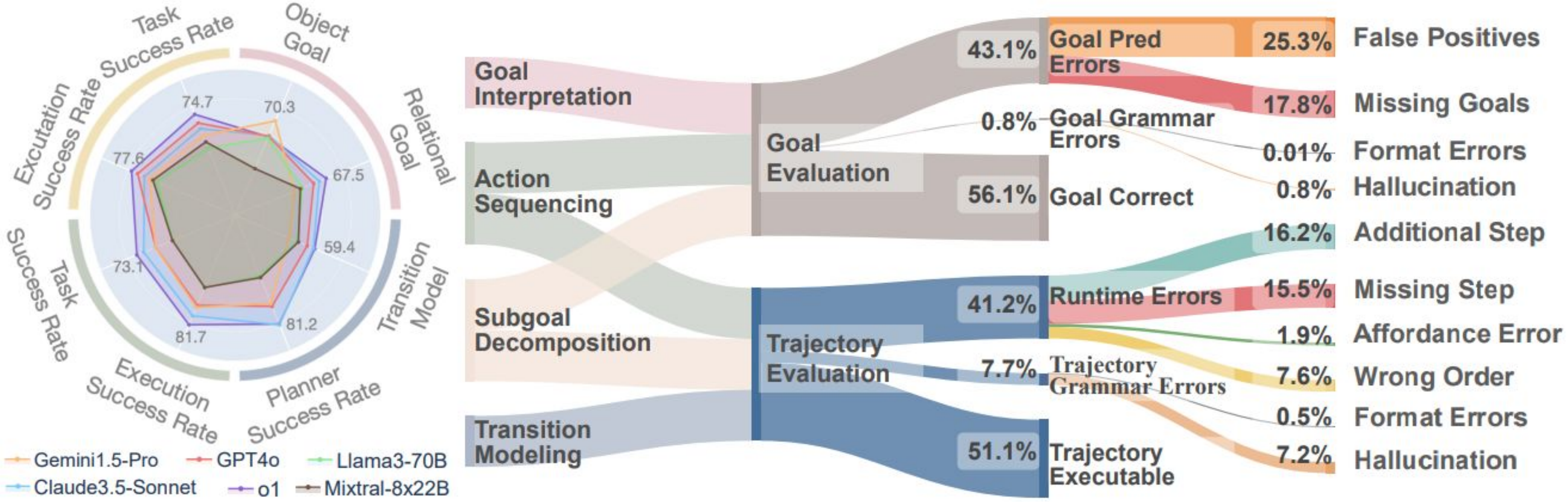
Preconditions & Effects

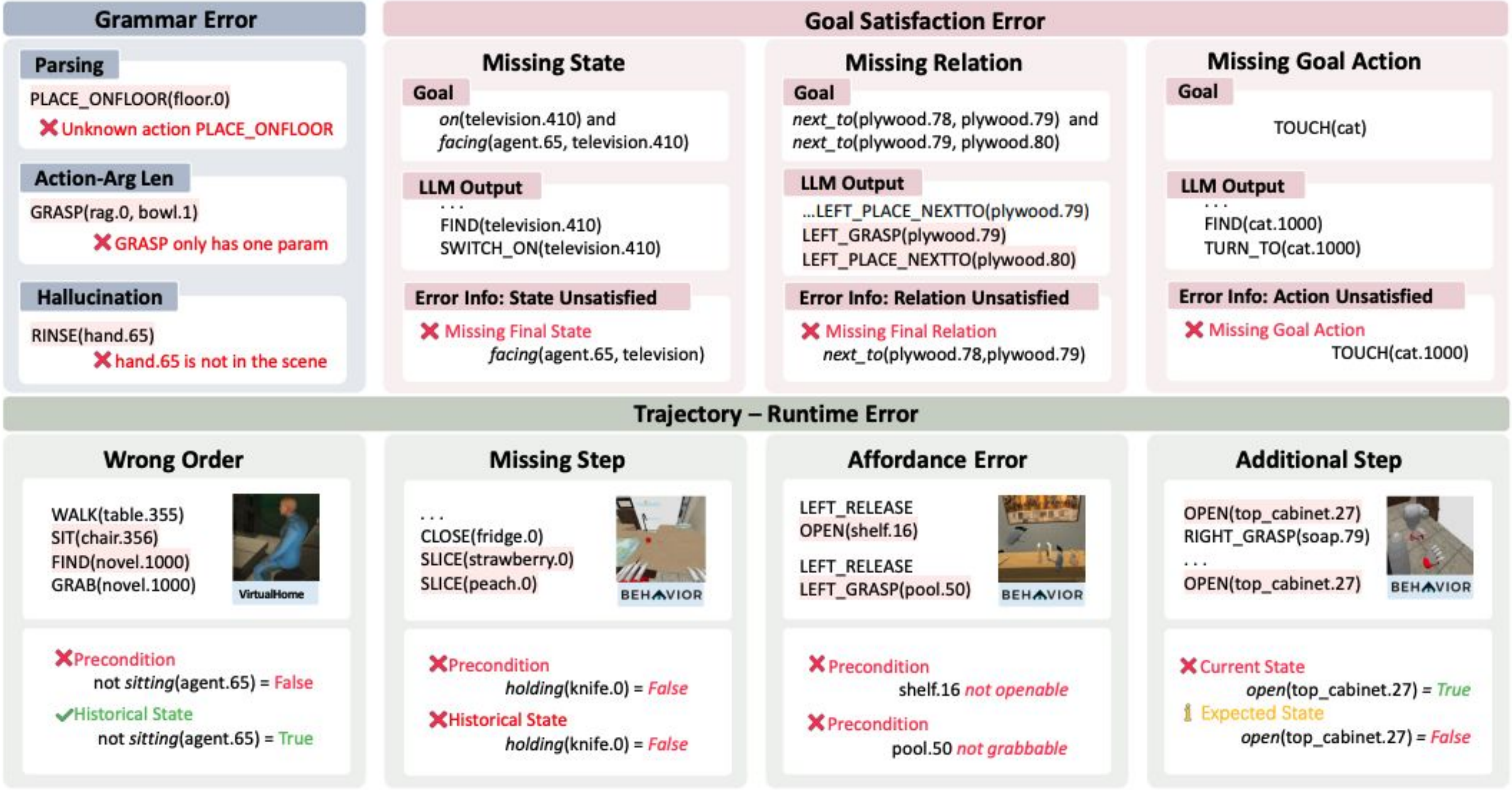
Transition Modeling

$\langle s_0, g \rangle, o$

Initial State | Goal | Operator Predicates

Evaluation


Task Success Rate
Object Goal
Relational Goal
Transition Model
Planner Success Rate
Execution Success Rate
Task Success Rate
Excutation Success Rate
74.7
70.3
67.5
59.4
81.2
81.7
73.1
77.6
Gemini1.5-Pro
GPT4o
Llama3-70B
Claude3.5-Sonnet
o1
Mixtral-8x22B
Goal Interpretation
Action Sequencing
Subgoal Decomposition
Transition Modeling
Goal Evaluation
Trajectory Evaluation
43.1%
Goal Pred Errors
0.8%
Goal Grammar Errors
56.1%
Goal Correct
41.2%
Runtime Errors
7.7%
Trajectory Grammar Errors
51.1%
Trajectory Executable
25.3%
False Positives
17.8%
Missing Goals
0.01%
Format Errors
0.8%
Hallucination
16.2%
Additional Step
15.5%
Missing Step
1.9%
Affordance Error
7.6%
Wrong Order
0.5%
Format Errors
7.2%
Hallucination



Grammar Error

Parsing

PLACE_ONFLOOR(floor.0)

✗ Unknown action PLACE_ONFLOOR

Action-Arg Len

GRASP(rag.0, bowl.1)

✗ GRASP only has one param

Hallucination

RINSE(hand.65)

✗ hand.65 is not in the scene

Goal Satisfaction Error

Missing State

Goal

on(television.410) and
facing(agent.65, television.410)

LLM Output

```
. . .
FIND(television.410)
SWITCH_ON(television.410)
```

Error Info: State Unsatisfied

✗ Missing Final State
facing(agent.65, television)

Missing Relation

Goal

next_to(plywood.78, plywood.79) and
next_to(plywood.79, plywood.80)

LLM Output

```
...LEFT_PLACE_NEXTTO(plywood.79)
LEFT_GRASP(plywood.79)
LEFT_PLACE_NEXTTO(plywood.80)
```

Error Info: Relation Unsatisfied

✗ Missing Final Relation
next_to(plywood.78,plywood.79)

Missing Goal Action

Goal

TOUCH(cat)

LLM Output

```
. . .
FIND(cat.1000)
TURN_TO(cat.1000)
```

Error Info: Action Unsatisfied

✗ Missing Goal Action
TOUCH(cat.1000)

Trajectory – Runtime Error

Wrong Order

```
WALK(table.355)
SIT(chair.356)
FIND(novel.1000)
GRAB(novel.1000)
```

✗ Precondition
not *sitting*(agent.65) = False

✓ Historical State
not *sitting*(agent.65) = True

Missing Step

```
. . .
CLOSE(fridge.0)
SLICE(strawberry.0)
SLICE(peach.0)
```

✗ Precondition
holding(knife.0) = *False*

✗ Historical State
holding(knife.0) = *False*

Affordance Error

```
LEFT_RELEASE
OPEN(shelf.16)

LEFT_RELEASE
LEFT_GRASP(pool.50)
```

✗ Precondition
shelf.16 *not openable*

✗ Precondition
pool.50 *not grabbable*

Additional Step

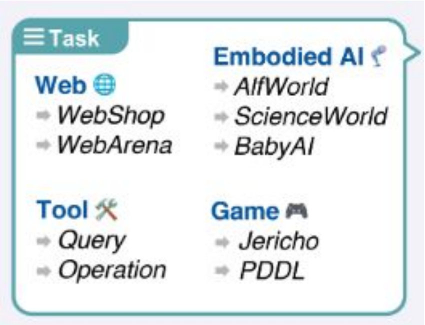
```
OPEN(top_cabinet.27)
RIGHT_GRASP(soap.79)
. . .
OPEN(top_cabinet.27)
```

✗ Current State
open(top_cabinet.27) = *True*

ℹ Expected State
open(top_cabinet.27) = *False*

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

- **Task diversity** is necessary to cover various agent tasks such as embodied, web, and tool agents.
- **Multi-round** interaction is critical to mimic realistic scenarios.
- Evaluating agents in **partially-observable** environments, where they must actively explore to understand their surroundings, is essential for practical assessments.



AlfWorld

Interactive text environment that require agents to explore surroundings and perform commonsense tasks like “put two soap bars in garbagecan".

Embodied
RotateRight
LookDown
MoveAhead
Pickup

Goal: heat some apple and put it in fridge.

> check valid actions
Choose an action from these valid actions: go to cabinet 1, go to cabinet 2, go to cabinet 3, …

> go to fridge 1
The fridge 1 is closed.

> open fridge 1
You open the fridge 1. The fridge 1 is open. In it, you see a apple 1, a bowl 3, a cup 2, a cup 1, a egg 3, a lettuce 1, a potato 2, a potato 1, and a tomato 1. (reward: 0.25)

…

ScienceWorld

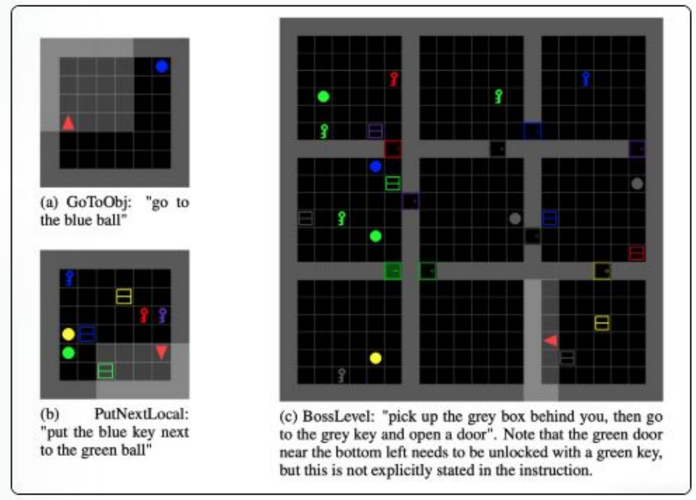
Interactive text environment testing scientific commonsense, e.g.“measure the melting point of the orange juice".

FOUNDRY
WORKSHOP
ART
LIVING
GREEN HOUSE
KITCHEN
BATH
BEDROOM
OUTSIDE

```
Goal: Your task is to find the animal with the longest life span.

...

> go to outside
You move to the outside. (reward: 0.3333333333333333)

> look around
This outside location is called the outside. Here you see: the
agent a substance called air an axe a baby brown bear ...

> focus on baby brown bear
You focus on the baby brown bear. (reward:
0.6666666666666666)
…
```

BabyAI

Interactive 20x20 grid environment where agents navigate and interact with objects within a limited sight range.



(a) GoToObj: "go to the blue ball"

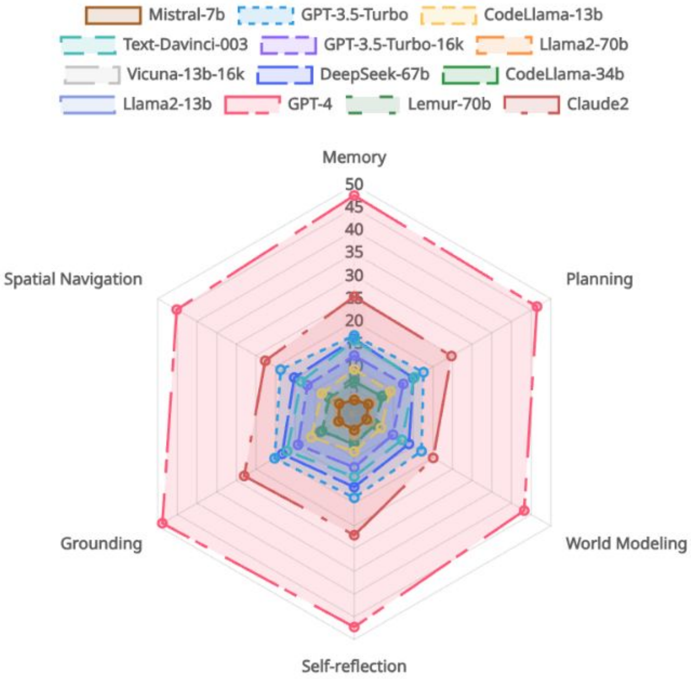
(b) PutNextLocal: "put the blue key next to the green ball"

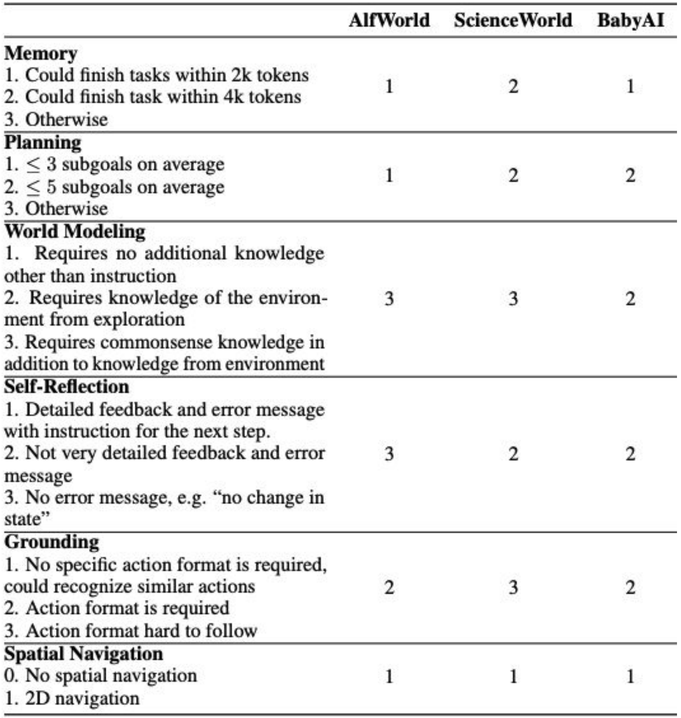
(c) BossLevel: "pick up the grey box behind you, then go to the grey key and open a door". Note that the green door near the bottom left needs to be unlocked with a green key, but this is not explicitly stated in the instruction.

Goal: Open the red door, and open the blue door

```
> check valid actions
You can take the following actions: turn left, turn right, move forward, toggle and go through blue closed door 1, go to blue closed door 1, check available actions

> go to blue closed door 1
In front of you in this room, you can see several objects: There is a blue closed door 1 right in front of you 1 steps away. The room has walls around you. You are not carrying anything. (reward: 0.25)

> toggle and go through blue door 1
The action is not recognized. Please check valid actions.
…
```

Evaluation





	AlfWorld	ScienceWorld	BabyAI
Memory 1. Could finish tasks within 2k tokens 2. Could finish task within 4k tokens 3. Otherwise	1	2	1
Planning 1. ≤ 3 subgoals on average 2. ≤ 5 subgoals on average 3. Otherwise	1	2	2
World Modeling 1. Requires no additional knowledge other than instruction 2. Requires knowledge of the environment from exploration 3. Requires commonsense knowledge in addition to knowledge from environment	3	3	2
Self-Reflection 1. Detailed feedback and error message with instruction for the next step. 2. Not very detailed feedback and error message 3. No error message, e.g. "no change in state"	3	2	2
Grounding 1. No specific action format is required, could recognize similar actions 2. Action format is required 3. Action format hard to follow	2	3	2
Spatial Navigation 0. No spatial navigation 1. 2D navigation	1	1	1