

AgentBoard

AgentBoard emphasizes **analytical evaluation** for Large Language Models (LLMs) as generalist agents to perceive and act within various environments. It outlines four principles for constructing a benchmark to evaluate LLMs as generalist agents:

1. **Task Diversity:** AgentBoard incorporates 9 distinct tasks to comprehensively understand the generalist ability of LLM agents, which is built upon LLM's extensive knowledge base and exceptional scenario comprehension.
2. **Multi-round Intercation:** AgentBoard provides multi-round interaction between agents and environment, which is necessary to reflect the evolutionary nature of human intelligence, which continuously receives information and adapts towards the environment.
3. **Partially-Observable Environments:** In AgentBoard, the complete state of the environment is not available to the agent, which assesses agent world modeling ability as additional knowledge needs to be acquired through online exploration.
4. **Analytical Evaluation:** AgentBoard is a systematic evaluation platform: it includes a user-friendly script to construct goal-oriented reflex agents for a range of models, and features a panel for visualizing and interpreting results across multiple dimensions of agent proficiency, including *fine-grained progress rates, grounding accuracy, performance breakdown for hard and easy examples, long-range in- teractions, detailed performance across various sub-skills, and trajectory with friendly visualization*

Dataset Link

[Github Repo](#)

[Official Website](#)

[Huggingface Dataset](#)

Data Card Author(s)

Chang Ma

Dataset Owners

Team(s)

AgentBoard Team

Contact Detail(s)

Affiliation: The Hong Kong University of Science and Technology

Group Email: llagentboard@gmail.com

Website: <https://hkust-nlp.github.io/agentboard/>

Authors:

Chang Ma, Junlei Zhang, Zhihao Zhu, Cheng Yang, Yujiu Yang, Yaohui Jin, Zhenzhong Lan, Lingpeng Kong, Junxian He

Funding Sources

Institution(s)

The Hong Kong University of Science and Technology

Funding or Grant Summary(ies)

N/A

Dataset Overview

Data Subject(s)

- Subgoal annotation for agent tasks
- test data for all tasks

Data Snapshots

Category	Data
Size of Dataset	5.7 GB
Number of Instances	1012
Number of Fields	6
Labeled Classes	N/A
Number of Labels	N/A
Average Labels Per Instance	N/A
Algorithmic Labels	N/A

Content Description

Each datapoint in the dataset contains one goal with corresponding subgoals as well as labeled difficulty. We keep all labels public to enhance open study on our benchmark.

Sensitivity of Data

No

Dataset Version and Maintenance Maintenance Status

Actively Maintained - No new versions will be made available, but this dataset will be actively maintained, including but not limited to updates to the data.

Version Details Current Version: 1.0

Last Updated: 02/2023

Release Date: 02/2023

Maintenance Plan: We will fix any error and provide help in github issues.

Versioning: N/A. AgentBoard is a static dataset. Minor releases correspond to any errors fixed in the dataset.

Feedback: For feedback, reach out to llagentboard@gmail.com or open an issue on our Github Repo

Example of Data Points

Primary Data Modality

Text Data

Sampling of Data Points

Explore AgentBoard on our official website.

Data Fields

Fields	Description
Task	Name of the task for the problem.
ID	ID of the problem in the task.
Goal	Goal for the agent task.
Subgoals	Decomposed subgoals, used in progress rate calculation.
Difficulty	Annotated difficulty level, used in easy/hard metrics calculation.
Additional Info	Other additional information, including descriptions or original ID for each task.

Typical Data Point

Below is an example from PDDL:

```
{
  "id": 0,
  "task": "pddl",
  "goal": "The goal is to satisfy the following conditions: ball1 is at roomb. ball2 is at roomb. ball3 is at roomb. ball4 is at roomb. ",
  "subgoals": [
    "ball1 is at roomb.",
    "ball2 is at roomb.",
    "ball3 is at roomb.",
    "ball4 is at roomb."
  ],
  "difficulty": "easy",
  "additional_info": {
    "subtask": "gripper"
  }
}
```

Atypical Data Point

The dataset does not contain atypical data points as far as we know.

Motivations & Intentions

Purpose(s)

Research

Domain(s) of Application

Machine Learning, Natural Language Processing, Autonomous Agent

Motivating Factor(s)

- Provide **analytical evaluation** for Large Language Models (LLMs) as generalist agents to perceive and act within various environments.
- Perform fine-grained evaluation and distinguish any minor improvement in open-source LLMs.

Intended Use

Dataset Use(s)

Safe for research use

Suitable Use Case(s)

Search for better hyperparameter and datasets during LLM training. For example, determine the optimal training strategy for an agent LLM.

Assess agentic abilities of LLMs.

Unsuitable Use Case(s)

The dataset is created for model evaluation. It is not intended to be used as pre-training data.

Citation Guidelines

BiBTeX:

```
@misc{ma2024agentboard,
  title={AgentBoard: An Analytical Evaluation Board of Multi-turn LLM Agents},
  author={Chang Ma and Junlei Zhang and Zhihao Zhu and Cheng Yang and Yujiu Yang and Yaohui Jin and Zhenzhong Lan and Lingpeng Kong and Junxian He},
  year={2024},
  eprint={2401.13178},
  archivePrefix={arXiv},
  primaryClass={cs.CL}
}
```

Provenance

Collection Method(s) Used

- Manually annotated
- Adapted from other environments.

Source Description(s)

We adapt from sources including:

- [Alfworld](#)
- [BabyAI](#)
- [Jericho](#)
- [PDDL Gym](#)
- [Scienceworld](#)
- [Webshop](#)
- [Webareana](#)

Data Processing

- We make sure all environments have a text-based interaction interface.
- We manually annotate subgoals for each task, and verify the subgoals through a strict process.

Use in ML or AI Systems Dataset Use(s)

Testing

Validation

Development or Production Use

Usage Guideline(s)

Please visit our [Github Repo](#) for detailed information.

Licenses

Code License Apache-2.0

The AgentBoard codebase is licensed under a [Apache-2.0 License](#).