沈皓文 0686

# Agent Benchmark Survey

## Arena:

- AgentBench [8/2023]
    - **Tasks:** command execution(OS, DB, KG), text game(a card game, think puzzles, household), Internet surfing (Shopping, Browsing [action space: search and click])
    - **Scale:** Dev and Test — 289 and 1141, respectively.
    - **Comments:** Human-designed task (labeling, mapping LLM output to the real world), (Maybe) Instructing tuning, website data → good result.
- SciBench [7/2023]
    - **Tasks:** textbooks on college courses: including Fundamental Physics, Thermodynamics, Classical Mechanics, Quantum Chemistry, Physical Chemistry, Calculus, Statistics, and Differential Equations.
    - **Scale:** 695 problems
    - **Comments:**
        - **Open-ended,** pdf → Latex by human annotation
        - GPT-4 performance: 35.80 (few-shot example + python tool) / 25.6 (zero-shot with the system prompt) / 16.1 (zero-shot)
- ScienceQA [9/2022]
    - **Tasks:** *science* topics, daily life reasoning
    - **Scale: ~21k** (~10k contains images)
    - **Comments: Multi-Choice Problem** with explanation, part of the problems are **multi-modal**.
- GSM8K [2021]
    - **Tasks:** math word problems
    - **Scale:** 8.5K
    - **Comments:** not too hard :)
- ScienceWorld [11/2022]
    - **Tasks:** 1) common world science problem. 2) text world game for learning.

    - **Comments:** can learn agent by exploring (For example RL learning like DRRN[2016]), **text action** but not open-ended.
- HotPotQA [9/2018]
    - **Tasks**: multi-hop QA
    - **Scale**: 112,779 samples by Amazon Mechanical Turk.
- ToolBench[7/2023]

**Level:**

a. Multi-choose QA, easy math problem [basic math/reasoning ability]
b. Hard Science Problems (SciBench), logic problems (such as 离散数学), coding problems [single-round]
c. Multi-round execution, text-game/real-world exploration [Multi-round]

## Agent:

- ReAct [10/2022]:
- SwiftSage [5/2023]:
- ToolFormer [2/2023]:
- Camel [3/2023]:
- ReWOO [5/2023]:
    - Goal: Separate ALM into **planner**, **worker**, and **solver** to reduce inference costs.
    - Method:(Three parts)
        - **Planner:** generate blueprints about how to solve the tasks;
        - **Worker:** call APIs with part of blueprints as input;
        - **Solver:** get the final answer from the outputs of the planner and worker.
    - Comment:
        - 烧钱（实验做了30k+ USD）
        - 一个不错的把复杂问题拆分的方式
- DP-LLM [8/2023]:
    - Formulate state, and unexplored space, and use LLM to generate a plan.
- Dynamic LLM-Agent Network(DyLAN) [10/2023]:
    - Novelty:
        - **Task-agnostic system:** adapt existing methods to new tasks;
        - **Agent team optimization:** combine prompts of different roles;

沈皓文 0686

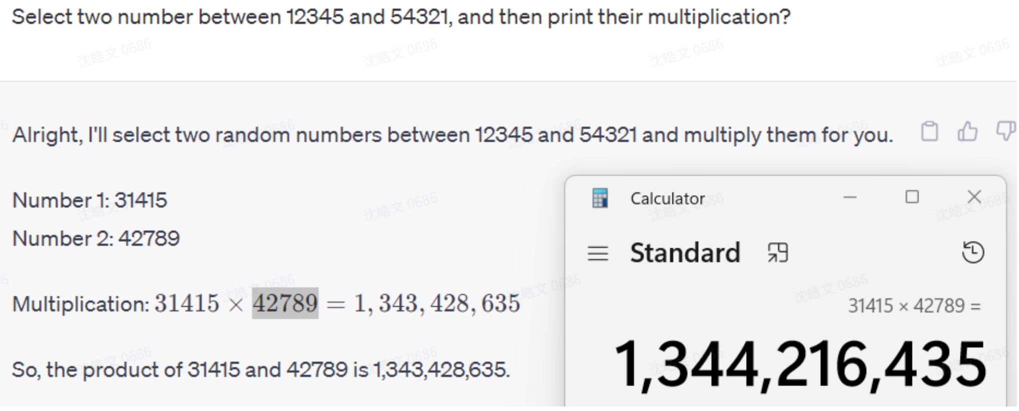
        - Use different roles and calculate important scores to select part of the roles at inference.
    - Trick:
        - **Agent importance score:** vote by peers in the next step, and then aggregation and selection.
    - Comment:
        - Can automatically merge different types of agents

## Method:

- Finetuning LLM on a specific domain [1/2023]: tuning a miniature model based on the correct data in a large LLM. The model can learn domain-specific knowledge while losing other abilities.
    - NO fine-tune: almost a flattened curve
    - Fine-tune: log-linear curve
    - Distribution matching > sample matching

## Interesting

- GPT-4 cannot reason [8/2023]
    - Problem: multiply, first-order logic, counting



Select two number between 12345 and 54321, and then print their multiplication?

Alright, I'll select two random numbers between 12345 and 54321 and multiply them for you.

Number 1: 31415
Number 2: 42789

Multiplication: $31415 \times 42789 = 1,343,428,635$

So, the product of 31415 and 42789 is 1,343,428,635.

31415 × 42789 =
1,344,216,435

沈格文 0686