ACL 2025 VIENNA

Enhancing Open-Domain Task-Solving Capability of LLMs via Autonomous Tool Integration from GitHub

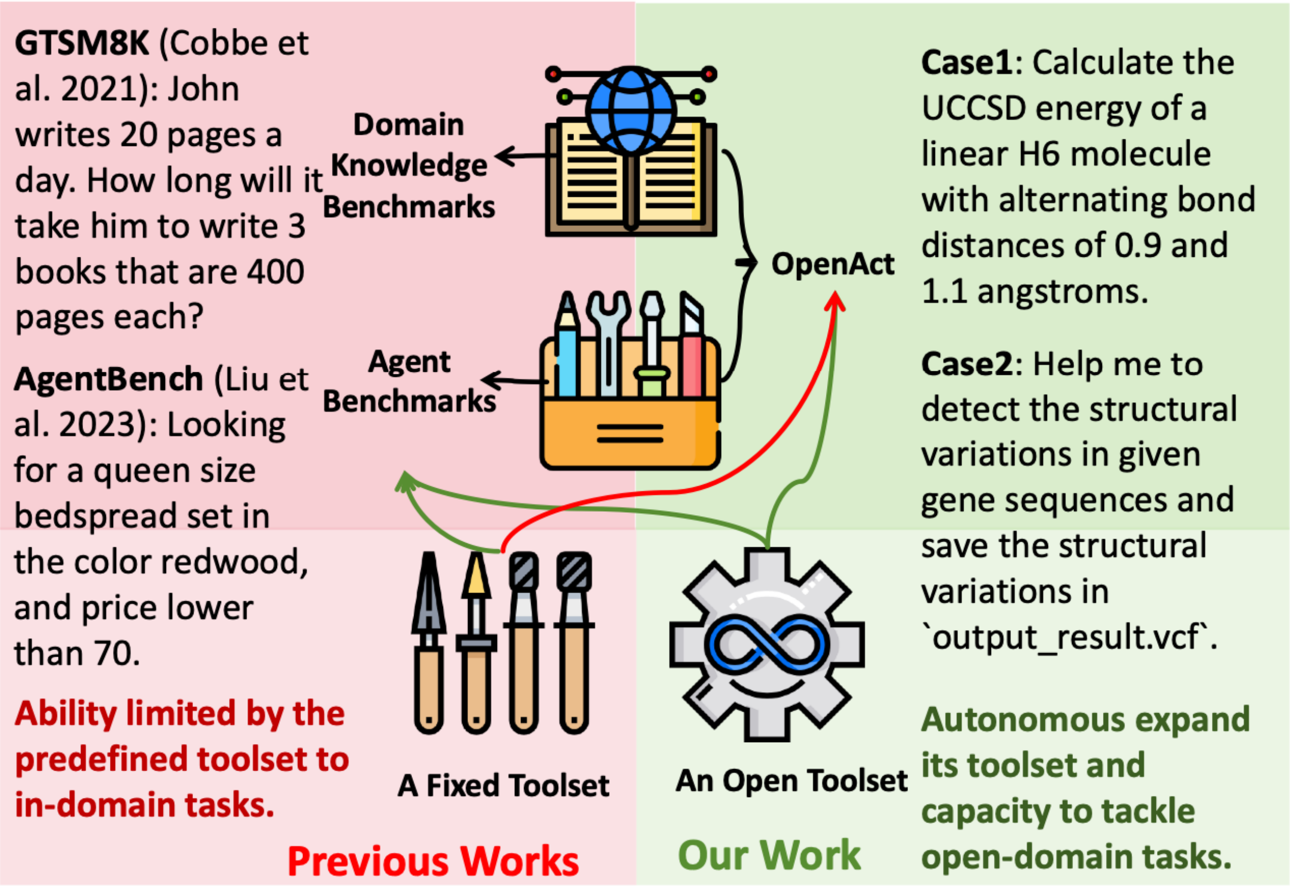
(Long Paper, Main Conference)

Bohan Lyu[1*], Xin Cong[1*†], Heyang Yu[1], Pan Yang[1], Cheng Qian[1,3], Zihe Wang[1], Yujia Qin[1], Yining Ye[1], Yaxi Lu[1], Chen Qian[1,4], Zhong Zhang[1], Yukun Yan[1], Yankai Lin[2], Zhiyuan Liu[1†], Maosong Sun[1]

[1] Department of Computer Science and Technology, Tsinghua University
[2] Gaoling School of Artificial Intelligence, Renmin University of China
[3] University of Illinois Urbana-Champaign
[4] School of Artificial Intelligence, Shanghai Jiao Tong University

lvbh22@mails.tsinghua.edu.cn,congxin1995@tsinghua.edu.cn

Background and Motivation



- LLM -> Agent
- Open-Ended Problems that require external tools
- Dataset: OpenAct
- Method: OpenAgent

Existing Benchmarks

Benchmark	Domain Num.	Task Source	Task Types	Code Use	Tool Use	Open End	Repository-Level
Minedojo (Fan et al., 2022)	-	Internet	Action	✓	✓	✓	✗
OSWorld (Xie et al., 2024)	-	Internet	Action	✗	✓	✓	✗
ToolBench (Qin et al., 2023b)	-	Tool	QA	✗	✓	✗	✗
MetaTool (Huang et al., 2024b)	-	Tool	QA	✗	✓	✗	✗
AgentBench (Liu et al., 2023)	-	Tool	QA	✓	✓	✗	✗
GTSM8K (Cobbe et al., 2021)	1	Domain	QA	✗	✓	✗	✗
ScienceQA (Lu et al., 2022)	3	Domain	QA	✓	✗	✗	✗
SciEval (Sun et al., 2023)	3	Domain	QA	✗	✗	✗	✗
SciBench (Wang et al., 2024b)	3	Domain	QA	✓	✗	✗	✗
SWE-Bench (Jimenez et al., 2024)	1	GitHub	Coding	✓	✓	✗	✓(12)
ML-Bench (Tang et al., 2024)	1	GitHub	Coding	✗	✓	✗	✓(14)
SUPER (Bogin et al., 2024)	-	GitHub	QA	✓	✓	✗	✓(45)
OpenAct (Ours)	7	Domain and Github	QA and Coding	✓	✓	✓	✓(21)

Table 1: Comparison of benchmarks for evaluating LLMs on domain knowledge and tool utilization. The "Domain Num." column indicates the number of domains evaluated by each benchmark, with "-" denoting benchmarks that do not assess domain knowledge. "Open End" denotes the presence of an open-ended environment for exploration within the benchmark. "Repository-Level" specifies whether the tasks in the benchmark are scoped at the repository level, with the number in the bracket denoting the number of repositories relevant to the benchmark.

Our Dataset: OpenAct

- Construction Method

- Consitution and Category

Domain	Num. of Repo.	Num. of Query
Finance	2	45
Chemistry	4	66
Bioinformatics	2	30
Computer Vision	6	90
Network Analysis	2	30
Security Analysis	2	30
Visualization	3	48
Total	21	339

Table: Statistics of our constructed OpenAct.

	App. Easy	**Medium**	**Hard**
Env. Easy	Pyflowchart, Bolt, yolov5	OCRmyPDF, Rembg	TenCirChem, ChemFor-mula, Chem-lib
Env. Medium	MultiQC, Photon, Smap	Bandit, recognize-anything	Aizynthfinder, mermaid-cli
Env. Hard	Latex-OCR	BOPBL	qlib, PlotNN

Table: GitHub repositories classified by difficulties.

Our Method: OpenAgent

OpenAgent tackles three main challenges: lack of quality assurance in GitHub repositories, alignment gaps between tools and queries, and workflow complexity. OpenAgent introduces two key innovations:

1. **Hierarchical Agent System**: A multi-level structure where tasks are broken down into subtasks, with agents either taking direct actions or delegating to sub-agents. See Figure 4 for details.
2. **Bi-Level Experience Learning**: Incorporates both in-task learning (using GitHub Issues/PRs) and cross-task learning (from past experiences). A specialized Issue/PR Agent handles experience-based problem-solving, while the system stores successful environments in Docker images for future use.

The system operates in three main phases:

- **Repository Search**: Identifies suitable repositories by checking stored options or searching GitHub;
- **Environment Setup**: Configures execution environment with necessary dependencies;
- **Tool Application**: Applies the repository to solve user queries.

Hierarchical Agent System

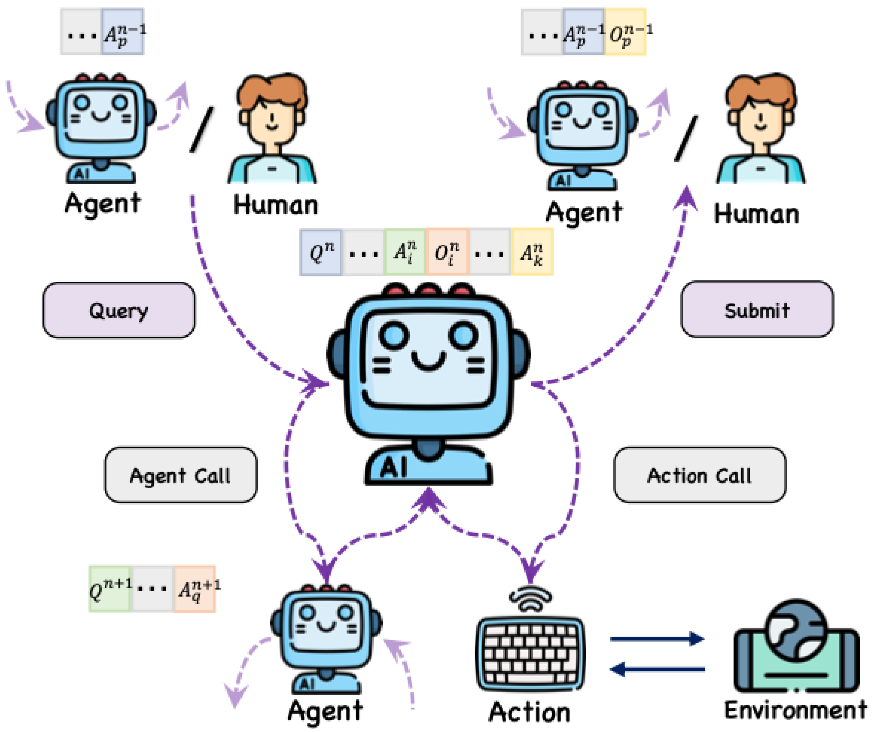


Figure: Illustration of the Hierarchical Agent System, where blocks mean memory list and the same background color denotes the same information.

Formally, we use $Agent_k^n$ to denote the k-th agent at level n of the hierarchy, A_i^n denotes the i-th action, which can be tool-using or designating inferior agents, by $Agent^n$. When $Agent_k^n$ receives query Q^n from its superior agent or human, the problem solving process can be formulated as $A_i^n = Agent_k^n(Q^n, A_j^n, O_j^n, \ldots, A_1^n, O_1^n)$, where O_j^n and A_j^n are respectively the observations and preceding actions that lead up to A_i^n. If A_i^n is calling a sub-agent, the query Q^{n+1} for $Agent_k^{n+1}$ is derived from A_i^n. When $Agent_k^{n+1}$ finishes its task, it will report the result A_q^{n+1} to $Agent_k^n$. Figure 4 demonstrated this hierarchical recursive process.

Experimental Results

Vanilla LLM, ReAct, ReAct + Summary, XAgent and OpenAgent on OpenAct, with 2 LLM backbones.

Methods	Finance	Chemistry	Bioinformatics	Computer Vision	Network Analysis	Security Analysis	Visualization	**Avg.**
				GPT-3.5-Turbo Based				
Vanilla	0.0	**36.4**	0.0	0.0	0.0	0.0	**31.3**	11.5
ReAct	2.2	3.0	3.3	6.7	0.0	0.0	0.0	2.4
ReAct + Sum.	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
OpenAgent (Ours)	**8.9**	24.2	**23.3**	**8.9**	**10.0**	**33.3**	20.1	**17.1**
				GPT-4 Based				
Vanilla	0.0	**68.2**	0.0	0.0	0.0	0.0	43.8	19.5
XAgent	0.0	40.9	0.0	0.0	**40.0**	0.0	**81.3**	23.0
ReAct	51.1	19.7	17.8	22.2	10.4	30.0	23.3	24.6
ReAct + Sum.	31.1	19.7	26.7	22.9	14.8	33.3	26.7	24.4
OpenAgent (Ours)	**68.9**	34.9	**86.7**	**45.6**	16.7	**43.3**	35.4	**47.3**

Table: Pass Rates (%) of different methods across various domains in the OpenAct dataset. Results are shown for both GPT-3.5-Turbo and GPT-4-based implementations. “Avg.” represents the average pass rate across all domains.

Ablation Study

For **in-task experience learning**, we remove the PRs/Issues actions to re-run the main experiments. It verifies the non-standardization problem of GitHub repositories and proves that learning from PRs/Issues can overcome this challenge. For **cross-task experience learning**, we select 2 repositories: `Qlib` and `AiZynthFinder`. We run their 51 queries and utilize the GPT-4-based OpenAgent to store the repositories with summarized practice experience. We then re-run these queries but OpenAgent would retrieve the stored repositories and utilize the summarized experience to accomplish the queries.

Method	Pass Rate
OpenAgent w/ PRs&Issues	47.3
OpenAgent w/o PRs&Issues	40.3

Table: Results of ablating in-task experience learning.

Method	w/o SelfExp	w/ SelfExp
GPT-3.5	17.6	58.8
GPT-4	47.0	82.3

Table: Results of employing cross-task experience learning.

Impact of Different Stages

We also analyze the impact of different phases in our model, examining search success rates across different prompt types and pass rates across varying setup/apply difficulties.

Prompt	Search Success Rate
Explicit Repo Prompt	96.0
Implicit Repo Prompt	66.0
No Repo Prompt	32.0

Table: Analysis for the search difficulty.

Setup/Apply Difficulty	Easy	Medium	Hard	Total
Easy	72.3	69.0	56.2	64.4
Medium	60.7	70.0	41.5	57.7
Hard	50.0	67.0	51.5	57.4
Total	64.1	68.7	51.4	60.7

Table: Analysis for the setup & apply difficulty.

Key Takeaways

- **Open-Domain Task-Solving is Challenging**: Even state-of-the-art LLMs struggle with complex domain-specific tasks that require specialized tools. Our experiments show significant performance gaps when dealing with open-domain problems across diverse fields.
- **Hierarchical Agent Structure is Effective**: Breaking down complex tasks into manageable subtasks through a hierarchical agent framework significantly improves performance. This approach allows specialized agents to focus on specific aspects of the workflow while maintaining overall coherence.
- **Experience Learning is Crucial**: The bi-level experience learning mechanism substantially enhances performance by:
  - Learning from human experiences (Issues/PRs) to overcome repository flaws
  - Accumulating and leveraging cross-task experiences to improve future problem-solving

Thank You for Your Attention!

Paper: https://openreview.net/forum?id=cDppq8dYFA
GitHub: https://github.com/OpenBMB/OpenAct

Questions? Contact me:

Bohan Lyu https://lyubh.cn

I am an undergraduate at Tsinghua University. I'm interested in ML and NLP topics. My works are published in ICML and ACL. I am seeking PhD opportunities starting in Fall 2026. Please feel free to reach out!