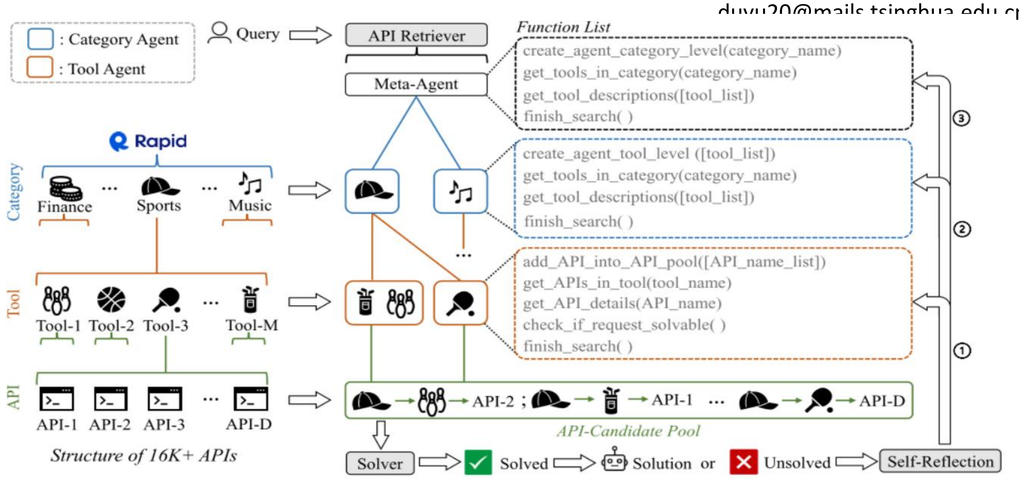
AnyTool: Self-Reflective, Hierarchical Agents for Large-Scale API Calls

Yu Du[1] Fangyun Wei[2†] Hongyang Zhang[3]

1Tsinghua University 2Microsoft Research Asia 3 University of Waterloo

duyu20@mails.tsinghua.edu.cn fawe@microsoft.com hongyang.zhang@uwaterloo.ca



Introduction

We introduce AnyTool, a large language model agent designed to revolutionize the utilization of a vast array of tools in addressing user queries. We utilize over 16,000 APIs from Rapid API, operating under the assumption that a subset of these APIs could potentially resolve the queries.

AnyTool primarily incorporates three elements:

1. an API retriever with a hierarchical structure
2. a solver aimed at resolving user queries using a selected set of API candidates
3. a self-reflection mechanism, which re-activates AnyTool if the initial solution proves impracticable.

We also revise the **evaluation protocol** to better reflect real-world application scenarios.

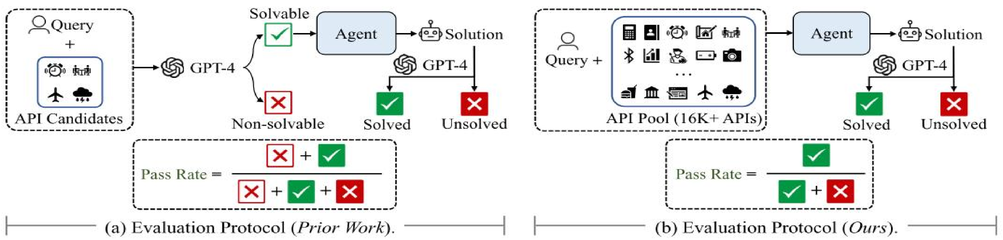


Illustration of the evaluation protocols used by: (a) ToolLLM [1]; and (b) ours. In (a), if the API retriever selects candidates completely unrelated to the user's query, GPT-4 may classify all queries as "non-solvable", leading to an artificially high pass rate, despite the queries remaining unsolved. In (b), we conduct a manual review of all queries and retain only those queries that can be resolved with specific APIs from the API pool for ToolBench [1].

Results

Table 1: Main results on the filtered ToolBench. We use pass rate defined in Eq 2 and illustrated in Figure 4(b), as the metric. All results are reproduced. *: OpenAI's text-embedding-ada-002; Ref.: reference; Avg.: average; SR: self-reflective.

Model	API Retriever	Solver	Use Ref. APIs	G1			G2		G3	Avg. (%)
				I (%)	T (%)	C (%)	I (%)	C (%)	I (%)	
ToolLLM	OpenAI TE*	ToolLLaMA w/ DFSDT		8.7	6.8	12.0	4.7	8.2	10.5	8.5
ToolLLM	ToolLLM's	ToolLLaMA w/ DFSDT		28.4	26.3	38.4	21.5	15.1	7.7	22.9
ToolLLM	ToolLLM's	GPT-4 w/ DFSDT		42.6	46.2	51.4	23.4	24.5	2.6	31.8
ToolLLM	None	ToolLLaMA w/ DFSDT	✓	29.4	31.8	37.1	19.6	22.4	13.2	25.6
GPT-4	None	GPT-4 w/ CoT	✓	31.3	34.8	47.1	27.1	34.7	2.6	29.6
GPT-4	None	GPT-4 w/ DFSDT	✓	36.5	49.2	51.4	38.3	39.8	18.4	38.9
GPT-4	Plain Agent	GPT-4 w/ DFSDT		13.9	23.5	17.6	13.9	9.2	13.2	15.2
GPT-4	AutoGen-RAG	GPT-4 w/ DFSDT		14.8	19.7	19.7	7.4	9.2	7.9	13.1
GPT-3.5	None	GPT-3.5 w/ CoT	✓	37.5	37.1	42.9	24.3	22.4	5.3	28.3
GPT-3.5	None	GPT-3.5 w/ DFSDT	✓	39.1	40.2	48.6	31.8	25.5	15.8	33.5
AnyTool (Ours)	SR Agent	SR GPT-4 w/ DFSDT		52.2	61.4	67.6	58.9	45.9	63.2	58.2

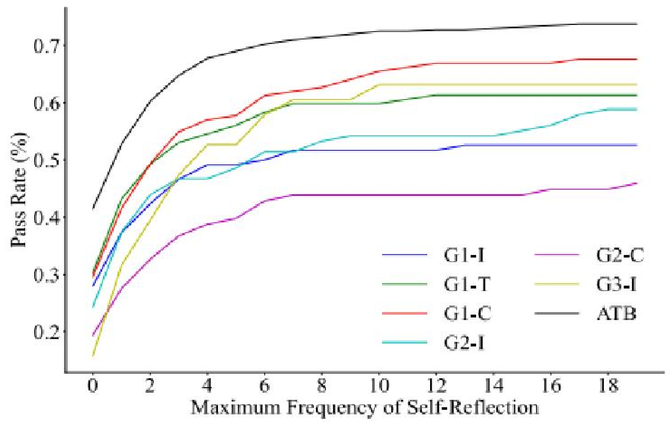


Figure 3: The performance of our AnyTool on different datasets (each denoted by a curve) improves as the number of self-reflection rounds increases. ATB: AnyToolBench.

Conclusion

Our innovation include

- A hierarchical API retriever coupled with a solver.
- A self-reflection mechanism, enhancing its proficiency in responding to user queries.
- A revised evaluation protocol to better reflect real-world application scenarios.
- Rigorous experiments conducted on ToolBench and our AnyToolBench demonstrate our approach's superiority over established models

Future research directions:

1) optimizing the organization of APIs for improved performance and efficiency
2) developing an advanced open-source LLM specifically for API utilization, which could facilitate local deployments

References

[1] Qin, Yujia, et al. "ToolLLM: Facilitating Large Language Models to Master 16000+ Real-world APIs." *The Twelfth International Conference on Learning Representations*.