Learning Evolving Tools for Large Language Models

Guoxin Chen[1], Zhong Zhang[2,*], Xin Cong[2,*], Fangda Guo[1],
Yesai Wu[2], Yankai Lin[3], Wenzheng Feng[4], Yasheng Wang[4]

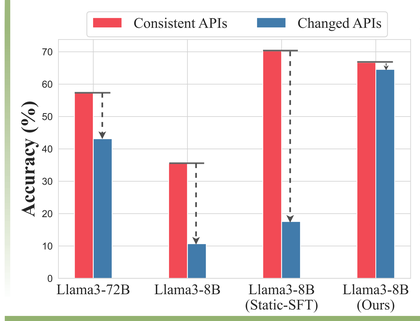
[1]Institute of Computing Technology, CAS; [2]Tsinghua University
[3]Renmin University of China; [4]Huawei Noah's Ark Lab

Paper HomePage

Motivation

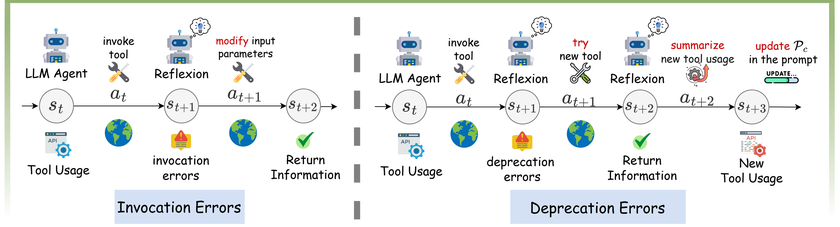
- The **rapid and frequent change of tools** is difficult to capture in a timely manner.
- There is a **discrepancy** between the APIs that LLMs have learned to use and those deployed in the real-world environment
- If **the specified tools in the prompt do not keep pace with changes in the external environment**, the model will incorrectly invoke the outdated APIs instead of the latest APIs.

Impact of Tool Variability

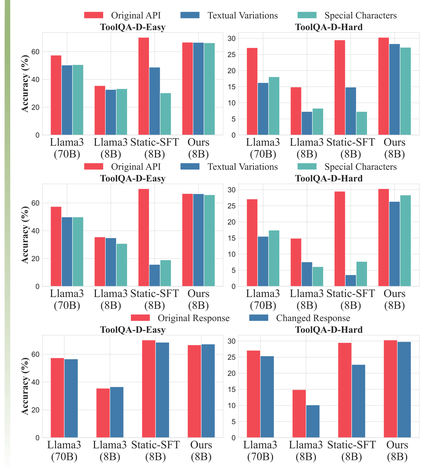


- "Consistent APIs" refer to APIs that are **consistent** between LLMs and servers
- "Changed APIs" refer to APIs accessible to LLMs that are **outdated** over time
- "Static-SFT" is supervised fine-tuning on tool usage data that has no adaptability to tool variability.

Self-reflection and Tool Update

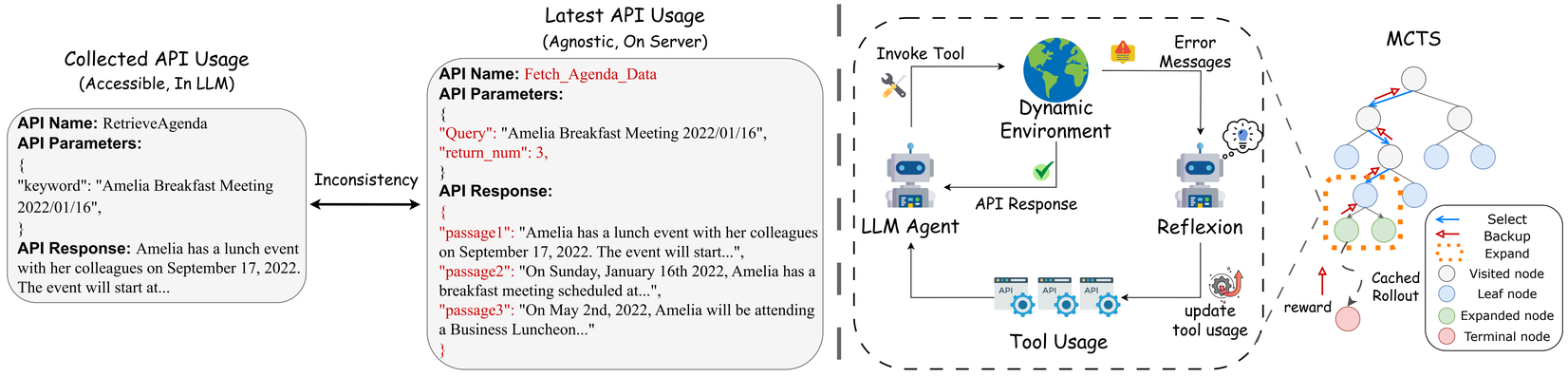


Analysis on Tool Variability



- **Static-SFT model** exhibits more fluctuations in the performance regarding tool variability.
- **Insertion of special characters** poses significant challenges for tool-using abilities.
- **Changes in API parameters** have a more substantial influence on performance.
- Even under the same changes, the model's performance tends to decline more significantly on **challenging tasks** relative to simpler ones.

Overview



Experiments & Results

Models	Agenda		Airbnb		Coffee		Dblp		Flights		Scirex		Yelp		Average	
	Easy	Hard	Easy	Hard	Easy	Hard	Easy	Hard	Easy	Hard	Easy	Hard	Easy	Hard	Easy	Hard
Static Environment ($\mathcal{P}_c$ in Prompt and $\mathcal{P}_c$ on Server)																
Llama3-70B-Instruct	55.0	55.0	76.0	27.0	94.0	**3.8**	32.0	32.0	59.0	26.0	0.0	2.0	86.0	44.0	57.4	27.1
Llama3-8B-Instruct	25.0	21.0	68.0	20.0	59.0	0.8	19.0	20.0	24.0	17.0	0.0	1.0	54.0	25.0	35.5	14.9
Static-SFT (Llama3-8B)	68.0	**65.0**	95.0	**33.0**	**100.0**	0.8	**55.0**	32.0	**86.0**	24.0	0.0	1.0	88.0	50.0	**70.2**	29.5
TOOLEVO (Llama3-8B)	**70.0**	57.0	**96.0**	30.0	88.0	1.5	49.0	**34.0**	69.0	**36.0**	**1.0**	**3.0**	**95.0**	**51.0**	66.7	**30.3**
Qwen2-72B-Instruct	55.0	46.0	79.0	32.0	95.0	1.5	42.0	**39.0**	70.0	26.0	**3.0**	1.0	81.0	45.0	60.7	27.2
Qwen2-7B-Instruct	40.0	35.0	59.0	12.0	94.0	3.1	31.0	25.0	46.0	13.0	0.0	1.0	68.0	15.0	48.2	14.8
Static-SFT (Qwen2-7B)	68.0	**55.0**	**97.0**	**34.0**	**98.0**	4.6	50.0	37.0	75.0	29.0	1.0	3.0	92.0	**53.0**	68.8	30.8
TOOLEVO (Qwen2-7B)	**76.0**	50.0	94.0	33.0	95.0	**6.1**	**51.0**	38.0	**84.0**	**45.0**	2.0	**8.0**	**93.0**	41.0	**70.7**	**31.5**
Dynamic Environment ($\mathcal{P}_c$ in Prompt and $\mathcal{P}_{s_{in}}$ on Server)																
Llama3-70B-Instruct	55.0	40.0	78.0	12.0	70.0	2.3	31.0	22.0	42.0	18.0	**4.0**	3.0	87.0	35.0	52.4	18.9
Llama3-8B-Instruct	23.0	21.0	63.0	10.0	44.0	0.0	21.0	13.0	16.0	10.0	1.0	2.0	53.0	13.0	31.5	9.8
Static-SFT (Llama3-8B)	53.0	10.0	49.0	6.0	14.0	0.0	44.0	11.0	15.0	29.0	0.0	1.0	86.0	34.0	37.2	13.0
TOOLEVO (Llama3-8B)	**61.0**	**53.0**	**95.0**	**26.0**	**88.0**	**4.6**	**50.0**	**32.0**	**74.0**	**34.0**	2.0	**5.0**	**93.0**	**48.0**	**66.2**	**28.9**
Qwen2-72B-Instruct	56.0	38.0	73.0	11.0	78.0	0.0	42.0	28.0	54.0	12.0	1.0	2.0	75.0	34.0	54.1	18.4
Qwen2-7B-Instruct	32.0	30.0	60.0	8.0	68.0	0.8	32.0	16.0	32.0	12.0	1.0	0.0	66.0	28.0	41.5	13.5
Static-SFT (Qwen2-7B)	49.0	27.0	52.0	21.0	35.0	0.8	41.0	18.0	31.0	25.0	0.0	1.0	78.0	29.0	40.8	17.3
TOOLEVO (Qwen2-7B)	**66.0**	**41.0**	**94.0**	**36.0**	**97.0**	**5.4**	**46.0**	**39.0**	**85.0**	**41.0**	**1.0**	**8.0**	**92.0**	**54.0**	**68.7**	**32.1**
OOD Dynamic Environment ($\mathcal{P}_c$ in Prompt and $\mathcal{P}_{s_{OOD}}$ on Server)																
Llama3-70B-Instruct	56.0	28.0	61.0	8.0	65.0	0.0	38.0	23.0	26.0	18.0	**3.0**	**8.0**	53.0	29.0	43.1	16.3
Llama3-8B-Instruct	27.0	7.0	7.0	9.0	1.0	0.0	28.0	13.0	2.0	9.0	1.0	4.0	9.0	14.0	10.7	8.0
Static-SFT (Llama3-8B)	58.0	9.0	23.0	8.0	11.0	0.0	24.0	6.0	8.0	11.0	0.0	3.0	19.0	26.0	20.4	9.0
TOOLEVO (Llama3-8B)	**71.0**	**49.0**	**65.0**	**29.0**	**88.0**	**2.3**	**51.0**	**34.0**	**63.0**	**33.0**	2.0	4.0	**91.0**	**47.0**	**61.6**	**28.3**
Qwen2-72B-Instruct	52.0	33.0	39.0	11.0	71.0	0.0	42.0	29.0	22.0	11.0	2.0	4.0	60.0	34.0	41.1	17.8
Qwen2-7B-Instruct	37.0	22.0	55.0	7.0	74.0	1.5	26.0	14.0	37.0	3.0	1.0	3.0	67.0	27.0	42.4	11.1
Static-SFT (Qwen2-7B)	37.0	31.0	63.0	8.0	68.0	0.7	35.0	17.0	58.0	22.0	1.0	1.0	73.0	39.0	47.8	16.9
TOOLEVO (Qwen2-7B)	**68.0**	**48.0**	**89.0**	**14.0**	**81.0**	**6.9**	**48.0**	**34.0**	**85.0**	**33.0**	**3.0**	**6.0**	**83.0**	**52.0**	**65.3**	**27.7**

Conclusion:

- Even without fine-tuning with tool trajectories of $\mathcal{P}_c$, **trial-and-error experiences** focus on tool variability can still enhance the tool-using capabilities in static environments.
- Our TOOLEVO empowers the model with the ability to **self-reflect and self-update its existing tool usage based on environmental feedback**, rather than merely **memorizing existing invocation patterns**.
- The **stereotypes** induced by Static-SFT lead to **extreme confidence** in the tool usage provided in the prompt, thereby rendering it ineffective in handling tool variability.
- Without dynamic environment, the model tends to **lazily** focus on how to use APIs provided in the prompt.
- tool variability have **a severely negative impact** on LLMs, which needs to be taken seriously in future work.
- Learning **how to use a tool in a dynamic environment** will yield more benefits than simply imitating tool usage in a static environment.