浙江大學
ZHEJIANG UNIVERSITY

LREC-COLING 2024

Towards Autonomous Tool Utilization in Language Models: A Unified, Efficient and Scalable Framework

Zhi Li, Yicheng Li, Hequan Ye and Yin Zhang
Zhejiang University, China
{zhili, yichengli, yehequan, zhangyin98}@zju.edu.cn

Background

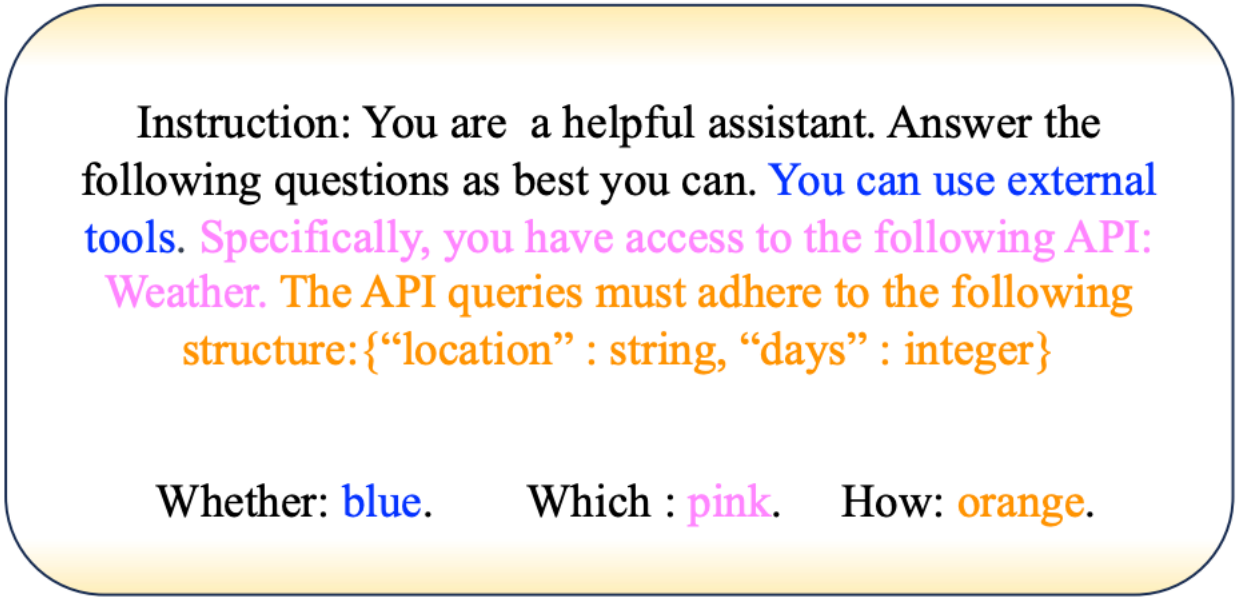


Figure 1: Using weather queries as an example, we showcase the prevailing methods of tool utilization.

Motivation

Potential Drawbacks:

- the limited context length
- additional retrieval steps leading to cumulative errors
- plays a passive role

Question:

Can we develop a language model capable of autonomously solving problems without the external prompt guidance ?

Comparison

Method	End to End Learning			Advantage		
	Whether	Which	How	Efficient	Scalable	Autonomous
Toolformer (Schick et al., 2023)	✓	✓	✓			✓
TRICE (Qiao et al., 2023)	✓		✓			✓
Gorilla(Patil et al., 2023)			✓	✓	✓	
ToolLLM (Qin et al., 2023)			✓	✓	✓	
GPT4Tools(Yang et al., 2023)	✓	✓	✓	✓	✓	
ToolAlpaca (Tang et al., 2023)	✓	✓	✓	✓	✓	
Our Work	✓	✓	✓	✓	✓	✓

Table 1: Comparative analysis of various studies. 'Autonomous' implies that there is no need to include any tool-related information in the prompt, as the model fully internalizes the knowledge and ability to use the tools.

Problem Formalization

Let's first agree on the abbreviations for the terms:

- Y denotes the target response.
- X stands for the query.
- $\mathrm{W_e}$ determines whether to use the tool.
- $\mathrm{W_i}$ indicates which tool to use.
- H indicates how to use the tool.

$$P(\mathrm{Y} \mid \mathrm{W_e}, \mathrm{W_i}, \mathrm{H}, \mathrm{X})$$

$$P(\mathrm{Y}, \mathrm{W_e}, \mathrm{W_i}, \mathrm{H} \mid \mathrm{X})$$

Unified Conditional Probability Decomposition

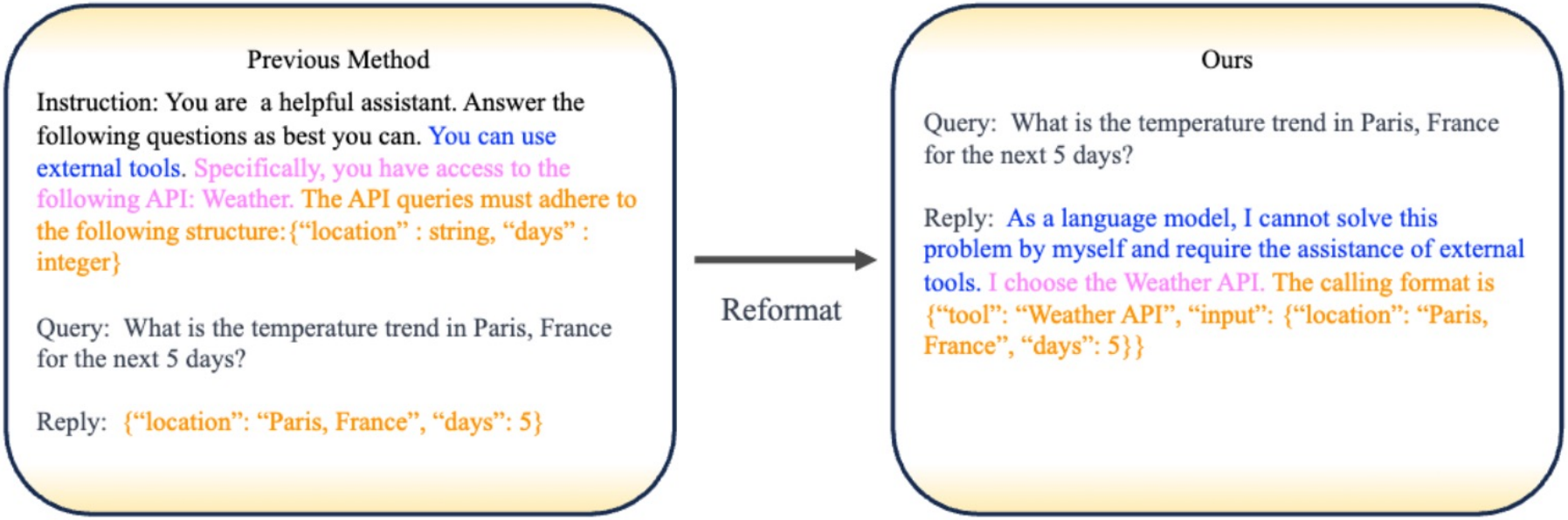
$$
\begin{aligned}
&P(\mathrm{Y}, \mathrm{W_e}, \mathrm{W_i}, \mathrm{H} \mid \mathrm{X}) \\
&\quad = P(\mathrm{W_e} \mid \mathrm{X}) \\
&\quad \times P(\mathrm{W_i} \mid \mathrm{W_e}, \mathrm{X}) \\
&\quad \times P(\mathrm{H} \mid \mathrm{W_e}, \mathrm{W_i}, \mathrm{X}) \\
&\quad \times P(\mathrm{Y} \mid \mathrm{H}, \mathrm{W_e}, \mathrm{W_i}, \mathrm{X})
\end{aligned}
$$

Query categories:

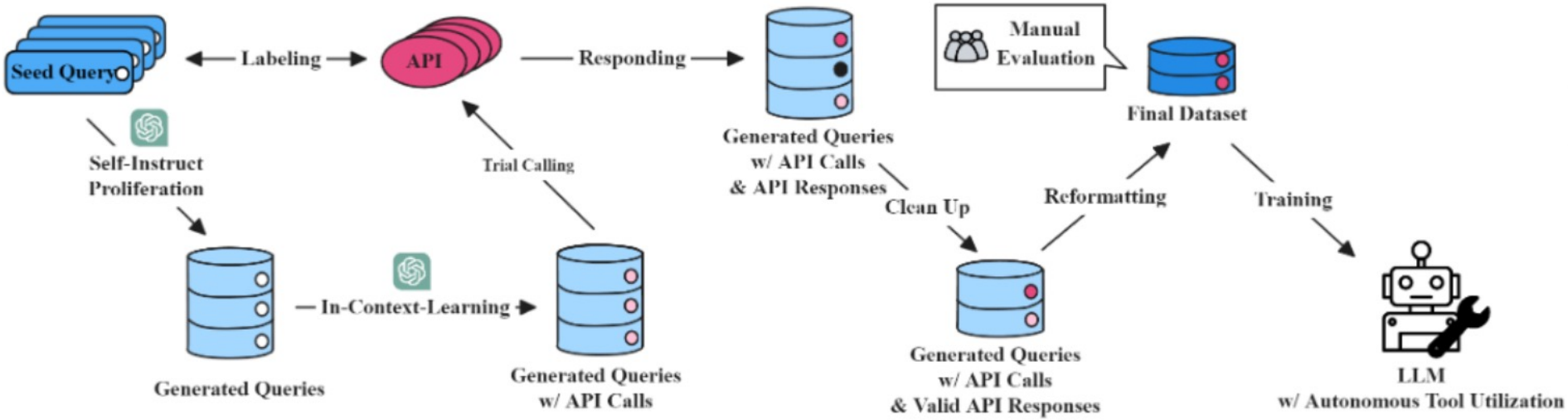
- Solvable: translation, sentiment analysis …
- Verifiable : mathematical reasoning, factual matters …
- Unsolvable: real-time, multimodal …

Efficient Dataset Construction

$$P(\mathrm{Y} \mid \mathrm{W_e}, \mathrm{W_i}, \mathrm{H}, \mathrm{X}) \longrightarrow P(\mathrm{Y}, \mathrm{W_e}, \mathrm{W_i}, \mathrm{H} \mid \mathrm{X})$$



Efficient Dataset Construction



Dynamic Rehearsal for Scalability

C_{new}: categories of newly introduced tool
C_{old} : categories of tools already learned
t : the current training epochs
$\alpha(t)$: the proportion of sampling from old tools in training dataset at epoch t

$$\alpha(t) = \min\left(\frac{C_{old}}{C_{new} + C_{old}}, \gamma \cdot t\right)$$

γ : The hyperparameter that controls the increment of $\alpha(t)$ each epoch until it equals to $\frac{C_{old}}{C_{new}+C_{old}}$

Tool Statistics

Type	Domain	Task	API
Solvable	NLP	Translation Sentiment Analysis Story Generation Summarization Grammar Correction etc	Language model itself
Verifiable	Reason	Math	Wolfram Alpha
	Fact	Entertainment	Bing Search
		Education	
		Culture and art	
		Politics	
		Sports	
		Map	Get route
			Get coordinates
			Get distance
			Search nearby
Unsolvable	Time	Recent Paper	Arxiv
		Stock	Get daily prices
			Get exchange rate
			Get open info
		Weather	Get weather today
			Forecast weather
	Mutimodal	Image Understanding	Image captioning
			Detect the Given Object
			Answer Question About The Image
			Segment the given object
			Remove Something From The Photo
		Image Generation	Generate Image Condition On Text
			Generate Image Condition On Normal Map
			Generate Image Condition On Pose Image
			Generate Image Condition On Depth
			Generate Image Condition On Sketch Image
			Generate Image Condition On Segmentations
			Generate Image Condition On Soft Hed Boundary Image
			Generate Image Condition On Canny Image

For each API, 800 for training, 100 for validation, and 100 for testing

Baseline && Evaluation

We chose Alpaca 7B as our base model to training on the dataset.

Baseline: Alpaca, Vicuna, Llama2Chat, GPT-3.5 and GPT-4.

We employ accuracy to evaluate the four aspects : Whether, Which, How and Success.

Result

Model	Whether	Which	How	Success
Alpaca	18.9	10.3	9.3	7.5
Vicuna	45.8	22.3	15.6	13.9
Llama2Chat	56.9	29.7	25.2	23.3
GPT-3.5	72.8	69.9	65.5	62.8
GPT-4	89.7	86.5	82.3	80.4
Ours	98.6	97.6	92.3	91.2

Table 3: Accuracy of 6 models across four critical aspects.

Result

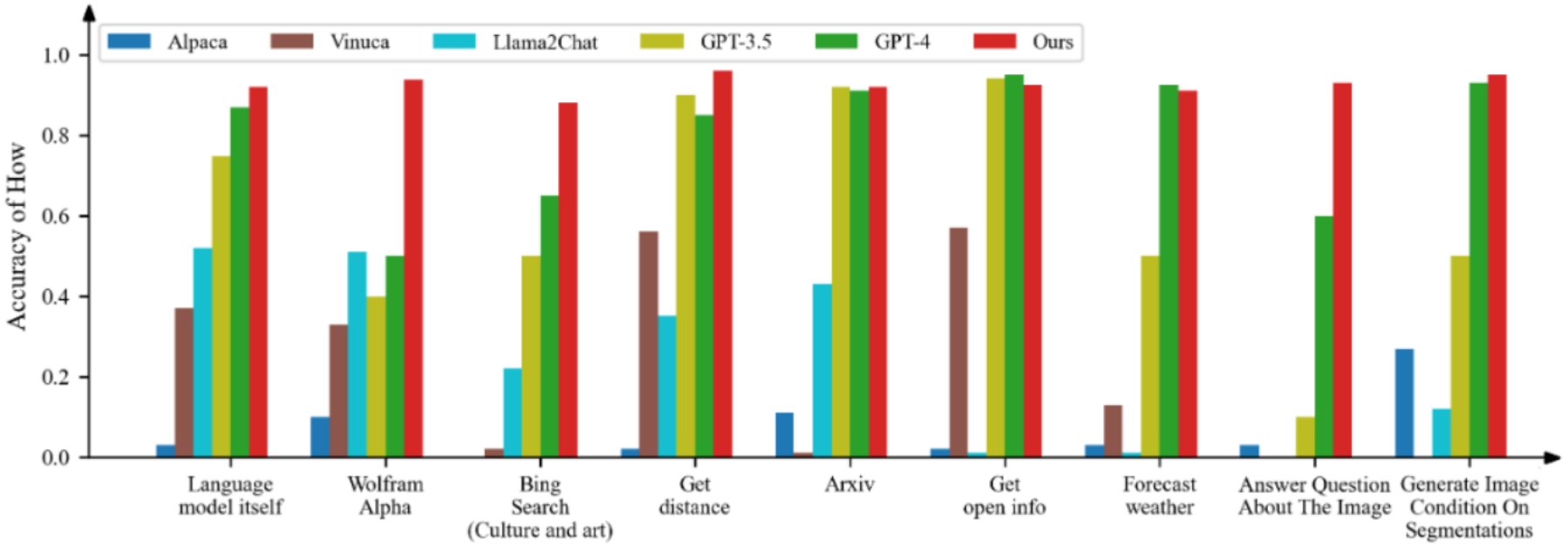


Figure 4: The performance of different models on various APIs.

Bad case analysis: 1.Overconfidence. 2.Disregard.

Dynamic Rehearsal for Scalability

Strategy	New APIs	Old APIs
without Replay	91.6	84.7
with Replay	91.1	90.88

Table 4: The success rate under the scenario of continual learning

Ablation

Method	How	Success
Ours	92.3	91.2
w/o Whether	78.7	75.4
w/o Which	66.4	62.0
w/o Whether+Which	55.6	52.9

Table 5:An in-depth ablation on the Unified Conditional Probability Decomposition

Conclusion

- **Unified** conditional probability decomposition
- Reformat for **Efficient** dataset construction
- Dynamic rehearsal strategy for **Scalability**

Thank you