

WTU-EVAL: A Whether-or-Not Tool Usage Evaluation Benchmark for Large Language Models

Anonymous ACL submission


Abstract


Although Large Language Models (LLMs) excel in NLP tasks, they still need external tools to extend their ability. Current research on tool learning with LLMs often assumes mandatory tool use, which does not always align with real-world situations, where the necessity for tools is uncertain, and incorrect or unnecessary use of tools can damage the general abilities of LLMs. Therefore, we propose to explore whether LLMs can discern their ability boundaries and use tools flexibly. We then introduce the Whether-or-not tool usage Evaluation benchmark (WTU-Eval) to assess LLMs with eleven datasets, where six of them are tool-usage datasets, and five are general datasets. LLMs are prompted to use tools according to their needs. The results of eight LLMs on WTU-Eval reveal that LLMs frequently struggle to determine tool use in general datasets, and LLMs' performance in tool-usage datasets improves when their ability is similar to ChatGPT. In both datasets, incorrect tool usage significantly impairs LLMs' performance. To mitigate this, we also develop the finetuning dataset to enhance tool decision-making. Fine-tuning Llama2-7B results in a 14% average performance improvement and a 16.8% decrease in incorrect tool usage. We will release the WTU-Eval benchmark.


1 Introduction

LLMs like GPT4 (OpenAI, 2023) excel in general-purpose NLP tasks, which however struggle in tasks requiring calling external tools, such as fetching up-to-date stock prices and booking flights (Qin et al., 2023a,b; Patil et al., 2023; Ruan et al., 2023). How to improve LLMs' interaction with tools is a current hot topic. Recent studies have explored tool usage fine-tuning (Qin et al., 2023b; Patil et al., 2023) and reinforcement learning (Li et al., 2023) techniques, showing promising results in areas like math reasoning and web search.

System Instruction: You have access to some tools to do the following task if you need.

 : The Quarter Pounder is a hamburger sold by international fast food chain McDonald's, so named for containing a patty with a precooked weight of a quarter of a pound (113.4 g)... **In 2015, McDonald's increased the precooked weight to 4.25 oz (120.5 g).**

 : Does a quarter pounder weight a quarter pound?

 : The passage mentions that a Quarter Pounder contains a patty with a precooked weight of a quarter of a pound, which is equivalent to 113.4 grams. To answer the question, I should verify this weight in grams. I can use the Calculator API to convert 1/4 pound to grams for comparison.

..... call tool


 : `Calculator[1/4 pound to grams]`

Figure 1: An example showing the failure of calling tools inappropriately.

Despite notable progress, prior studies (Patil et al., 2023; Zhuang et al., 2023) mainly focused on scenarios mandating tool use by LLMs. However, in a real-world application, the necessity for tool usage is uncertain. Moreover, we observe that inappropriate tool invocation can lead to errors, adversely affecting outcomes. For example, Figure 1 provides an example of using ChatGPT (0613) to answer a question. Despite the context hinting at the answer: a quarter pounder's weight has been increased to 120.5g, not a quarter pound (113.4g), ChatGPT still invokes an external tool, *Calculator*, and due to incorrect parameter settings, it produces an erroneous response and redundant response time.

With the above observations, we want to explore an intriguing question: whether LLMs can discern their ability boundaries, and if LLMs have the option to decide whether to use tools, would their performance improve in general and tool-usage datasets?

To this end, we propose a Whether-or-not tool usage Evaluation benchmark (WTU-Eval), which contains six tool-usage datasets that explicitly require tool usage and five general datasets that can be answered without tools. As illustrated in the accompanying Figure 2, Region1 (R1) and Region3 (R3) are baselines that test LLMs without tools, and Region2 (R2) and Region4 (R4) evaluate LLMs that have the option to use tools flexibly according to their needs.

Furthermore, we also develop a dataset from the WTU-Eval benchmark training sets, resulting in a finetuning dataset with a size of 4000. This dataset is used to enhance the model’s decision-making capability regarding tool use, resulting in a 14% average performance improvement and a 16.8% decrease in incorrect tool usage with finetuning Llama2-7B, which also gains a significant improvement by up to 40% for the PIQA’s Search Engine—and reduces the tool invocation rate (§5.3).

The contributions of this paper are as follows:

- We propose to explore whether LLMs can discern their ability boundaries and use tools flexibly and introduce the WTU-Eval, which is the first benchmark to evaluate whether to use tools accurately.
- We rigorously evaluate the performance of eight well-known LLMs and highlight their limitations. Most LLMs struggle to recognize their capability boundaries and lack of tool usage decision-making capability.
- Based on the above insights, we also introduce a finetuning dataset, particularly for enhancing the model’s decision-making capability regarding tool use, showing its positive effects.

2 Related Work

Integrating tool calls into LLMs spans three critical areas: API collection and search, tool assistant strategy, and performance evaluations.

API Collection and Search. APIBench (Patil et al., 2023), featuring APIs from HuggingFace, TorchHub, and TensorHub, assesses its proficiency. ToolBench (Qin et al., 2023b) features 16000+ real-world APIs across 49 categories from RapidAPI Hub, and develops a depth-first search decision tree (DFSDT), improving LLMs’ search and reasoning capabilities.

Tool Assistant Strategy. SelfAsk (Press et al., 2022) simplifies tasks into sub-questions for tool invocation, akin to DemonstrateSearch-Predict (Khattab et al., 2023). Similarly, Toolformer (Schick et al., 2023), ART (Paranjape et al., 2023), and others (Gao et al., 2023; Lyu et al., 2023; Chen et al., 2022) using specific tokens to guide tool usage, halting to invoke tools, and incorporating their outputs for continued generation. But they only focus on tool-usage tools, and can not apply to real-world scenarios.

Tool Usage Evaluation. Jacovi et al. (2023) focuses on mathematical reasoning and reveals the influence of tool use versus non-use is less pronounced in larger LLMs compared to smaller ones. MetaTool (Huang et al., 2023) assesses LLMs’ decisions on whether to utilize external tools and which tool to use, but does not address the effects of incorrect or unnecessary tool usage.

Different from previous works, WTU-Eval aims to bridge this gap by investigating whether models recognize the need for tool use in real-world scenarios and how improper tool integration might affect the foundational efficiency of LLMs, as detailed in §3.

3 The WTU-Eval Benchmark

The desired diagram of WTU-Eval is shown in Figure 2. In R1, the user asks a real-time question, but LLM cannot access this information without the search engine, so it fails to answer. In R2, when faced with the same question, LLM has access to tool pools and knows that the tool usage is necessary, so it decides to call *Search Engine* to find the real-time information and gives the correct answer. In R3, the user asks a general question, and LLM answers it with its knowledge. In R4, when presented with the same question, the LLM can access tool pools. Recognizing that tool usage is unnecessary, it decides to provide an answer directly.

By comparing the results between R1 and R2, we can determine whether LLMs recognize when a question exceeds their capabilities and thus requires the use of tools, and quantify the impact of using tools. By comparing the results of R3 and R4, we can determine whether the LLMs, when given the option to use tools, recognize that the current question can be answered without tools. Additionally, we can quantify the damage when they choose to use tools unnecessarily.

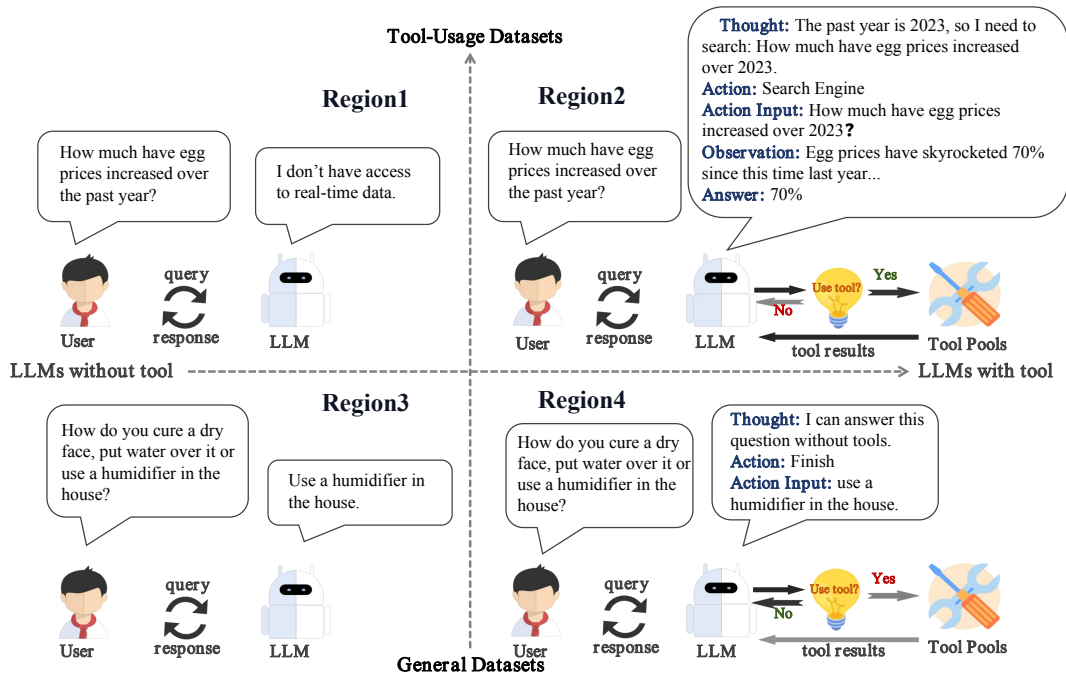


Figure 2: Illustrative diagram depicting user interaction scenarios with and without access to tool pools. LLMs need to respond to the user’s query in *Region1* (R1) and *Region3* (R3). In *Region2* (R2) and *Region4* (R4), LLMs must judge based on the nature of the task whether a tool is required. If so, the corresponding tool from the tool pool is invoked; if not, the answer is provided using its knowledge. If the judgment is correct, then the corresponding choice is highlighted in **green**; otherwise, it is in **red**.

3.1 Evaluation Settings

We show **WTU-Eval** settings from datasets, tool pools, LLMs, and evaluation metrics.

Datasets. We partition the datasets into the tool datasets (for tasks requiring specific tools), and the general datasets (for tasks solvable with LLMs’ own ability). The tool datasets include MLQA (Lewis et al., 2019), ASDiv (Miao et al., 2021), GSM8K (Cobbe et al., 2021), MathQA (Amini et al., 2019), HotpotQA (Yang et al., 2018), and RealtimeQA (Kasai et al., 2022), focusing on machine translation, math reasoning, Wikipedia search, and web search. The general datasets contain BoolQ (Clark et al., 2019), RACE (Lai et al., 2017), PIQA (Bisk et al., 2020), RTE (Dagan et al., 2005), and HellaSwag (Zellers et al., 2019), focusing on reading comprehension, commonsense reasoning, and sentence completion. More details about the datasets are discussed in the Appendix A.

Tool Pools. Following BMTools (Qin et al., 2023a), we select the tools used in the evaluation, where machine translator and calculator are single-action tools, and search engine and Wikipedia search are multiple-action tools.

- **Machine Translator:** We select Baidu Translator¹, as a current mainstream translation API with good performance, for testing.
- **Calculator:** We choose the WolframAlpha API² as our calculator.
- **Search Engine:** We choose the Bing Search³ API as the web search tool for LLMs to browse current events, fiction stories, history facts, etc.
- **Wikipedia Search:** Besides a simple Wikipedia API, as WikiSearch and WikiLoadPage are designed, we define an additional action – WikiDisambiguation. When the search entity cannot return the expected result, the model can access the interface to get a similar entity to the current search result and self-correct the search parameters.

LLMs. We test LLMs from both commercial and open-source sectors for a broad evaluation, including Text-Davinci-003, ChatGPT (0613), Llama2,

¹<https://fanyi-api.baidu.com/?fr=pcHeader>

²<https://developer.wolframalpha.com/>

³<https://www.microsoft.com/en-us/bing/apis/bing-web-search-api>

ChatGLM3-6B, and Zephyr-7B. ChatGLM3-6B is notable for its unique agent-tuning with tool interaction insights. Zephyr-7B, evolved from Mistral-7B, employs Direct Distilled Preference Optimization (DPO) to better align with user preferences in language tasks.

Evaluation Metrics. In WTU-Eval, we prioritize accuracy using advanced methods beyond exact matches, categorizing datasets into numerical and free-text responses. We check numerical answers with specific data and transform free-text responses into labels. For example, in PIQA (which provides two solutions for a given task), we label these two solutions as 1 and 2. When we cannot match labels or text content, we manually check the responses.⁴

Additionally, tool usage is marked incorrect in the general dataset and a correct example is shown in Figure 2 R4. To balance comparisons, we introduce the Call Rate, considering the initial use of a tool as a call, ensuring a thorough evaluation.

3.2 Evaluation Prompt

In WTU-Eval, we utilize ReACT (Yao et al., 2022) for zero-shot and few-shot experiments in scenarios with access to tool pools (R2 and R4). The ReACT is structured into four stages: *Thought*, *Action*, *Observation*, and *Final Answer*, performed in a limited loop. To ensure fairness, all LLMs are evaluated under the same settings during the assessment. Further details on prompts are provided in the Appendix D.

Zero-shots. We introduce tool names, descriptions, and parameters to guide the LLMs to use tools.

Few-shots. We introduce tool names, descriptions, parameters, and examples of: a) one tool usage scenario, and b) one general scenario where tools are not used.

4 Experiments and Results

4.1 LLMs' Performance in Tool Datasets

When LLMs can determine whether to use tools and LLMs' ability is similar to ChatGPT, their performance in tool-usage datasets improves.

⁴If both solutions are deemed unsuitable: "answer": "Neither solution is suitable"
If the model discusses both solutions: "answer": "Solution 1 is..., solution 2 is..., I think solution 1 is better"

In Table 1 R2, when LLMs have access to tools, Llama2-13B's zero-shot performance on most tool questions drops to 0, while ChatGPT and Text-Davinci-003 exhibit significant improvements (by up to 25% in GSM8K), exceeding their performance in R1. It is observed that the use of tools does not unconditionally enhance LLMs' performance and the enhancement depends on LLMs' ability. Considering the scale gap between ChatGPT, Text-Davinci-003, and Llama2, we believe that properly using tools demands models' ability to deal with complex and extensive tool prompts without demonstrations.

This trend alters a little with the adoption of the few-shot methodology. In R2, ChatGPT and Text-Davinci-003's performance also improve (by up to 40% in GSM8K) with the few-shot setting, exceeding their performance in R1. In contrast, Llama2 only shows improvement on a small portion of tool datasets, with performance declining on the rest compared to R1. It is concluded that the efficacy of tool invocation in augmenting performance is contingent upon the ability of the model.

4.2 Impact of Different Tools on LLMs' Performance in Tool Datasets

In most tool-usage datasets, the proficiency of LLMs diminishes as the complexity of tools increases. In Table 1 R2, we especially introduce the Translator to MLQA, the Calculator to ASDiv, GSM8K, and MathQA, the Search Engine to Real-timeQA, and the Wikipedia Search to HotPotQA.

Tool usage impact is closely linked to tool complexity. LLMs efficiently manage translation tasks due to the Translator's simplicity. However, when faced with complex tools like the WolframAlpha Calculator, Llama2's performance drops significantly. Similarly, tasks using BingSearch and WikipediaSearch see only modest improvements due to more complex tool instructions, particularly in the few-shot setting across all LLMs.

Moreover, the few-shot setting remarkably outperforms the zero-shot setting, with improvements reaching up to 76% in some cases. In zero-shot settings, such as ChatGPT's use of a range of tools from Translator to Wikipedia Search, there is a clear trend: as tool's complexity increases, LLMs' proficiency decreases. This indicates that tasks requiring a deeper understanding of tool usage present more significant challenges for LLMs, underscoring the increased interpretive burden in nav-

Test Set	Model w/o Tool						Model w/ Tool					
	T003	ChatGPT	Llama2-13B		Llama2-7B		T003	ChatGPT	Llama2-13B		Llama2-7B	
			Base	Chat	Base	Chat			Base	Chat	Base	Chat
Tool datasets			R1				R2					
MLQA	54.17	53.13	52.08	57.29	55.21	62.50	58.33	50.00	0.00	12.50	1.04	11.45
							70.83	65.62	26.04	60.41	50.00	48.95
ASDiv	48.67	79.33	13.00	50.00	23.00	45.67	70.66	83.00	9.00	23.66	46.66	38.66
							68.33	83.00	9.00	45.00	43.00	47.66
GSM8K	14.00	67.00	2.00	9.00	9.00	12.00	39.00	58.00	0.00	20.00	2.00	8.00
							52.00	53.00	0.00	15.00	14.00	5.00
MathQA	33.00	18.00	12.00	17.00	19.00	26.00	37.00	39.00	8.00	11.00	4.00	11.00
							39.00	36.00	6.00	12.00	10.00	5.00
RealtimeQA	36.66	40.00	20.00	30.00	23.34	40.00	56.66	36.66	0.00	23.33	3.30	40.00
							36.66	40.00	0.00	33.33	26.66	26.66
HotPotQA	33.50	34.50	11.50	33.00	20.00	36.00	28.50	39.00	0.00	18.00	0.00	20.50
							47.95	41.50	5.00	26.50	18.50	20.50
General datasets			R3				R4					
BoolQ	79.00	89.00	56.00	46.00	46.00	57.00	20.00	6.00	0.00	0.00	0.00	2.00
							58.50	76.25	54.00	61.25	55.00	32.50
RACE	68.96	79.09	14.80	22.93	32.00	33.87	6.00	30.00	0.00	0.00	14.00	0.00
							82.93	77.46	62.05	52.53	58.40	50.40
PIQA	58.00	84.00	16.00	32.00	25.00	49.00	3.00	39.00	0.00	0.00	0.00	0.00
							50.25	58.75	27.25	43.50	18.75	31.00
RTE	59.00	78.00	68.00	63.00	54.00	58.00	3.00	12.00	0.00	0.00	0.00	0.00
							66.25	50.00	13.50	34.50	45.50	23.25
HellaSwag	54.00	75.00	21.00	49.00	54.00	44.00	23.00	28.00	0.00	0.00	1.00	0.00
							50.75	50.00	4.25	23.25	20.00	15.75

Table 1: The accuracy of experiments executed general datasets and tool datasets whether or not have access to tool pools, where "T003" means "Text-Davinci-003", and " " indicates few-shot results, while cells without background color indicate zero-shot results.

igating tool-specific instructions.

4.3 LLMs' Performance in General Datasets

LLMs' performance in general datasets declines when they can determine whether to use tools, indicating LLMs do not know their ability boundary. By comparing R4 to R3 in Table 1, we can observe that LLMs' performance decreases in all general datasets. Analyzing the incorrect answer, we note that LLMs tend to use tools, and due to wrong tool invocation, their performance declines. The whole incorrect answer study will be discussed in §5.2.

Table 1 R4 demonstrates a significant reduction in zero-shot performance when accessing tools compared with R3, particularly evident in Llama2. Notably, the most substantial decrease observed is 83% in BoolQ (Text-Davinci-003), and Llama2's performance nearly falls to 0. This is primarily due to LLMs' frequent misuse of tools in general queries. Error analysis §5.2 suggests that the com-

plexity of following tool instructions complicates the adherence to the ReACT framework, thus impacting the *Thought* process.

To mitigate this issue, we add demonstrations, leading to the few-shot results in R4. These experiments largely echo the zero-shot findings, but slight improvements are observed compared with R3. Importantly, this increase is mainly observed in Llama2, which shows a 10% to 30% improvement. We believe the demonstrations not only inspire LLMs' Chain of Thought (COT) ability but also correct their response formats.

To further explore the impact of the COT and ReACT's format, we conduct few-shot trials in R3, focusing on the COT process depicted in Figure 3. The results, as shown in Table 2, reveal that COT significantly aids smaller-scale Llama models. However, for larger models such as ChatGPT, COT does not lead to improvements and might even result in performance declines on BoolQ and HellaSwag.

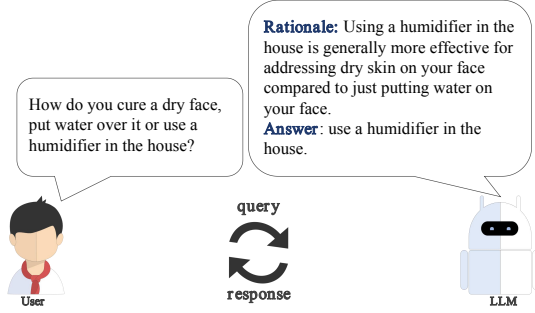


Figure 3: Illustrative diagram depicting user interaction scenarios with LLMs in COT setting without the integration of a tool set.

Test set	ChatGPT		Llama2-7B		Llama2-7B-Chat	
	Zero-shot	COT	Zero-shot	COT	Zero-shot	COT
BoolQ	89.00	81.00	46.00	74.00	57.00	51.00
RACE	79.09	83.47	60.30	65.07	33.87	67.73
PIQA	84.00	86.00	25.00	56.00	49.00	54.00
RTE	78.00	78.00	54.00	57.00	58.00	47.00
HellaSwag	75.00	66.00	54.00	24.00	44.00	39.00

Table 2: Accuracy in general datasets without tool access in COT and zero-shot settings.

4.4 Impact of Different Tools on LLMs' Performance in General Datasets

In general datasets, LLMs experience a decline in performance when various external tools are introduced, with the impact order being (Wikipedia Search, Search Engine) more significant than (Translator, Calculator). We test LLMs by introducing all tools and tool usage instructions. In the zero-shot setting, the collective impact of the tool pool is evaluated by introducing information for all tools simultaneously in the instructions. In the few-shot setting, due to the input length constraints of LLMs, we present each tool's name, description, and demonstration in individual prompts. The average accuracy across all tools reflects the cumulative effect of the tool pools.

As illustrated in Table 3, it is observed that due to the increased complexity introduced by the addition of tool instruction, zero-shot performance is considerably lower than few-shot, and reveals that different tool demonstrations have varying effects on the model's performance in general datasets, with the order of impact being (Wikipedia Search, Search Engine) more pronounced than (Translator, Calculator). The complete experiment of all general datasets is shown in Appendix E.

LLM	MT	Cal	SE	Wiki	All
T003	56.00	47.00	31.00	42.00	20.00
	68.00	83.00	45.00	38.00	58.50
ChatGPT	11.00	5.00	8.00	8.00	6.00
	80.00	85.00	70.00	70.00	76.25
Llama2-7B-Base	0.00	0.00	0.00	0.00	0.00
	64.00	59.00	41.00	56.00	54.00
Llama2-7B-Chat	0.00	0.00	0.00	0.00	0.00
	45.00	42.00	9.00	34.00	32.50
Llama2-13B-Base	0.00	0.00	0.00	0.00	0.00
	62.00	56.00	46.00	52.00	55.00
Llama2-13B-Chat	0.00	0.00	1.00	0.00	0.00
	77.00	52.00	60.00	56.00	32.50
Zephyr-7B	35.00	33.00	35.00	34.00	17.00
	52.00	8.00	53.00	77.00	47.50
ChatGLM3-6B	10.00	7.00	8.00	18.00	20.00
	31.00	43.00	23.00	14.00	27.75

Table 3: Detailed Results of BoolQ Experiment: Performance of each LLM in few-shot and zero-shot settings, where *MT* means *Machine Translator*, *Cal* means *Calculator*, *SE* means *Search Engine*, *Wiki* means *Wikipedia Search*, and " " indicates few-shot results, while cells without background color indicate zero-shot results.

5 Discussion

For a deeper understanding of the results, we explore different finetuning methods' impacts, conduct the error analysis, and make some improvements on WTU-Eval.

5.1 Quantitative Results on Different Fine-tuning Methods

We conduct R2 and R4's evaluation in ChatGLM3-6B and Zephyr-7B for their different fine-tuning methods from Llama2.

Poor Performance of ChatGLM3-6B in Decision on Whether to Use the Tool. The above analysis reveals that LLMs' performance drops in general datasets with tool access but improves in tool usage datasets, dependent on LLM ability. LLMs lacking effective tool-usage training show weaker decision-making on tool employment. Thus, we experiment with ChatGLM3-6B in general and tool-usage datasets, leveraging its agent tuning for better tool usage.

As illustrated in Table 5, compared with the Llama2-7B, ChatGLM3-6B shows superior performance in the zero-shot settings, indicating the validity of agent tuning. However, the results also show its poor performance in the decision on whether

Error Type	Example	Analysis
Incorrect or Unnecessary Tool Invocation	Passage: ...People often get take-away meals—they buy the food outside Question: When they get a takeaway meal, they often eat it ()? Options: A. at home B. in the school C. outside D. in the bars Thought: There is no answer in the passage. Action: BingSearch Action Input: Where do people often eat takeaway meals?	The LLMs’ judgment on when to invoke tools is inaccurate, often opting for tool usage in the first <i>Thought</i> step.
NULL Content	I have 3 hundreds, 8 tens, and 3 ones. What number am I? Thought: I should try to find a way to convert the numbers into a number that I can add. Action:	<i>Action</i> is null. Insufficient training in framework comprehension leads to failures in producing the required content.
Call Correct Tool but Not Reasoning	Jerry was weighing boxes for moving. The first was 2 pounds, the second was 11 pounds and the last was 5 pounds. What is the combined weight of all three boxes? Thought: I need to calculate the combined weight of the boxes. Action: Calculator Action Input: 2 + 11 + 5 Observation: {'@title': 'Result', '@primary': 'true', 'subpod': {'@title': '', 'img': {'@src': '{given link}', '@alt': '18', '@title': '18', '@contenttype': 'image/gif'}} Thought: I need to calculate the combined weight of the boxes.	The model pays more attention to tool usage rather than understanding the returned observation and moving to the next state.
Call Invalid Tool Repeatedly After Reminding	14 red plums and four green plums are in the basket. How many plums are in the basket? Thought: I should use a calculator for Addition to finish this task. Action: Calculator Action Input: 14 (red plums) + 4 (green plums) Observation: WolframAlpha API cannot parse the input query. Thought: Hmm, I think I need to separate the red and green plums. Action: Calculator Action Input: 14 (red plums) + 4 (green plums)	The model cannot comprehend the instruction and forgets how to consistently use the given tools.
Endless Retry	The McCarran Airport Connector serves what airport named after late U.S. Senator Pat McCarran? Thought: I need to find information about the McCarran Airport Connector and the airport it serves, which is named after a late U.S. Senator. Action: WikiSearch Action Input: McCarran Airport Connector Observation: State Route 171. The Harry Reid Airport Connector is ... Thought: I have found information about the McCarran Airport Connector, but I need to find information about the airport it serves, which is named after a late U.S. Senator. Action: WikiSearch Action Input: Pat McCarran Observation: Patrick Anthony McCarran (August 8, 1876 – September 28, 1954) ... Thought: ...	The model keeps using tools in a loop without reaching a conclusion and needs to improve its ability to give up and restart when stuck in a loop.

Table 4: Examples and analysis of five error types encountered in the failure cases.

Test Set	ChatGLM3-6B		Llama2-7B	
	Zero-Shot	Few-Shot	Zero-Shot	Few-Shot
Tool Datasets				
MLQA	14.58	42.70	1.04	50.00
ASDiv	38.33	52.66	43.66	43.00
GSM8K	14.00	26.00	2.00	14.00
MathQA	13.00	9.00	4.00	10.00
RealtimeQA	13.33	23.33	3.30	26.66
HotPotQA	4.00	11.50	0.00	18.50
General Datasets				
BoolQ	20.00	27.75	0.00	55.00
RACE	22.00	66.06	14.00	58.40
PIQA	3.00	19.00	0.00	18.75
RTE	35.00	46.25	0.00	45.00
HellaSwag	11.00	23.25	1.00	20.00

Table 5: Accuracy of ChatGLM3-6B with access to tools in tool-usage datasets and general datasets.

Test set	Model w/o Tool	Model w/ Tool	
	Zero-Shot	Zero-Shot	Few-Shot
MLQA	46.15	39.58	33.33
AsDiv	64.00	46.33	0.00
GSM8K	18.00	35.00	7.00
MathQA	21.00	7.00	5.00
RealtimeQA	50.00	23.33	10.00
HotpotQA	30.00	18.00	16.23
BoolQ	76.00	17.00	47.50
RACE	75.46	17.00	59.40
PIQA	68.00	4.00	15.75
RTE	66.00	1.00	15.50
HellaSwag	31.00	4.00	15.50

Table 6: Accuracy of Zephyr-7B in tool usage and general datasets with and without tool access.

to use the tool for similar results to Llama2-7B in most general datasets, which indicates that current tool training methods have not adequately addressed the question of whether to invoke a tool.

Counterintuitive Results in Zephyr-7B. In our study, Zephyr-7B exhibits unique performance trends compared to other LLMs, particularly underlined by a decrease in efficacy when using tools in tool datasets, as detailed in Table 6. Its few-shot

394
395
396
397
398

performance falls short of its zero-shot capabilities, a discrepancy most evident in calculator-involved tasks, dropping to as low as 0 in ASDiv. Moreover, within general datasets, the calculator’s negative impact on Zephyr-7B is notably worse than that of other tools (§4.4), resulting in a 38.5% lower average accuracy. Analysis of errors in these datasets reveals a recurrent issue: Zephyr-7B frequently misapplies the calculator in calculation-related tasks, leading to response inaccuracies.

Test Set	Model	MT	Cal	SE	Wiki	All
BoolQ	Baseline(Acc)	45.00	42.00	9.00	34.00	32.50
	Ours(Acc)	45.00	53.00	44.00	37.00	44.75
	Baseline(CR)	0.00	4.00	41.00	10.00	13.75
	Ours(CR)	0.00	1.00	3.00	1.00	1.25
PIQA	Baseline(Acc)	51.00	14.00	0.00	10.00	18.75
	Ours(Acc)	37.00	47.00	40.00	43.00	41.75
	Baseline(CR)	2.00	80.00	53.00	11.00	36.50
	Ours(CR)	0.00	6.00	2.00	1.00	2.25
HellaSwag	Baseline(Acc)	19.00	21.00	21.00	19.00	20.00
	Ours(Acc)	25.00	28.00	30.00	25.00	27.00
	Baseline(CR)	0.00	0.00	6.00	14.00	5.00
	Ours(CR)	0.00	4.00	0.00	1.00	1.25

Table 7: Performance on general datasets improves through SFT of Llama2-7B model, where CR means Call Rate.

5.2 Error Analysis

In this section, we make a deep sampling of the failure cases. Besides the wrong answer, we set five error types of these cases, show examples and analysis in Table 4, and more details in Appendix C.

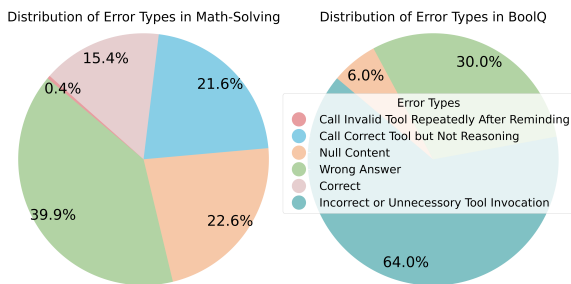


Figure 4: Distribuion of Error Types in Tool-Usage and General Datasets with Zero-Shot Setting in Llama2-7B

The proportions of these error types vary across different models, datasets, and settings. For instance, Figure 4 shows the distribution of error types in math-solving questions (ASDiV, GSM8K and MathQA) and commonsense reason questions (BoolQ) with Llama2-7B in R2 and R4. It can be inferred that incorrect/unnecessary tool invocation

is preferred to appear in general datasets, while the other error types about tool invocation steps appear in tool-usage datasets.

5.3 Supervised Fine-Tuning for Tool-Usage Decision-Making

Based on our findings, LLMs’ indecision on tool usage not only undermines their overall performance but also adversely affects their effectiveness on general datasets. To mitigate this, we curate a specialized dataset with a size of 4000 from the general datasets’ training sets. Based on observation of step *Thought*’s importance for the decision on tool usage in Table 4, we train the first *Thought* and second *Action* steps, aiming at improving decision-making ability regarding tool usage. We apply GPT-4 to generate the first *Thought* step and select the correct action for the general questions.

After supervised fine-tuning, Llama2-7B’s performance improves by an average of 14%, and incorrect tool use drops by 16.8% in general datasets. Specifically, in the PIQA, accuracy in the *Search Engine* improves by 40%, and the *Calculator* call rate decreases by 74%, as detailed in Table 7.

6 Conclusion

In this paper, we explore whether LLMs can discern their ability boundaries and use tools flexibly. We introduce the WTU-Eval to assess LLMs with eleven datasets and four tools. The results of WTU-Eval reveal that LLMs frequently struggle to determine tool use in general datasets, and their performance in tool-usage datasets improves when their ability is similar to ChatGPT. In both datasets, incorrect tool usage significantly impairs LLMs’ performance. After detailed analysis, we also introduce a dataset focused on improving decision-making in tool usage, which successfully enhances Llama2-7B’s performance and reduces unnecessary tool invocations.

Our work points out the overlooked shortcomings in tool usage by LLMs, i.e., they struggle to recognize their capability boundaries and lack of tool usage decision-making capability. We use the WTU-Eval to test eight LLMs, which is the first benchmark to evaluate whether LLMs can use tools accurately. Future works include adding more datasets and tools, and testing more types of LLMs.

468 **Limitations**

469 This study’s limitations arise from computational
470 constraints, limiting our model selection to exclude
471 larger variants like Llama2-70B, and from the mod-
472 els’ slow processing of tool directives, leading us to
473 evaluate a sampled subset of the test set, potentially
474 causing result discrepancies with other studies.

475 **Ethics Statement and Broader Impacts**

476 This study exclusively utilized datasets and toolsets
477 that are publicly available and previously published,
478 ensuring they contain no offensive or harmful con-
479 tent. We rigorously adhere to ethical standards,
480 including a thorough review of materials to safe-
481 guard privacy and integrity.

482 This study is pivotal for the practical application
483 of LLMs, as it aims at reducing unnecessary tool
484 invocations, thereby enhancing the efficiency of
485 tool usage. This optimization in tool interaction
486 not only advances the development of AI but also
487 ensures more effective and streamlined AI oper-
488 ations, leading to smarter and more efficient AI
489 systems that better serve the needs across different
490 sectors and research disciplines.

491 **References**

492 Aida Amini, Saadia Gabriel, Peter Lin, Rik Koncel-
493 Kedzioriski, Yejin Choi, and Hannaneh Hajishirzi.
494 2019. Mathqa: Towards interpretable math word
495 problem solving with operation-based formalisms.
496 *arXiv preprint arXiv:1905.13319*.

497 Yonatan Bisk, Rowan Zellers, Jianfeng Gao, Yejin Choi,
498 et al. 2020. Piqa: Reasoning about physical com-
499 monsense in natural language. In *Proceedings of the*
500 *AAAI conference on artificial intelligence*, volume 34,
501 pages 7432–7439.

502 Wenhui Chen, Xueguang Ma, Xinyi Wang, and
503 William W Cohen. 2022. Program of thoughts
504 prompting: Disentangling computation from reason-
505 ing for numerical reasoning tasks. *arXiv preprint*
506 *arXiv:2211.12588*.

507 Christopher Clark, Kenton Lee, Ming-Wei Chang,
508 Tom Kwiatkowski, Michael Collins, and Kristina
509 Toutanova. 2019. Boolq: Exploring the surprising
510 difficulty of natural yes/no questions. *arXiv preprint*
511 *arXiv:1905.10044*.

512 Karl Cobbe, Vineet Kosaraju, Mohammad Bavarian,
513 Mark Chen, Heewoo Jun, Lukasz Kaiser, Matthias
514 Plappert, Jerry Tworek, Jacob Hilton, Reiichiro
515 Nakano, et al. 2021. Training verifiers to solve math
516 word problems. *arXiv preprint arXiv:2110.14168*.

517 Ido Dagan, Oren Glickman, and Bernardo Magnini.
518 2005. The pascal recognising textual entailment chal-
519 lenge. In *Machine learning challenges workshop*,
520 pages 177–190. Springer.

521 Luyu Gao, Aman Madaan, Shuyan Zhou, Uri Alon,
522 Pengfei Liu, Yiming Yang, Jamie Callan, and Gra-
523 ham Neubig. 2023. Pal: Program-aided language
524 models. In *International Conference on Machine*
525 *Learning*, pages 10764–10799. PMLR.

526 Yue Huang, Jiawen Shi, Yuan Li, Chenrui Fan, Siyuan
527 Wu, Qihui Zhang, Yixin Liu, Pan Zhou, Yao Wan,
528 Neil Zhenqiang Gong, et al. 2023. Metatool bench-
529 mark for large language models: Deciding whether
530 to use tools and which to use. *arXiv preprint*
531 *arXiv:2310.03128*.

532 Alon Jacovi, Avi Caciularu, Jonathan Herzig, Roei
533 Aharoni, Bernd Bohnet, and Mor Geva. 2023. A
534 comprehensive evaluation of tool-assisted generation
535 strategies. *arXiv preprint arXiv:2310.10062*.

536 Jungo Kasai, Keisuke Sakaguchi, Yoichi Takahashi,
537 Ronan Le Bras, Akari Asai, Xinyan Yu, Dragomir
538 Radev, Noah A Smith, Yejin Choi, and Kentaro Inui.
539 2022. Realtime qa: What’s the answer right now?
540 *arXiv preprint arXiv:2207.13332*.

541 Guokun Lai, Qizhe Xie, Hanxiao Liu, Yiming Yang,
542 and Eduard Hovy. 2017. Race: Large-scale reading
543 comprehension dataset from examinations. *arXiv*
544 *preprint arXiv:1704.04683*.

545 Patrick Lewis, Barlas Oğuz, Ruty Rinott, Sebastian
546 Riedel, and Holger Schwenk. 2019. Mlqa: Eval-
547 uating cross-lingual extractive question answering.
548 *arXiv preprint arXiv:1910.07475*.

549 Lei Li, Yekun Chai, Shuohuan Wang, Yu Sun, Hao Tian,
550 Ningyu Zhang, and Hua Wu. 2023. Tool-augmented
551 reward modeling. *arXiv preprint arXiv:2310.01045*.

552 Qing Lyu, Shreya Havaldar, Adam Stein, Li Zhang,
553 Delip Rao, Eric Wong, Marianna Apidianaki, and
554 Chris Callison-Burch. 2023. Faithful chain-of-
555 thought reasoning. *arXiv preprint arXiv:2301.13379*.

556 Shen-Yun Miao, Chao-Chun Liang, and Keh-Yih Su.
557 2021. A diverse corpus for evaluating and developing
558 english math word problem solvers. *arXiv preprint*
559 *arXiv:2106.15772*.

560 OpenAI. 2023. [Gpt-4 technical report](#). 560

561 Bhargavi Paranjape, Scott Lundberg, Sameer Singh,
562 Hannaneh Hajishirzi, Luke Zettlemoyer, and
563 Marco Tulio Ribeiro. 2023. Art: Automatic multi-
564 step reasoning and tool-use for large language mod-
565 els. *arXiv preprint arXiv:2303.09014*.

566 Shishir G Patil, Tianjun Zhang, Xin Wang, and
567 Joseph E Gonzalez. 2023. Gorilla: Large language
568 model connected with massive apis. *arXiv preprint*
569 *arXiv:2305.15334*.

Ofir Press, Muru Zhang, Sewon Min, Ludwig Schmidt, Noah A Smith, and Mike Lewis. 2022. Measuring and narrowing the compositionality gap in language models. *arXiv preprint arXiv:2210.03350*.

Yujia Qin, Shengding Hu, Yankai Lin, Weize Chen, Ning Ding, Ganqu Cui, Zheni Zeng, Yufei Huang, Chaojun Xiao, Chi Han, et al. 2023a. Tool learning with foundation models. *arXiv preprint arXiv:2304.08354*.

Yujia Qin, Shihao Liang, Yining Ye, Kunlun Zhu, Lan Yan, Yaxi Lu, Yankai Lin, Xin Cong, Xiangru Tang, Bill Qian, et al. 2023b. Toolllm: Facilitating large language models to master 16000+ real-world apis. *arXiv preprint arXiv:2307.16789*.

Jingqing Ruan, Yihong Chen, Bin Zhang, Zhiwei Xu, Tianpeng Bao, Guoqing Du, Shiwei Shi, Hangyu Mao, Xingyu Zeng, and Rui Zhao. 2023. Tptu: Task planning and tool usage of large language model-based ai agents. *arXiv preprint arXiv:2308.03427*.

Timo Schick, Jane Dwivedi-Yu, Roberto Dessì, Roberta Raileanu, Maria Lomeli, Luke Zettlemoyer, Nicola Cancedda, and Thomas Scialom. 2023. Toolformer: Language models can teach themselves to use tools. *arXiv preprint arXiv:2302.04761*.

Zhilin Yang, Peng Qi, Saizheng Zhang, Yoshua Bengio, William W Cohen, Ruslan Salakhutdinov, and Christopher D Manning. 2018. Hotpotqa: A dataset for diverse, explainable multi-hop question answering. *arXiv preprint arXiv:1809.09600*.

Shunyu Yao, Jeffrey Zhao, Dian Yu, Nan Du, Izhak Shafran, Karthik Narasimhan, and Yuan Cao. 2022. React: Synergizing reasoning and acting in language models. *arXiv preprint arXiv:2210.03629*.

Rowan Zellers, Ari Holtzman, Yonatan Bisk, Ali Farhadi, and Yejin Choi. 2019. Hellaswag: Can a machine really finish your sentence? *arXiv preprint arXiv:1905.07830*.

Yuchen Zhuang, Yue Yu, Kuan Wang, Haotian Sun, and Chao Zhang. 2023. Toolqa: A dataset for llm question answering with external tools. *arXiv preprint arXiv:2306.13304*.

A Hardware Configuration and Datasets

In this study, the hardware configuration comprises an NVIDIA GeForce RTX 3090 GPU with 20 GB of memory for a tool-usage task that costs several minutes, and an NVIDIA A100 GPU with 80 GB of memory for the fine-tuning task, requiring six hours to complete nine epochs tuning.

Table 8 outlines each dataset’s size, action type, and tool call steps. Test sizes are chosen based on reasoning complexity and time costs, leading to random sampling from original test sets, which may

Test Set	Test Size	Action Type	Call Step
Tool Dataset			
MLQA(2019)	96	Single	Single
ASDiv(2021)	300	Single	Multiple
GSM8K(2021)	100	Single	Multiple
MathQA(2019)	100	Single	Multiple
HotpotQA(2018)	200	Multiple	Multiple
RealtimeQA(2022)	120	Multiple	Multiple
General Dataset			
BoolQ(2019)	100	-	-
RACE(2017)	100	-	-
PIQA(2020)	100	-	-
RTE ⁵	100	-	-
HellaSwag(2019)	100	-	-

Table 8: Details of all datasets evaluated in WTU-Eval, include the type of tool action, and tool call steps involved. Action Type and Tool Call Step are not listed for general datasets as they theoretically involve only a single step.

cause different results from some public benchmarks. Action types relate to tool interactions, and call steps indicate if a task requires a single or multiple tool call step.

B Examples of Successful Types

• Correct Tool Usage Example of Calculator

Task Description:

Given a context and related question, try your best to answer the question. · Context:

Being his favorite, he saved checking on the grapevines for his last stop. He was told by one the pickers that they fill 324 drums of grapes per day.

·Question: How many drums of grapes would be filled in 9 days?

Thought: I should use calculator for Multiplication to finish this task.

Action: Calculator

Action Input: 324*9

Observation: {'@title': 'Input', 'subpod': {'@title': '', 'img': {'@src': '{given link}', '@alt': '324 * 9', '@title': '324 * 9', '@contenttype': 'image/gif'}, 'plaintext': '324 * 9'}}, {'@title': 'Result', '@primary': 'true', 'subpod': {'@title': '', 'img': {'@src': '{given link}', '@alt': '2916', '@title': '2916', '@contenttype': 'image/gif'}, 'plaintext': '2916'}, 'states': {'@count': '1', 'state': {'@name': 'Step-by-step solution', '@input': 'Result_Step-by-step solution', '@stepbystep': 'true', '@buttonstyle': 'StepByStepSolution'}}}

Thought: I think I got the final answer.

Action: Finish

656	<i>Action Input:</i> 2916 (drums)	the third time in its history due to a major	705
657	<i>Final Answer:</i> 2916 (drums)"	storm?	706
658		· Option:['Greece', 'New Zealand', 'Mexico',	707
		'Finland']	708
659	• Correct Tool Usage Example of Machine	<i>Thought:</i> To answer the question, I can	709
660	Translator	use the BingSearch API to search for news	710
661	Task Description:	articles specifically about the given options:	711
662	Given an English context and related Arabic	Greece, New Zealand, Mexico, and Finland,	712
663	question (you can translate the question into	along with keywords like national state of	713
664	English), answer the question in English.	emergency and major storm:	714
665	English Context: In 1884, a paper was signed	<i>Action:</i> BingSearch	715
666	at Togoville with the King Mlapa III, whereby	<i>Action Input:</i> 'Greece national state of	716
667	Germany claimed a protectorate over a stretch	emergency major storm'	717
668	of territory along the coast and gradually ex-	<i>Observation:</i> page: 1	718
669	tended its control inland. Its borders were	title: Storm Elias: Greece hit by several	719
670	defined after the capture of hinterland by	months of rain in a day, just ... - CNN	720
671	German forces and signing agreements with	summary: CNN Storm Elias dumped several	721
672	France and Britain. In 1905, this became the	months-worth of rain in less than a day as	722
673	German colony of Togoland. The local pop-	it swept across Greece on Wednesday, just	723
674	ulation was forced to work, cultivate cotton,	weeks after Storm Daniel killed 17 people in	724
675	coffee and cocoa and pay high taxes. A rail-	the southern European nation.	725
676	way and the port of Lomé were built for	page: 2	726
677	export of agricultural products. The Germans	title: Greek Government Holds Emergency	727
678	introduced modern techniques of cultivation	Meeting As Dangerous Storm Front Set ...	728
679	of cocoa, coffee and cotton and developed the	summary: An emergency weather forecast	729
680	infrastructure.	issued on Friday by the Hellenic National	730
681	· Arabic Question:	Meteorological Service (EMY) speaks of	731
682	متى تم توقيع الأوراق؟	heavy rainfall and storms over western,	732
683		central and northern Greece starting Monday,	733
684	<i>Thought:</i> I need to translate the question	before spreading to more areas by mid-next	734
685	متى تم توقيع الأوراق؟	week. Tags: greece, Greek, Greek news,	735
686		greek storm, Vassilis Kikilias Previous Article	736
687		page: 3	737
688	<i>Action:</i> Translator	title: Greek floods and fires expose Europe's	738
689	<i>Action Input:</i>	frail climate defences	739
690	متى تم توقيع الأوراق؟	summary: Following the previous major	740
691		storm that flooded Thessaly in 2020, Greece's	741
692	<i>Observation:</i> When were these documents	conservative government promised to prevent	742
693	signed?	a repeat of the disaster. ... Greece's emer-	743
694	<i>Thought:</i> I think I got the final answer	gency plan issued in October 2022 ...	744
695	<i>Action:</i> Finish	<i>Thought:</i> Based on the search results, Greece	745
696	<i>Action Input:</i> 1884	has been hit by several major storms, with	746
697	<i>Final Answer:</i> 1884	heavy rainfall and floods causing significant	747
698		damage. The articles mention a national	748
		state of emergency being declared due to	749
		these storms, which aligns with the question.	750
699	• Correct Tool Usage Example of Search	Therefore, the answer to the question is	751
700	Engine	Greece. <i>Action:</i> Finish	752
701	Task Description: Choose the best option to	<i>Action Input:</i> Greece	753
702	answer the given question	<i>Final Answer:</i> Greece	754
703	· Question: Which country declared a		755
704	National State of Emergency this week for		

756	• Correct Tool Usage Example of Wikipedia Search		
757	Task Description:		
758	Answer the given question in a short sentence:		
759	Were Michael Winner and Guy Hamilton both	granddaughter of the Christian campaigner Sir	807
760	English film directors?	Culling Eardley, 3rd Baronet, and of the politi-	808
761	<i>Thought:</i> I need to find information about	cian Thomas Fremantle, 1st Baron Cottesloe.	809
762	Michael Winner and Guy Hamilton to deter-	His parents divorced in 1923, and Hamilton	810
763	mine if they were both English film directors.	attended school in England.[2]. [3] His first	811
764	<i>Action:</i> WikiSearch	exposure to the film industry came in 1938,	812
765	<i>Action Input:</i> "Michael Winner" <i>Observation:</i>	when he was a clapperboard boy at the Vic-	813
766	Robert Michael Winner (30 October 1935	torine Studios in Nice.	814
767	2013 21 January 2013) was a British film-	<i>Thought:</i> I have found information about both	815
768	maker, writer, and media personality. He is	Michael Winner and Guy Hamilton. Now I	816
769	known for directing numerous action, thriller,	need to determine if they were both English	817
770	and black comedy films in the 1960s, 1970s	film directors.	818
771	and 1980s, including several collaborations	<i>Action:</i> Finish	819
772	with actors Oliver Reed and Charles Bron-	<i>Action Input:</i> Yes, they were both English film	820
773	son.[1][2]. Winner's best-known works in-	directors.	821
774	clude Death Wish (1974) and its first two	<i>Final Answer:</i> Yes, they were both English	822
775	sequels, the World War II comedy Hannibal	film directors.	823
776	Brooks (1969), the hitman thriller The Me-		
777	chanic (1972), the supernatural horror film	• Correct Example of General Datasets	824
778	The Sentinel (1977), the neo-noir The Big	Task Description:	825
779	Sleep (1978), the satirical comedy Won Ton	Given a passage and a question, try your best	826
780	Ton, the Dog Who Saved Hollywood (1976),	to answer the question true or false according	827
781	and the Revisionist Westerns Lawman (1971)	to the passage.	828
782	and Chato's Land (1972).. Winner was known	· Passage: In statistics, ordinary least squares	829
783	as a media personality in the United King-	(OLS) is a type of linear least squares method	830
784	dom, appearing regularly on television talk	for estimating the unknown parameters in	831
785	programmes and publishing a restaurant re-	a linear regression model. OLS chooses	832
786	view column for The Sunday Times. He was	the parameters of a linear function of a set	833
787	also a founder of the Police Memorial Trust..	of explanatory variables by the principle	834
788	<i>Thought:</i> I found information about Michael	of least squares: minimizing the sum of	835
789	Winner, but I still need to find information	the squares of the differences between the	836
790	about Guy Hamilton.	observed dependent variable (values of the	837
791	<i>Action:</i> WikiSearch	variable being predicted) in the given dataset	838
792	<i>Action Input:</i> "Guy Hamilton"	and those predicted by the linear function.	839
793	<i>Observation:</i> Mervyn Ian Guy Hamilton,	· Question: are least squares the same as	840
794	DSC (16 September 1922 2013 20 April 2016)	linear regression? <i>Thought:</i> I can answer the	841
795	was an English film director. He directed 22	question without any tools. According to	842
796	films from the 1950s to the 1980s, including	the passage, least squares are not the same	843
797	four James Bond films.. Hamilton was born	as linear regression. Rather, least squares is	844
798	in Paris on 16 September 1922, son of Freder-	a method used within linear regression for	845
799	ick William Guy Hamilton (1895-1988), press	estimating the unknown parameters.	846
800	attach00e9 to the British embassy in Paris	<i>Action:</i> Finish	847
801	and Captain in the King's Own Royal Lan-	<i>Action Input:</i> false	848
802	caster Regiment, and Winifred Grace Culling	<i>Final Answer:</i> false	849
803	(1895-1970), daughter of William Archibald		850
804	Culling Fremantle, of the Church Missionary	C Examples of Failure Type	851
805	Society in India.[1] His mother was a great-	• Incorrect or Unnecessary Tool Invocation	852
806		Model: ChatGPT	853
		· Context:	854
		English breakfast is a very big meal—eggs,	855
		tomatoes, tea, coffee... For many people,	856

857	lunch is a quick meal. In cities, there are a lot	'StepByStepSolution'}}}	908
858	of sandwich bars. People can buy sandwiches	<i>Thought:</i>	909
859	there. Students can have a hot meal at school,	<i>Error Raise:</i> Thought string is null, model	910
860	but many just take a sandwich, a drink, and	refused to answer	911
861	some fruit from home. "Tea" means two	Action Null	912
862	things. It is a drink and a meal! Some people	· Context: I have 3 hundreds, 8 tens, and 3	913
863	have afternoon tea, with sandwiches, cakes,	ones.	914
864	and a cup of tea. They usually have dinner	· Question: What number am I?	915
865	quite early, between 6:00 and 8:00(.....), and	<i>Thought:</i> I should try to find a way to convert	916
866	often all the family eat together.	the numbers into a number that I can add.	917
867	People often get take-away meals—they buy	<i>Action:</i>	918
868	the food outside	<i>Error Raise:</i> Action string is null, model	919
869	· Question:	refused to answer	920
870	When they get a takeaway meal, they often	Action Input Null	921
871	eat it ()?	· Context:	922
872	· Options:	Three carpenters are installing some doors for	923
873	A. "at home"	a building.	924
874	B. "in the school"	Carpenter A installs a door every 40 min.	925
875	C. "outside"	Carpenter B installs a door every 1 h 20 min.	926
876	D. "in the bars"	Carpenter C installs a door every 1 h 40 min.	927
877	<i>Thought:</i> There is no answer in the passage.	If they start installing three doors at the same	928
878	<i>Action:</i> BingSearch	time at 9:15,	929
879	<i>Action Input:</i> Where do people often eat	· Question:	930
880	takeaway meals?	at what time will they install doors at the	931
881	<i>Error Raise:</i> It's general data, you can not	same again?	932
882	use the tools.	<i>Thought:</i> I can use the calculator to calculate	933
883		the time.	934
884		<i>Action:</i> Calculator	935
885	• Null Content	<i>Error Raise:</i> Action Input string is null,	936
886	Model: Llama2-13B-hf	model refused to answer	937
887	Thought Null		938
888	· Context:		
889	Olivia picked up sixteen pieces of paper from	• Recall correct tool not reasoning	939
890	the floor. If Edward picked up three pieces,	Model: Llama2-13B-hf	940
891	· Question: how many did they pick up in	· Context:	941
892	total?	Jerry was weighing boxes for moving. The	942
893	<i>Thought:</i> I should try to calculate the total	first was 2 pounds, the second was 11 pounds	943
894	number of pieces of paper picked up.	and the last was 5 pounds.	944
895	<i>Action:</i> Calculator	· Question:	945
896	<i>Action Input:</i> 16 + 3	What is the combined weight of all three	946
897	<i>Observation:</i> {'@title': 'Input', 'sub-	boxes?	947
898	pod': {'@title': '', 'img': {'@src': {given	<i>Thought:</i> I need to calculate the combined	948
899	link}, '@alt': '16 + 3', '@title': '16 +	weight of the boxes.	949
900	3', '@contenttype': 'image/gif'}, 'plain-	<i>Action:</i> Calculator	950
901	text': '16 + 3'}}, {'@title': 'Result',	<i>Action Input:</i> 2 + 11 + 5	951
902	'@primary': 'true', 'subpod': {'@title': '',	<i>Observation:</i> {'@title': 'Result', '@pri-	952
903	'img': {'@src': {given link}, '@alt': '19',	mary': 'true', 'subpod': {'@title': '', 'img':	953
904	'@title': '19', '@contenttype': 'image/gif'},	{'@src': {given link}, '@alt': '18', '@title':	954
905	'plaintext': '19'}, 'states': {'@count': '1',	'18', '@contenttype': 'image/gif'}	955
906	'state': {'@name': 'Step-by-step solution',	<i>Thought:</i> I need to calculate the combined	956
907	'@input': 'Result_Step-by-step solution',	weight of the boxes.	957
	'@stepbystep': 'true', '@buttonstyle':	<i>Action:</i> Calculator	958

959	<i>Action Input:</i> 2 + 11 + 5	glish and then use the translator to answer the	1009
960	<i>Observation:</i> { '@title': 'Result', '@primary': 'true', 'subpod': { '@title': '', 'img':	question.	1010
961	{ '@src': {given link}, '@alt': '18', '@title':	<i>Action:</i> Translate the question into English	1011
962	'18', '@contenttype': 'image/gif' }	<i>Action Input:</i>	1012
963	<i>Thought:</i> I need to calculate the combined	ماذا كان الاسم الأخير لأهرون؟	1013
964	weight of the boxes.		1014
965	<i>Action:</i> Calculator	<i>Observation:</i> Please choose a valid tool in	1015
966	<i>Action Input:</i> 2 + 11 + 5	Translator!	1016
967	<i>Observation:</i> { '@title': 'Result', '@primary': 'true', 'subpod': { '@title': '', 'img':	<i>Error Raise:</i> Continue to use the invalid	1017
968	{ '@src': {given link}, '@alt': '18', '@title':	action tool, the model refused to answer.	1018
969	'18', '@contenttype': 'image/gif' }		1019
970	<i>Thought:</i> I need to calculate the combined		
971	weight of the boxes.	• Similar invalid parameter recycle	1020
972	<i>Action:</i> Calculator	Model: Llama2-13B-Chat-hf	1021
973	<i>Action Input:</i> 2 + 11 + 5	· Context:	1022
974	<i>Error Raise:</i> Continue to try the same tool	Uncle Dave bought 143 ice cream sandwiches.	1023
975	and parameter, model refused to answer	If he wants to give them to his 11 hungry	1024
976		nieces,	1025
977		· Question:	1026
978		how many can each niece get?	1027
979	• Recall invalid tool after reminding	<i>Thought:</i> Hmm, I wonder if I can use the cal-	1028
980	Model: Zephyr-7B	culator to calculate the number of ice cream	1029
981	· English Context:	sandwiches each niece can get.	1030
982	"Das aschkenasische Rabbinat : Studien über	<i>Action:</i> Calculator	1031
983	Glaube und Schicksal" (Julius Carlebach)	<i>Action Input:</i> 143 ice cream sandwiches / 11	1032
984	/ Die Rolle des Rabbiners in Deutschland	nieces	1033
985	heute (Ahron Daum)	<i>Observation:</i> WolframAlpha API cannot	1034
986	· Arabic Question:	parse the input query.	1035
987		<i>Thought:</i> Oh no! Maybe I need to rephrase	1036
988	ماذا كان الاسم الأخير لأهرون؟	the question.	1037
989		<i>Action:</i> Calculator	1038
990	<i>Thought:</i> I can translate the question into En-	<i>Action Input:</i> 143 ice cream sandwiches / 11	1039
991	glish and then use the translator to answer the	nieces = how many ice cream sandwiches per	1040
992	question.	niece?	1041
993	<i>Action:</i> Translate the question into English.	<i>Observation:</i> WolframAlpha API cannot	1042
994	<i>Action Input:</i>	parse the input query.	1043
995	ماذا كان الاسم الأخير لأهرون؟	<i>Thought:</i> Hmm, maybe I need to convert the	1044
996		number of ice cream sandwiches to a more	1045
997	<i>Observation:</i> Please choose a valid tool in	manageable unit.	1046
998	Translator!	<i>Action:</i> Calculator	1047
999	<i>Thought:</i> I can translate the question into En-	<i>Action Input:</i> 143 ice cream sandwiches / 11	1048
1000	glish and then use the translator to answer the	nieces = how many ice cream sandwiches per	1049
1001	question.	niece (in dozens)?	1050
1002	<i>Action:</i> Translate the question into English	<i>Observation:</i> WolframAlpha API cannot	1051
1003	<i>Action Input:</i>	parse the input query.	1052
1004	ماذا كان الاسم الأخير لأهرون؟	<i>Error Raise:</i> Observation return same content,	1053
1005		model refused to answer.	1054
1006	<i>Observation:</i> Please choose a valid tool in	• Endless retry	1055
1007	Translator!	Model: Zephyr-7B	1056
1008	<i>Thought:</i> I can translate the question into En-	Question: The McCarran Airport Connector	1057
		serves what airport named after late U.S. Sen-	1058

ator Pat McCarran?
Thought: I need to find information about the McCarran Airport Connector and the airport it serves, which is named after a late U.S. Senator.

Action: WikiSearch

Action Input: McCarran Airport Connector
Observation: State Route 171. The Harry Reid Airport Connector (RAC) is a limited-access roadway system located in Paradise, an unincorporated town in the Las Vegas Valley, Clark County, Nevada, United States. Composed of State Route 171 (SR 171), the Airport Tunnel and arterial streets, the airport connector provides vehicular access to the passenger terminals at Harry Reid International Airport. Despite being completely owned by Clark County, the first 0.685 miles (1.102 km) of the Harry Reid Airport Connector is maintained by NDOT as unsigned SR 171, while the remaining section is maintained by Clark County.[1]. The Harry Reid Airport Connector was constructed and opened to traffic in 1994, in conjunction with the completion of I-215 between Interstate 15 and Warm Springs Road.

Thought: I have found information about the McCarran Airport Connector, but I need to find information about the airport it serves, which is named after a late U.S. Senator.

Action: WikiSearch

Action Input: Pat McCarran

Observation: Patrick Anthony McCarran (August 8, 1876 – September 28, 1954) was an American farmer, attorney, judge, and Democratic politician who represented Nevada in the United States Senate from 1933 until 1954. McCarran was born in Reno, Nevada, attended Nevada State University (now the University of Nevada, Reno), and was a farmer and rancher. In 1902, he won election to the Nevada Assembly but left office in 1905 after an unsuccessful campaign for the Nevada State Senate. He studied law privately and was admitted to the bar in 1905, then won election as Nye County District Attorney. He served a two-year term, after which he returned to Reno.

Error Raise: The text is too long

D Prompts

The prompts used in the evaluation are shown in this section.

System's Instruction

You are AutoGPT, you can use many tools (functions) to do the following task if you need.

First I will give you the task description, and your task start.

At each step, you need to give your thought to analyze the status now and what to do next, with a function call to actually excute your step.

After the call, you will get the call result, and you are now in a new state.

Then you will analyze your status now, then decide what to do next...

After many (Thought-call) pairs, you finally perform the task, then you can give your finial answer.

Remember:

1.the state change is irreversible, you can't go back to one of the former state, if you want to restart the task, follow this:

Action: Finish

Action Input: I give up and restart

2.All the thought is short, at most in 5 sentence.

Few-shot Prompt

Instruction:

Answer the following questions as best you can. Specifically, you have access to the following APIs if you need:

{tool_description}

Use the following format:

Task description: the task you must finish.

Thought: you should always reason about the current situation.

Action: the action to take, should be one of {tool_list} and Finish.

Action Input: required parameter to the action

Observation: execution of the action

... (this Thought/Action/Action Input/Observation can repeat N times)

If you believe that you can answer the task, please call:

Action: Finish

Action Input: your answer string.

Here are some examples:

{example}

Begin!

Task Description: {user_instruction}

Zero-shot Prompt

Instruction:

Answer the following questions as best you can. Specifically, you have access to the following APIs if you need:

{function_description}

Use the following format:

Task description: the task you must finish.

Thought: you should always reason about the current situation.

Action: the action to take, should be one of {func_list} and Finish.

Action Input: required parameter to the action

Observation: execution of the action

... (this Thought/Action/Action Input/Observation can repeat N times)

If you believe that you can answer the task, please call:

Action: Finish

Action Input: your answer string.

Begin!

Task Description: {user_instruction}

Tool-Usage Example

Task Description:

Given a context and related question, try your best to answer the question.

· Context:

There are 96 cupcakes for 8 children to share.

· Question:

How much will each person get if they share the cupcakes equally?

Thought: I should use calculator for Common-Division to finish this task.

Action: Calculator

Action Input: 96/8

Observation: 12

Thought: I think I get the final answer.

Action: Finish

Action Input: 12 (cupcakes)

Final Answer: 12 (cupcakes)

Tool Description

· **WikiSearch:** A tool to search entities on Wikipedia. You should only use exactly a simple entity name as input for the search. Some Tips to use the API better:

1. When the search API doesn't find the corresponding page, you should search for a related entity in the return list.

2. You can only search one entity name in each action, so, don't contact multiple entity names in one search input.

3. Don't over-rely on this simple tool, you may figure out the next action based on your own knowledge.

· **WikiLookUP:** A tool to view content on Wikipedia. The keyword input to this tool should be simple like one or two words. Some Tips to use the tool better:

1. The WikiLookUP tool can only be used after search API since it depends on the result page of the search.

2. Don't over-rely on this simple tool, you may figure out the next action based on your own knowledge.

· **WikiDisambiguation:** A tool to disambiguate entities on Wikipedia. You should only use exactly a simple entity name as input for disambiguate. Some Tips to use the tool better:

1. When WikiSearch API results in an entity page that is not related, you should use

WikiDisambiguation to disambiguate the searched entity to find other entities with the same name.

2. Don't over-rely on this simple tool, you may figure out the next action based on your own knowledge.

· **BingSearch:** Use BingSearch to get the top 3 search results after inputting the key to search. The input must be the exact keyword you need to search.

· **BingLoadPage:** Use BingLoadPage to load the detailed page of the search result. The input must be an integer of the page index.

· **Translator:** Translate a given Arabic text to English.

· **Calculator:** Use this tool to calculate an equation. Input must be a math problem or a calculation query.

COT Prompt

Answer the following questions as best you can.

Use the following format:

Task description: the task you must finish

Rationale: you should always reason about question according to given information

Answer: you should give the final answer according to Rationale

Here is an example:

{example}

Begin!

Task Description: {user_instruction}

General Example

Task Description:

Given a passage and a question, try your best to answer the question true or false according to the passage.

· Passage:

In trigonometry, the law of sines, sine law, sine formula, or sine rule is an equation relating the lengths of the sides of a triangle (any shape) to the sines of its angles. According to the law,

· Question:

Can law of sines be used on any triangle?

Thought: I can answer the question without any tools. The passage states that the law of sines in trigonometry is an equation that relates the lengths of the sides of a triangle to the sines of its angles. It specifies that this law can be applied to triangles of any shape. Therefore, the law of sines can be used on any triangle.

Action: Finish

Action Input: true

Final Answer: true

COT Example

Task Description:

Given a goal and two solution, try your best to select the most suitable solution to achieve the goal.

· Goal: how do you dispense of something?

· Solution0: put it in a safe.

· Solution1: throw it away.

Rationale: Solution2 is more suitable as it directly addresses the goal of disposing of something, which typically involves discarding or throwing it away.

Answer: Solution1

E Tool Impact on Model Performance across General Datasets

Test set	Tool	GPT3	ChatGPT	Model w/ Tool				Zephyr-7B	Chatglm3-6b
				Llama2-7B		Llama2-13b			
				Base	Chat	Base	Chat		
Boolq	Translator	68.00	80.00	64.00	45.00	62.00	77.00	52.00	31.00
		56.00	11.00	0.00	0.00	0.00	0.00	35.00	10.00
	Calculator	83.00	85.00	59.00	42.00	56.00	52.00	8.00	43.00
		47.00	5.00	0.00	0.00	0.00	0.00	33.00	7.00
	Search Engine	45.00	70.00	41.00	9.00	46.00	60.00	53.00	23.00
		31.00	8.00	0.00	0.00	0.00	1.00	35.00	8.00
	WikiPedia Search	38.00	70.00	56.00	34.00	52.00	56.00	77.00	14.00
		42.00	8.00	0.00	0.00	0.00	0.00	34.00	18.00
	All	58.50	76.25	55.00	32.50	54.00	61.25	47.50	27.75
		20.00	6.00	0.00	2.00	0.00	0.00	17.00	20.00
RACE	Translator	86.93	79.47	71.47	40.00	70.07	58.67	55.20	59.73
		52.00	27.00	5.00	0.00	0.00	0.00	29.00	22.00
	Calculator	86.93	76.26	69.60	67.20	40.26	35.20	40.80	74.13
		57.00	28.00	6.00	0.00	0.00	1.00	33.00	30.00
	Search Engine	78.13	77.60	41.60	40.26	78.40	62.13	74.13	62.93
		25.00	23.00	0.00	0.00	0.00	1.00	37.00	32.00
	WikiPedia Search	79.73	76.53	50.93	54.13	59.47	54.13	67.47	67.47
		30.00	33.00	0.00	0.00	0.00	0.00	51.00	21.00
	All	82.93	77.47	58.40	50.40	62.05	52.53	59.40	66.06
		6.00	30.00	14.00	0.00	0.00	0.00	17.00	22.00
PIQA	Translator	77.00	57.00	51.00	41.00	41.00	48.00	38.00	37.00
		63.00	40.00	5.00	0.00	0.00	0.00	23.00	1.00
	Calculator	68.00	61.00	14.00	47.00	42.00	32.00	3.0	21.00
		51.00	40.00	0.00	0.00	0.00	0.00	18.00	10.00
	Search Engine	25.00	55.00	0.00	19.00	25.00	44.00	3.00	18.00
		13.00	44.00	0.00	0.00	0.00	0.00	49.00	5.00
	WikiPedia Search	31.00	62.00	10.00	17.00	1.00	50.00	19.00	0.00
		26.00	40.00	0.00	0.00	0.00	0.00	47.00	6.00
	All	50.25	58.75	18.75	31.00	27.25	43.50	15.75	19.00
		3.00	39.00	14.00	0.00	0.00	0.00	4.00	3.00
RTE	Translator	75.00	51.00	51.00	44.00	34.00	36.00	15.00	46.00
		31.00	21.00	0.00	0.00	0.00	0.00	45.00	21.00
	Calculator	71.00	60.00	53.00	32.00	7.00	42.00	3.0	47.00
		36.00	29.00	2.00	0.00	0.00	0.00	41.00	26.00
	Search Engine	61.00	45.00	48.00	13.00	13.00	34.00	2.00	42.00
		10.00	22.00	0.00	1.00	0.00	0.00	28.00	37.00
	WikiPedia Search	58.00	44.00	30.00	4.00	0.00	26.00	42.00	50.00
		10.00	15.00	0.00	0.00	0.00	0.00	28.00	35.00
	All	66.25	50.00	45.50	23.25	13.50	34.50	15.50	46.25
		3.00	12.00	0.00	0.00	0.00	0.00	1.00	35.00
HellaSwag	Translator	62.00	55.00	19.00	19.00	6.00	29.00	28.00	18.00
		51.00	22.00	0.00	0.00	0.00	0.00	24.00	7.00
	Calculator	60.00	64.00	21.00	16.00	7.00	16.00	3.0	25.00
		52.00	22.00	1.00	0.00	0.00	3.00	25.00	7.00
	Search Engine	40.00	50.00	21.00	10.00	4.00	27.00	0.00	29.00
		59.00	23.00	1.00	0.00	0.00	0.00	27.00	7.00
	WikiPedia Search	41.00	31.00	19.00	18.00	0.00	21.00	31.00	21.00
		39.00	26.00	0.00	0.00	0.00	0.00	23.00	15.00
	All	50.75	50.00	20.00	15.75	4.25	23.25	15.50	23.25
		23.00	28.00	1.00	0.00	0.00	0.00	4.00	11.00

Table 9: Detailed results of all general datasets experiment, where *T003* means *Text-Davinci-003*, and " " indicates few-shot results, while cells without background color indicate zero-shot results.