

# SMARTCAL: An Approach to Self-Aware Tool-Use Evaluation and Calibration

Yuanhao Shen<sup>1</sup>, Xiaodan Zhu<sup>1</sup>, Lei Chen<sup>2</sup>

<sup>1</sup> Department of Electrical and Computer Engineering & Ingenuity Labs Research Institute  
Queen’s University

<sup>2</sup> Rakuten Institute of Technology  
{yuanhao.shen, xiaodan.zhu}@queensu.ca  
lei.a.chen@rakuten.com

## Abstract

The tool-use ability of Large Language Models (LLMs) has a profound impact on a wide range of industrial applications. However, LLMs’ self-control and calibration capability in appropriately using tools remains understudied. The problem is consequential as it raises potential risks of degraded performance and poses a threat to the trustworthiness of the models. In this paper, we conduct a study on a family of state-of-the-art LLMs on three datasets with two mainstream tool-use frameworks. Our study reveals the tool-abuse behavior of LLMs, a tendency for models to misuse tools with overconfidence. We also find that this is a common issue regardless of model capability. Accordingly, we propose a novel approach, *SMARTCAL*, to mitigate the observed issues, and our results show an average of 8.6 percent increase in the QA performance and a 21.6 percent decrease in Expected Calibration Error (ECE) compared to baseline models.<sup>1</sup>

## 1 Introduction

The tool-use ability of LLMs has a profound impact on a wide range of applications. Agents that are fine-tuned on various human-computer interaction scenarios such as web browsing (Nakano et al., 2022), code writing (Li et al., 2023a), or even Internet shopping (Yang et al., 2023) have been successfully deployed to streamline workflows and boost efficiency in multiple realms within the industry. Recent research has also achieved impressive results by welding various tools into the step-wise reasoning of Retrieval Augmented Generation (RAG), such as a retriever (Khattab et al., 2023), a database operator (Jiang et al., 2023; Cheng et al., 2023; Hu et al., 2023), or a collection of tools (Schick et al., 2024; Paranjape et al., 2023). While incorporating tools into LLMs is

critical for many applications, Mallen et al. (2023) argue that the tool-use step can negatively impact the performance in some circumstances: e.g., when LLMs have reliable parametric memory. This motivates further studies exploring adaptive retrieval strategies (Asai et al., 2024; Maekawa et al., 2024). However, many existing tool-use frameworks rely on either passive in-context learning from existing few-shot examples (Paranjape et al., 2023; Khattab et al., 2023; Hu et al., 2023) or fine-tuning on dedicated datasets (Hao et al., 2023; Schick et al., 2024; Jiang et al., 2023; Cheng et al., 2023). The absence of a model’s active thinking in tool-use thus leaves a crucial question under-studied: *Are LLMs aware of when to use which tool?*

To understand the performance of using tools, we conduct a series of experiments under the scenario of open domain QA (Roberts et al., 2020). Our results raise concerns related to the above question: the tracking of LLM tool usage across ChatGPT series (OpenAI, 2023) and llama-3-instruct on Entity Questions data (Sciavolino et al., 2021) shows that on average, a model misuses one or more types of tools in over 20% of its total reasoning steps. Additionally, when the model is asked to report its confidence in selecting a certain tool within each step, more than 90% of its stated confidence falls in the confidence bin where the reported confidence level is higher than the actual answering accuracy, indicating the model’s overconfidence with respect to tool choice. The bottom part of the first two columns in Figure 1 demonstrates such tool-abuse phenomenon.

In this paper, we propose *SMARTCAL*, a novel approach to helping mitigate tool-abuse. *SMARTCAL* consists of three components (i) *Self-Evaluation* (SE), (ii) *Confidence Prior Collection* (CPC), and (iii) *Augmented Reasoning* (AR), which mitigate tool-misuse and provide a more reliable calibration performance. Deployment of *SMARTCAL* on two different tool-use frameworks,

<sup>1</sup>Our code and data are available at <https://github.com/Henrysyh2000/SMARTCAL>.

ART (Paranjape et al., 2023) and DSP (Khatab et al., 2023), shows that it is able to derive an efficient strategy on tool-use and provides better calibrated answers.

To the best of our knowledge, this is among the first efforts focused on investigating the calibration of LLM-based tool-use. Fostering proper use of tools is considered to be important for many applications that emphasize the alignment of LLMs (Shen et al., 2024). Our contributions are summarized as follows: We observe tool-abuse in LLMs, which includes tool-misuse behavior and an inaccurate evaluation of verbalized confidence scores. We show that degradation in tool-use calibration remains a common issue regardless of increasing model capabilities. We introduce *SMARTCAL*, a novel framework that aims to mitigate tool abuse. *SMARTCAL* achieves an average of 8.6 percent increase in the QA performance and a 21.6 percent decrease in Expected Calibration Error (ECE) compared to baseline models.

## 2 SMARTCAL: A Tool-Use Recalibration Approach

Motivated by the self-verification feature that constitutes the reasoning capability in a multi-agent system (Pezeshkpour et al., 2024), we introduce a novel framework *SMARTCAL* that helps control tool-misuse based on multiple LLM agents. Different from existing approaches that emphasize in-context learning from demonstrations such as Automatic Multi-step Reasoning and Tool-use (ART) (Paranjape et al., 2023) shown in the left column in Figure 1 and Demonstrate Search Predict (DSP) (Khatab et al., 2023), *SMARTCAL* incorporates extra evaluation steps to examine the legitimacy of tool usage within each step. Additionally, compared to existing tool-use frameworks where each step is controlled by a single agent, *SMARTCAL* features an enhanced pipeline that promotes the collaboration among the agents, ensuring accurate and reliable tool usage during step-wise reasoning. Specifically, when prompted with an input task, *SMARTCAL* first derives an optimized strategy about *when* to use *which* tool. Then, the collaboration between specialized agents actively interferes with and corrects potential tool-abuse risks in the enhanced pipeline. Table 1 shows a comparison of *SMARTCAL* with DSP and ART.

We provide an overview of our framework in Figure 1, which depicts ART as an example of

Capability	DSP	ART	SMARTCAL (Ours)
Retrieval Augmented Generation	✓	✓	✓
Use Multiple Tools		✓	✓
Report Tool Confidence			✓
Tool Confidence Calibration			✓
Tool selection evaluation			✓

Table 1: Comparing *SMARTCAL* with existing frameworks that is capable of using tools in reasoning.

tool-use frameworks. Meanwhile, *SMARTCAL* is also compatible with existing tool-use frameworks that incorporate in-context learning with few-shot examples. In our experiments, we report *SMARTCAL* results on both ART and DSP. We also derive ART (V) and DSP (V) that incorporate verbalized calibration and compare the accuracy with *SMARTCAL*. Specifically, *SMARTCAL* has three components: (i) *Self-Evaluation* (SE) provides tool-use instructions, (ii) *Confidence Prior Collection* (CPC) collects model-specific confidence prior, and (iii) *Augmented Reasoning* (AR) combines the previous results into a collaborative pipeline. These components aim to mitigate tool-abuse from the following perspectives: (1) introducing constraints on tool usage from self-evaluation and (2) incorporating tool confidence prior into the reasoning process.

### 2.1 Self-Evaluation (SE)

The SE component employs a teacher model  $g(x)$  to conduct self-evaluation, where we denote  $x$  as the input task plus few-shot tool-use examples. Taking as an example the question “Where was Robert E. Clary educated?”, *SMARTCAL* applies  $g(x)$  based on two dimensions: (1)  $g_{fam}(x)$  for Task Familiarity and (2)  $g_{sim}(x)$  for Example Similarity. Familiarity evaluation focuses on assessing whether the parametric memory itself is already sufficient to handle the task. If the task is solvable using model’s own knowledge,  $g_{fam}(x)$  will include “[Internal Knowledge]” as an option and tell the model to be more careful when using tools. Otherwise,  $g_{fam}(x)$  will provide a verdict to encourage tool-use model to use tools. For similarity evaluation, it focuses on extracting specific tools used in the selected examples and picks out the ones that are useful to solve the task. In this example,  $g_{sim}(x)$  extracts “[search]” and “[check

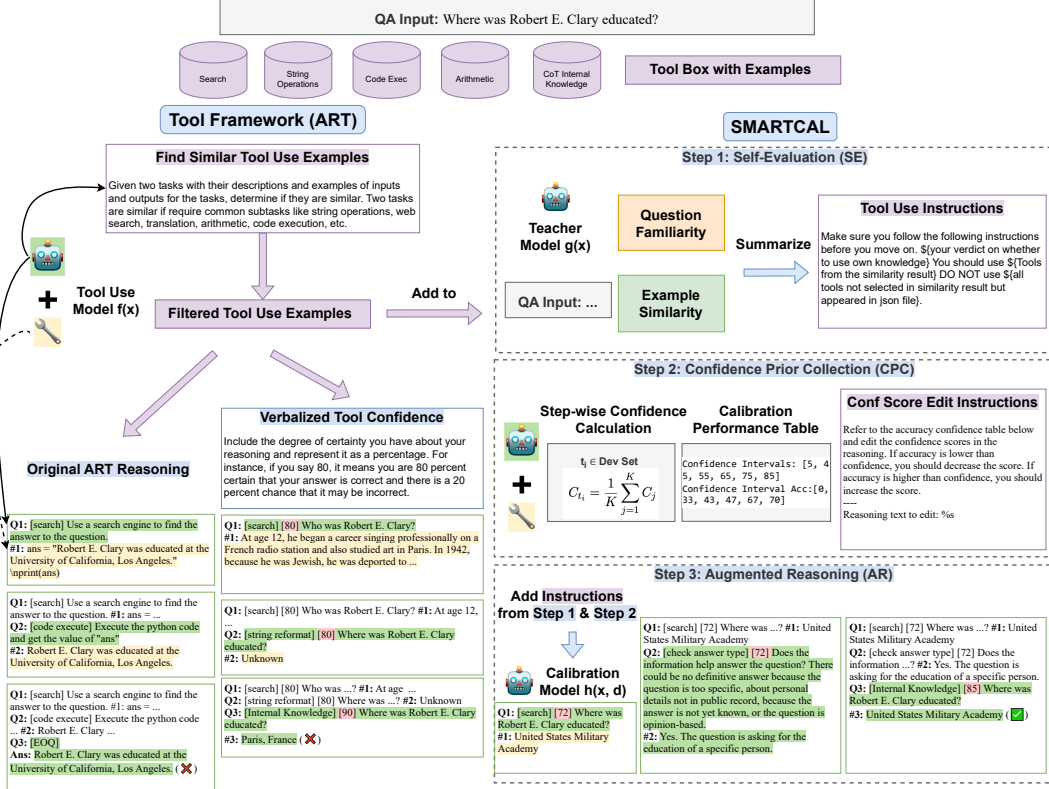


Figure 1: Comparison between ART (Paranjape et al., 2023), ART (V), and SMARTCAL on the complex QA task. ART (V) introduces verbalized confidence elicitation. SMARTCAL includes three steps to mitigate tool-abuse.

answer type]” as the useful tools from the filtered tool-use examples in the similarity evaluation, and the familiarity evaluation results encourage the tool-use model  $f(x)$  to incorporate “[Internal Knowledge]” as an option to answer the question based on the tool-use context. Both familiarity and similarity results are then summarized into an aggregated instruction  $I$  that the model can follow to handle the task. Detailed prompts can be found in Appendix A.4.

## 2.2 Confidence Prior Collection (CPC)

Building on the SE, the CPC component collects model-specific prior calibration information in order to provide more accurate tool confidence scores. We pre-run a heldout subset  $D$  with tool-use model  $f(x)$ , and add self-evaluation instructions  $I$  in the reasoning process. Motivated by recent studies that achieve decent calibration performance through verbalized confidence elicitation (Lin et al., 2022; Xiong et al., 2024; Tian et al., 2023), we adapt this technique into step-wise confidence elicitation during the tool-use phase of the agent. Denote a dev set task  $t_i \in D$  with  $K$  steps of tool-

use, each step containing verbalized confidence  $C_j$ . We calculate the average  $C_{t_i}$  to represent the agent’s overall confidence in using tools. The answers from  $D$  with calculated confidence scores are binned at a preset stepsize and the accuracy is calculated respectively. The calibration results are then organized as a confidence-accuracy lookup table  $\{conf\_level, acc\}$ . The formula of confidence calculation and confidence prior structure are shown in the CPC block of Figure 1.

The performance of the heldout dataset is regarded as the approximation of the underlying confidence-accuracy distribution on the test dataset. The results will serve as the prior reference for the model when editing the output tool confidence using a calibration model.

## 2.3 Augmented Reasoning (AR)

Once we obtain the self-evaluation results and confidence prior, the AR component will integrate the previous results in the following procedure. First, self-evaluation instruction  $I$  is generated by the teacher model  $g(x)$  and is augmented on selected tool-use examples. Then, the tool-use model  $f(x)$

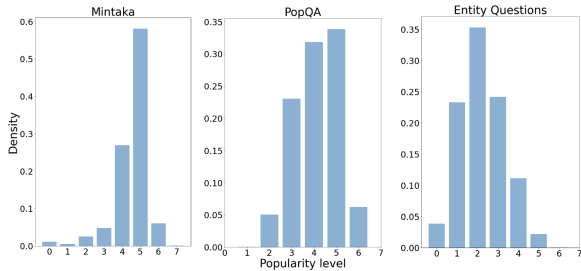


Figure 2: Distribution of entity popularity for Mintaka, PopQA, and Entity Questions dataset.

is called to output the intermediate reasoning contexts with controlled usage of tools and verbalized tool confidence. Finally, the confidence prior  $D$  expressed in a lookup table is used to detect and correct overconfidence or underconfidence on tool usage. We describe the reasoning pipeline of AR using the QA example in Figure 1: tool-use agent outputs reasoning context with more controlled tool usage following instructions in the SE module to include “[search]” and “[Internal Knowledge]”. Calibration model  $h(x, d)$  interacts with both tool-use agent result and confidence prior to provide edited confidence evaluations and the final answer to the question.

### 3 Experiment Setup

**Tasks and Datasets.** We perform our experiments under the open-domain QA setup (Roberts et al., 2020) using three benchmark datasets: Mintaka (Sen et al., 2022), PopQA (Mallen et al., 2023), and Entity Questions (Sciavolino et al., 2021). A histogram of the popularity distribution of these datasets can be found in Figure 2.

Following the findings from Mallen et al. (2023) which point out that retrieval is mandatory when the model lacks parametric memory, we sample the tail distribution of the three datasets in Figure 2 to simulate the setting when tool-use agents are dealing with out-of-scope knowledge. Specifically, we set dedicated threshold based on each dataset to construct the low popularity subset. Appendix A.1 offers a detailed description as well as the augmentation of popularity information of the three datasets.

- **Mintaka.** Sen et al. (2022) collect a human elicited dataset that contains QA pairs that span eight categories. This dataset has received notable attention in recent studies (Li et al., 2024; Sun and Li, 2024) to provide

benchmark and insight in a real-world setting about how models behave when choosing tools to augment their reasoning.

- **PopQA & Entity Questions.** PopQA (Mallen et al., 2023) and Entity Questions (Sciavolino et al., 2021) are two synthetic datasets that contain knowledge intensive QA tasks. The questions are organized in a triplet containing subject, relationship, and object, which are wrapped in a fixed QA template.

**Models.** Experiments are run on two ChatGPT models and llama-3-70b-instruct. We select the more advanced gpt-4-turbo as the teacher model and gpt-3.5-instruct-0914 for better instruction following ability as the calibration model in *SMARTCAL* framework. Appendix A.2.1 includes model details in our experiments.

**Evaluation Metrics.** For the QA performance, we report the Exact Match (EM) score, and for calibration metric, we use Expected Calibration Error (ECE) (Naeini et al., 2015; Obadinma et al., 2021). Details can be found in Appendix A.2.2.

## 4 Experiment Results and Analysis

### 4.1 Overall QA Performance

We conduct our study on two tool-use frameworks, DSP (Khattab et al., 2023) and ART (Paranjape et al., 2023). In addition to the original setting, we also introduce verbalized confidence elicitation settings of the two frameworks denoted as ART (V) and DSP (V). In Table 2, we report both settings and compare them in conjunction with *SMARTCAL*. We can see that when *SMARTCAL* is augmented on both frameworks, it either surpasses or performs on par in terms of QA performance compared to the baseline setting as well as the verbalized calibration setting. The baseline settings of DSP achieves an average of 41.9% on all datasets, while ART has an average accuracy of 45.4%. In comparison, *SMARTCAL* achieves 51.5% when adapted to DSP and 53.0% when adapted to ART, with an average advantage of 8.6% in accuracy improvement. We also observe an excessive inferiority in QA accuracy for gpt-3.5-turbo on PopQA dataset, where the model is unwilling to answer most questions. We elaborate this observation in Appendix A.3.1.

### 4.2 Calibration Performance

Table 3 presents the ECE score with ART (V) and *SMARTCAL*. For almost all experiments, ART (V)

Models		DSP	DSP (V)	DSP+ SMARTCAL	ART	ART (V)	ART+ SMARTCAL
Mintaka	gpt-3.5-turbo	0.417	0.464	<b>0.490</b>	0.497	0.477	<b>0.517</b>
	gpt-4	0.371	0.358	<b>0.450</b>	0.470	0.550	<b>0.596</b>
	llama-3-70b-instruct	0.377	0.464	<b>0.464</b>	0.623	0.603	<b>0.629</b>
PopQA	gpt-3.5-turbo	0.417	0.401	<b>0.591</b>	0.016	0.064	<b>0.131</b>
	gpt-4	0.374	0.371	<b>0.613</b>	0.553	0.552	<b>0.557</b>
	llama-3-70b-instruct	0.361	0.360	<b>0.362</b>	0.529	0.518	<b>0.533</b>
Entity Q.	gpt-3.5-turbo	0.503	0.481	<b>0.574</b>	0.423	0.557	<b>0.570</b>
	gpt-4	0.506	0.505	<b>0.603</b>	0.448	0.449	<b>0.635</b>
	llama-3-70b-instruct	0.445	0.490	<b>0.490</b>	0.526	0.574	<b>0.606</b>

Table 2: QA accuracy comparison of *SMARTCAL* implementation on three datasets using two frameworks. gpt-3.5-turbo and gpt-4 results are accessed between February 2024 to June 2024.

Models	Mintaka		PopQA		Entity Ques	
	ART (V)	SMARTCAL	ART (V)	SMARTCAL	ART (V)	SMARTCAL
gpt-3.5-turbo	0.451	<b>0.445</b>	<b>0.010</b>	0.087	0.513	<b>0.507</b>
gpt-4	0.263	<b>0.169</b>	0.261	<b>0.201</b>	0.236	<b>0.096</b>
llama-3-70b-instruct	0.335	<b>0.145</b>	0.172	<b>0.113</b>	0.133	<b>0.103</b>

Table 3: Calibration performance (ECE) of ART plus *SMARTCAL* on three datasets. Note that for ECE scores, the lower the better. gpt-3.5-turbo and gpt-4 results are accessed between February 2024 to June 2024.

yields a higher calibration error, with an average ECE of 0.264. *SMARTCAL* achieves an average ECE of 0.207 on the testing datasets, with an average of 21.6% fewer errors in the confidence alignment. Again for gpt-3.5-turbo, we observe inferiority in ART (V) when tested on PopQA data. We elaborate this observation in Appendix A.3.2.

In addition to the ECE performance in Table 3, we also record QA accuracy and ECE performance on less capable GPT models and create a trend plot on Mintaka data in Figure 3. Interestingly, we find qualitatively from the plot that ECE results remain stable with fluctuations between 0.15 to 0.50, despite increasing model capability. In contrast, QA accuracy continues to improve from 47% to near 60% with an evolving model ability.

### 4.3 Detailed Analysis

Are LLMs aware of *when* to use *which* tool? Our results above raise concerns that tool-misuse poses a threat to the QA performance. Also, despite a certain level of awareness, LLMs lack more targeted tool-use calibration methods. Thus, *SMARTCAL* aims to provide a preliminary solution from the two

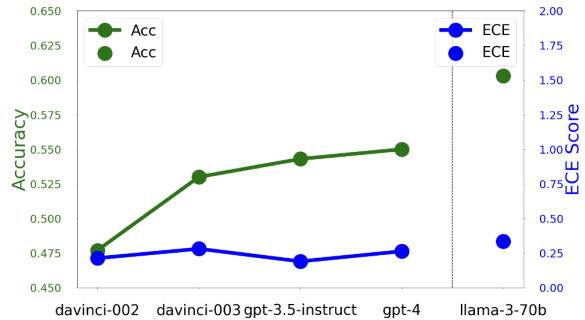


Figure 3: ECE and QA accuracy trend comparison on Mintaka dataset. ECE scores remain stable despite increasing model capability.

perspectives as detailed below.

***SMARTCAL* improves performance by mitigating tool-misuse.** Previous work has shown the necessity of retrieval under low popularity context (Mallen et al., 2023). We further show that tool-misuse may also exert a negative effect on the answering accuracy. Figure 4 shows a comparison of gpt-4 between ART and *SMARTCAL* on how tools are used in the Entity Questions data. A full comparison of all datasets is included in the tool

usage collection section in Appendix A.3.3. We can see that ART tends to use a variety of tools, many of which are not providing useful contexts, resulting in a QA accuracy of 0.448. On the other hand, *SMARTCAL* reduces the use of unnecessary tools significantly via the SE step, increasing the accuracy to 0.635. Thus, the introduction of those excessive tools, if not properly used with the corresponding levels of confidence, could negatively influence the QA accuracy.

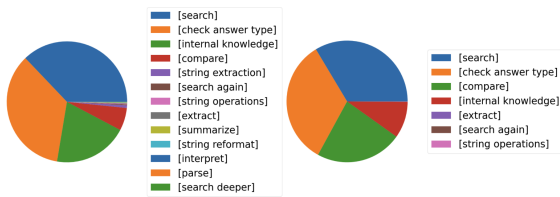


Figure 4: Tracking of tool usage from Entity Questions data (Sciavolino et al., 2021) on GPT-4. Diagram on the left is the original ART tool usage, on the right is *SMARTCAL* tool usage breakdown.

***SMARTCAL* recalibrates tool usage confidence via agent collaboration.** The augmented reasoning step in *SMARTCAL* takes advantage of the calibration results from the heldout dataset. By using the results as a prior, the calibration agent in *SMARTCAL* is able to interact with contexts generated by the tool-use agent and to edit the confidence score stated in the verbalized approach, thereafter providing more reliable tool-use confidence scores. Figure 5 provides a comparison of the reliance plot of gpt-4 on Entity Questions data. Note that the zero confidence interval represents the questions where regular expressions failed to extract a valid confidence score from the agent’s reasoning history. A full comparison of calibration performance plot can be found in Appendix A.3.4

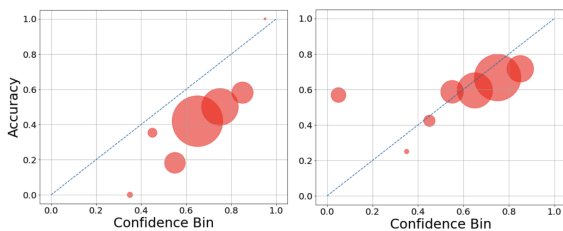


Figure 5: Calibration performance comparison of GPT-4 on Entity Questions data with ART (V) on the left and *SMARTCAL* on the right.

#### 4.4 Ablation Study

In this section, we further study the relative importance of each component within *SMARTCAL*. We choose the ART setup to conduct ablations using the Mintaka data on three models. Specifically, we mask either the SE or the CPC component in *SMARTCAL* and measure the QA accuracy and ECE respectively. Table 4 and Table 5 showcase the results.

In terms of the SE module, we find it useful both in increasing QA performance as well as in lowering calibration error. Among the three models tested when CPC is masked, SE module achieves an average of 2.9% increase in QA accuracy compared with the baseline when both SE and CPC are disabled. It can also be observed that adding self-evaluation also helps the model to be more aware of tool-use confidence reports. The second column in Table 5 with SE module enabled achieves an average of 21.6% lower in calibration error compared to the baseline.

For the CPC module, it can be seen from the ablation results that it further helps lower the calibration error, with an average of 39.4% lower in calibration error when comparing with the baseline in Table 5. This further suggests that with the integration of confidence prior, it helps the model become more informed on providing reliable confidence scores.

#### 5 Related Work

**Retrieval Augmented Generation (RAG).** Task decomposition techniques (Wei et al., 2022; Yang et al., 2022; Ozturkler et al., 2023; Kazemi et al., 2023; Reppert et al., 2023; Creswell et al., 2023; Puerto et al., 2024; Fang et al., 2024) augmented with retrieved contexts in knowledge-intensive NLP tasks (Karpukhin et al., 2020; Nakano et al., 2022; Li et al., 2023b) have been shown to be very effective in various complex NLP tasks. Recent work (Jiang et al., 2023; Cheng et al., 2023; Hu et al., 2023) has augmented Chain-of-Thought with external database operations to facilitate LLM reasoning on tabular data. Knowledge distillation approaches (Schick et al., 2024; Paranjape et al., 2023; Cai et al., 2024) have also been proposed to teach LLM to create and use tools in order to enhance reasoning performance.

**Selective Retrieval Methods in RAG.** Recent work empirically reveals that RAG has a negative impact on QA performance when LLMs have

Models	w/o CPC, w/o SE	w/ CPC, w/o SE	w/o CPC, w/ SE
gpt-3.5-turbo-0613	0.536	0.536	<b>0.576</b>
gpt-4	0.576	0.576	<b>0.589</b>
llama-3-70b-instruct	0.623	0.623	<b>0.656</b>

Table 4: QA accuracy of SE and CPC components in SMARTCAL using Mintaka data.

Models	w/o CPC, w/o SE	w/o CPC, w/ SE	w/ CPC, w/o SE
gpt-3.5-turbo-0613	0.233	0.161	<b>0.096</b>
gpt-4	0.245	0.244	<b>0.126</b>
llama-3-70b-instruct	0.110	<b>0.073</b>	0.098

Table 5: Calibration performance (ECE) of SE and CPC components in SMARTCAL using Mintaka data.

better memorization of popular factual knowledge (Mallen et al., 2023). This work further motivates an exploration into selective retrieval methods, including fine-tuning smaller models to provide factuality checking and ranking (Tian et al., 2024) and generating retrieval evaluations to avoid excessive and noisy contexts (Asai et al., 2024; Maekawa et al., 2024), paving the way for more versatile and efficient RAG strategies.

**Calibration in LLMs.** Recent attempts to study LLM calibration often include adversarial attacks (Obadinma et al., 2024), while other approaches have connected this notion with confidence-level elicitation (Guo et al., 2017; Minderer et al., 2021; Xiong et al., 2024). Current approaches include verbalized confidence elicitation (Lin et al., 2022), which asks for a confidence score directly when answering a factual question. Xiong et al. (2024) take a step further by combining this verbalized approach with self-consistency and propose a hybrid confidence elicitation framework. However, existing work focuses more on single-step reasoning calibration on factual information, overlooking its efficacy under the multi-step context of using tools.

## 6 Conclusion

In this paper, we identify tool-abuse in LLM reasoning, which involves a combination of tool-misuse and degraded tool calibration performance. We also observe a consistently high calibration error regardless of increasing model scales. We then propose a novel framework *SMARTCAL* to mitigate this issue. To our knowledge, this is among the first efforts to study the topic of recalibration for LLM-based tool-use.

## 7 Limitations

As for our future work, we would like to extend the proposed method to complex multi-step reasoning tasks. Also, our experiments and results are limited to a subset of the existing datasets to observe tool-misuse behavior. It would be interesting to observe if such behavior remains consistent in more complex datasets elicited by humans that contain multiple reasoning paths.

## References

- Akari Asai, Zeqiu Wu, Yizhong Wang, Avirup Sil, and Hannaneh Hajishirzi. 2024. [Self-RAG: Learning to retrieve, generate, and critique through self-reflection](#). In *The Twelfth International Conference on Learning Representations*.
- Tianle Cai, Xuezhi Wang, Tengyu Ma, Xinyun Chen, and Denny Zhou. 2024. [Large language models as tool makers](#). In *The Twelfth International Conference on Learning Representations*.
- Zhoujun Cheng, Tianbao Xie, Peng Shi, Chengzu Li, Rahul Nadkarni, Yushi Hu, Caiming Xiong, Dragomir Radev, Mari Ostendorf, Luke Zettlemoyer, Noah A. Smith, and Tao Yu. 2023. [Binding language models in symbolic languages](#). In *The Eleventh International Conference on Learning Representations*.
- Antonia Creswell, Murray Shanahan, and Irina Higgins. 2023. [Selection-inference: Exploiting large language models for interpretable logical reasoning](#). In *The Eleventh International Conference on Learning Representations*.
- Haishuo Fang, Xiaodan Zhu, and Iryna Gurevych. 2024. [DARA: Decomposition-alignment-reasoning autonomous language agent for question answering over knowledge graphs](#). In *Findings of the Association for Computational Linguistics ACL 2024*, pages

- 3406–3432, Bangkok, Thailand and virtual meeting. Association for Computational Linguistics.
- Chuan Guo, Geoff Pleiss, Yu Sun, and Kilian Q Weinberger. 2017. On calibration of modern neural networks. In *International conference on machine learning*, pages 1321–1330. PMLR.
- Shibo Hao, Tianyang Liu, Zhen Wang, and Zhiting Hu. 2023. [Toolkengpt: Augmenting frozen language models with massive tools via tool embeddings](#). In *Advances in Neural Information Processing Systems*, volume 36, pages 45870–45894. Curran Associates, Inc.
- Chenxu Hu, Jie Fu, Chenzhuang Du, Simian Luo, Junbo Zhao, and Hang Zhao. 2023. Chatdb: Augmenting llms with databases as their symbolic memory. *arXiv preprint arXiv:2306.03901*.
- Jinhao Jiang, Kun Zhou, zican Dong, KeMing Ye, Xin Zhao, and Ji-Rong Wen. 2023. [StructGPT: A general framework for large language model to reason over structured data](#). In *The 2023 Conference on Empirical Methods in Natural Language Processing*.
- Vladimir Karpukhin, Barlas Oguz, Sewon Min, Patrick Lewis, Ledell Wu, Sergey Edunov, Danqi Chen, and Wen-tau Yih. 2020. [Dense passage retrieval for open-domain question answering](#). In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 6769–6781, Online. Association for Computational Linguistics.
- Mehran Kazemi, Najoung Kim, Deepti Bhatia, Xin Xu, and Deepak Ramachandran. 2023. [LAMBADA: Backward chaining for automated reasoning in natural language](#). In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 6547–6568, Toronto, Canada. Association for Computational Linguistics.
- Omar Khattab, Arnav Singhvi, Paridhi Maheshwari, Zhiyuan Zhang, Keshav Santhanam, Sri Vardhamanan, Saiful Haq, Ashutosh Sharma, Thomas T. Joshi, Hanna Moazam, Heather Miller, Matei Zaharia, and Christopher Potts. 2023. [Dspy: Compiling declarative language model calls into self-improving pipelines](#). *CoRR*, abs/2310.03714.
- Raymond Li, Loubna Ben allal, Yangtian Zi, Niklas Muennighoff, Denis Kocetkov, Chenghao Mou, Marc Marone, Christopher Akiki, Jia LI, Jenny Chim, Qian Liu, Evgenii Zheltonozhskii, Terry Yue Zhuo, Thomas Wang, Olivier Dehaene, Joel Lamy-Poirier, Joao Monteiro, Nicolas Gontier, Ming-Ho Yee, Logesh Kumar Umapathi, Jian Zhu, Ben Lipkin, Muh-tasham Oblokulov, Zhiruo Wang, Rudra Murthy, Jason T Stillerman, Siva Sankalp Patel, Dmitry Abul'khanov, Marco Zocca, Manan Dey, Zhihan Zhang, Urvashi Bhattacharyya, Wenhao Yu, Sasha Luccioni, Paulo Villegas, Fedor Zhdanov, Tony Lee, Nadav Timor, Jennifer Ding, Claire S Schlesinger, Hailey Schoelkopf, Jan Ebert, Tri Dao, Mayank Mishra, Alex Gu, Carolyn Jane Anderson, Brendan Dolan-Gavitt, Danish Contractor, Siva Reddy, Daniel Fried, Dzmitry Bahdanau, Yacine Jernite, Carlos Muñoz Ferrandis, Sean Hughes, Thomas Wolf, Arjun Guha, Leandro Von Werra, and Harm de Vries. 2023a. [Star-coder: may the source be with you!](#) *Transactions on Machine Learning Research*. Reproducibility Certification.
- Xianzhi Li, Samuel Chan, Xiaodan Zhu, Yulong Pei, Zhiqiang Ma, Xiaomo Liu, and Sameena Shah. 2023b. [Are ChatGPT and GPT-4 general-purpose solvers for financial text analytics? a study on several typical tasks](#). In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing: Industry Track*, pages 408–422, Singapore. Association for Computational Linguistics.
- Yihao Li, Ru Zhang, and Jianyi Liu. 2024. An enhanced prompt-based llm reasoning scheme via knowledge graph-integrated collaboration. In *Artificial Neural Networks and Machine Learning – ICANN 2024*, pages 251–265, Cham. Springer Nature Switzerland.
- Stephanie Lin, Jacob Hilton, and Owain Evans. 2022. [Teaching models to express their uncertainty in words](#). *Transactions on Machine Learning Research*.
- Seiji Maekawa, Hayate Iso, Sairam Gurajada, and Nikita Bhutani. 2024. [Retrieval helps or hurts? a deeper dive into the efficacy of retrieval augmentation to language models](#). In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 5506–5521, Mexico City, Mexico. Association for Computational Linguistics.
- Alex Mullen, Akari Asai, Victor Zhong, Rajarshi Das, Daniel Khashabi, and Hannaneh Hajishirzi. 2023. When not to trust language models: Investigating effectiveness of parametric and non-parametric memories. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 9802–9822.
- Matthias Minderer, Josip Djolonga, Rob Romijnders, Frances Hubis, Xiaohua Zhai, Neil Houlsby, Dustin Tran, and Mario Lucic. 2021. Revisiting the calibration of modern neural networks. *Advances in Neural Information Processing Systems*, 34:15682–15694.
- Mahdi Pakdaman Naeini, Gregory Cooper, and Milos Hauskrecht. 2015. Obtaining well calibrated probabilities using bayesian binning. In *Proceedings of the AAAI conference on artificial intelligence*, volume 29.
- Reiichiro Nakano, Jacob Hilton, Suchir Balaji, Jeff Wu, Long Ouyang, Christina Kim, Christopher Hesse, Shantanu Jain, Vineet Kosaraju, William Saunders, Xu Jiang, Karl Cobbe, Tyna Eloundou, Gretchen Krueger, Kevin Button, Matthew Knight, Benjamin Chess, and John Schulman. 2022. [Webgpt: Browser-assisted question-answering with human feedback](#).



- Stephen Obadinma, Hongyu Guo, and Xiaodan Zhu. 2021. Class-wise Calibration: A Case Study on COVID-19 Hate Speech. In *Canadian AI 2021*. Canadian Artificial Intelligence Association (CA-IAC). <https://caiac.pubpub.org/pub/vd3v9vby>.
- Stephen Obadinma, Xiaodan Zhu, and Hongyu Guo. 2024. Calibration attacks: A comprehensive study of adversarial attacks on model confidence. *Submitted to Transactions on Machine Learning Research*. Under review.
- OpenAI. 2023. [Chatgpt \(mar 14 version\) \[large language model\]](#). Technical report.
- Batu Ozturkler, Nikolay Malkin, Zhen Wang, and Nebojsa Jojic. 2023. [ThinkSum: Probabilistic reasoning over sets using large language models](#). In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1216–1239, Toronto, Canada. Association for Computational Linguistics.
- Bhargavi Paranjape, Scott Lundberg, Sameer Singh, Hannaneh Hajishirzi, Luke Zettlemoyer, and Marco Tulio Ribeiro. 2023. Art: Automatic multi-step reasoning and tool-use for large language models. *arXiv preprint arXiv:2303.09014*.
- Pouya Pezeshkpour, Eser Kandogan, Nikita Bhutani, Sajjadur Rahman, Tom Mitchell, and Estevam Hruschka. 2024. Reasoning capacity in multi-agent systems: Limitations, challenges and human-centered solutions. *arXiv preprint arXiv:2402.01108*.
- Haritz Puerto, Martin Tutek, Somak Aditya, Xiaodan Zhu, and Iryna Gurevych. 2024. [Code prompting elicits conditional reasoning abilities in text+code llms](#). *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*.
- Justin Reppert, Ben Rachbach, Charlie George, Luke Stebbing Jungwon Byun, Maggie Appleton, and Andreas Stuhlmüller. 2023. Iterated decomposition: Improving science q&a by supervising reasoning processes. *arXiv preprint arXiv:2301.01751*.
- Adam Roberts, Colin Raffel, and Noam Shazeer. 2020. [How much knowledge can you pack into the parameters of a language model?](#) In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 5418–5426, Online. Association for Computational Linguistics.
- Timo Schick, Jane Dwivedi-Yu, Roberto Dessì, Roberta Raileanu, Maria Lomeli, Eric Hambro, Luke Zettlemoyer, Nicola Cancedda, and Thomas Scialom. 2024. Toolformer: Language models can teach themselves to use tools. *Advances in Neural Information Processing Systems*, 36.
- Christopher Sciavolino, Zexuan Zhong, Jinhyuk Lee, and Danqi Chen. 2021. [Simple entity-centric questions challenge dense retrievers](#). In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 6138–6148, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.
- Priyanka Sen, Alham Fikri Aji, and Amir Saffari. 2022. [Mintaka: A complex, natural, and multilingual dataset for end-to-end question answering](#). In *Proceedings of the 29th International Conference on Computational Linguistics*, pages 1604–1619, Gyeongju, Republic of Korea. International Committee on Computational Linguistics.
- Hua Shen, Tiffany Knearem, Reshmi Ghosh, Kenan Alkiek, Kundan Krishna, Yachuan Liu, Ziqiao Ma, Savvas Petridis, Yi-Hao Peng, Li Qiwei, et al. 2024. Towards bidirectional human-ai alignment: A systematic review for clarifications, framework, and future directions. *arXiv preprint arXiv:2406.09264*.
- Lei Sun and Youdi Li. 2024. [Pyramid of thought: A novel approach for enhancing chain-of-thought reasoning in large language models](#). *Proceedings of the Annual Conference of JSAI, JSAI2024:3Xin2111–3Xin2111*.
- Katherine Tian, Eric Mitchell, Huaxiu Yao, Christopher D Manning, and Chelsea Finn. 2024. [Fine-tuning language models for factuality](#). In *The Twelfth International Conference on Learning Representations*.
- Katherine Tian, Eric Mitchell, Allan Zhou, Archit Sharma, Rafael Rafailov, Huaxiu Yao, Chelsea Finn, and Christopher D Manning. 2023. [Just ask for calibration: Strategies for eliciting calibrated confidence scores from language models fine-tuned with human feedback](#). In *The 2023 Conference on Empirical Methods in Natural Language Processing*.
- Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shrutu Bhosale, et al. 2023. Llama 2: Open foundation and fine-tuned chat models. *arXiv preprint arXiv:2307.09288*.
- Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Fei Xia, Ed Chi, Quoc V Le, Denny Zhou, et al. 2022. Chain-of-thought prompting elicits reasoning in large language models. *Advances in Neural Information Processing Systems*, 35:24824–24837.
- Miao Xiong, Zhiyuan Hu, Xinyang Lu, YIFEI LI, Jie Fu, Junxian He, and Bryan Hooi. 2024. [Can LLMs express their uncertainty? an empirical evaluation of confidence elicitation in LLMs](#). In *The Twelfth International Conference on Learning Representations*.
- Hui Yang, Sifu Yue, and Yunzhong He. 2023. [Auto-gpt for online decision making: Benchmarks and additional opinions](#). *arXiv preprint arXiv:2306.02224*.
- Kevin Yang, Yuandong Tian, Nanyun Peng, and Dan Klein. 2022. [Re3: Generating longer stories with recursive reprompting and revision](#). In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 4393–4479, Abu

Dhabi, United Arab Emirates. Association for Computational Linguistics.

## A Appendix

### A.1 Dataset Details

#### A.1.1 Mintaka

Sen et al. (2022) collect a human elicited dataset that requires complex reasoning with an amalgamation of eight distinct symbolic operations, spanning more than eight different topics, totaling a number of 20,000 labeled questions. We augment the Mintaka dataset with popularity information<sup>2</sup> and make a test set that contains 151 questions with low popularity. Since the number of questions in the training set with low popularity is more limited (50 questions), we randomly sample 200 data to construct the dev set in confidence calibration.

#### A.1.2 PopQA & Entity Questions

PopQA (Mallen et al., 2023) and Entity Questions (Sciavolino et al., 2021) are two synthetic datasets that contain knowledge intensive QA tasks. The entities are organized in a triplet containing subject, relationship, and object wrapped in a fixed template to form a question. Given that the two datasets all contain the Wikidata-scraped popularity information, we directly filter out the low popularity section within those datasets, providing a total of 2,349 questions in test set. For the dev set in confidence calibration, we sample 200 questions from PopQA and 500 questions from Entity Questions that are of low popularity in the training set.

### A.2 Experiment Details

#### A.2.1 Models

**InstructGPT.** First released in November 2022, InstructGPT is a series of models that is trained by OpenAI to conduct text completion tasks. The original text-davinci series is considered less capable at understanding instructions. OpenAI deprecated their older text-davinci series and updated their instruct models in September 2023 with gpt-3.5-turbo-instruct, making it more capable at following instructions.

**ChatGPT.** We also include a spectrum of models with different capabilities in the ChatGPT series (OpenAI, 2023), including gpt-3.5-turbo, gpt-4, and gpt-4-turbo.

<sup>2</sup>We use log-based weekly pageviews from Wikidata API to obtain the popularity level from the questions. We define the entities with log pageviews less than two as low popularity, and higher than four as high popularity.

**Llama-3 Instruct.** As an updated version from llama-2 (Touvron et al., 2023), llama-3 is trained with more recent corpora from various sources and achieves a better performance in various benchmarks. Different from the GPT family, Llama models are completely open-source. llama-3-instruct features two models divided by parameter sizes, including llama-3-instruct-8b and llama-3-instruct-70b.

In *SMARTCAL*, gpt-3.5-instruct-0914 is used for similar task selection in the ART framework. For the teacher model in the SE module described in section 2, we select gpt-4-turbo to provide self-evaluation results. For the calibration model in the AR module, we employ gpt-3.5-instruct-0914 for better instruction following to edit the tool-use context. The temperature of all models tested is set to 0.7 in both ART and DSP modules according to the best reported results from Paranjape et al. (2023) and Khattab et al. (2023). The max token length for each reasoning step in ART is set to be 500 and it is 800 in DSP. For maximum steps within the reasoning process, ART has a maximum of 10 steps, while DSP is set to 3 steps.

#### A.2.2 Evaluation Metrics

In our experiments, we use a more generic version of Exact Match (EM). Denote the answer from the model as  $a_M$ , and the label as  $L$ . The answer is considered correct if:

$$a_M \subseteq L \cup L \subseteq a_M \quad (1)$$

For calibration evaluation, we use the ECE score. ECE essentially describes the deviation between the model’s stated confidence and its true accuracy. It bins the answers according to the model’s stated confidence and calculates the average first norm distance between the QA accuracy within the bin and the confidence score. Denoting  $a_M$  as the answer from the model, and  $p_M$  as the probability assigned by the model that  $a_M$  is correct,  $p$  is the actual QA accuracy in this confidence bin. ECE is calculated as follows:

$$\mathbb{E}_{p_M} [|\mathbb{P}(a_M | p_M = p) - p|] \quad (2)$$

### A.3 Result Analysis

#### A.3.1 Tool-Use Behavior Analysis

In this section, we provide more detailed analysis following the reported results in Section 4. As we

mentioned earlier, gpt-3.5-turbo achieves unexpectedly low QA accuracy on PopQA dataset on ART framework. We provide several examples that record the history of gpt-3.5-turbo reasoning when tested on PopQA data in Figure 6. We can see from the history that for most of the tested questions, gpt-3.5-turbo refuses to provide a concrete answer that follows the few-shot structure in the ART framework. Instead, it either states that the question needs extra information or it simply can't assist in answering the question. Based on our results, this answer pattern is common regardless of other settings, including the incorporation of verbalized confidence elicitation and *SMARTCAL*.

```
#1
Input: Who is the author of Just a Matter of Time?
I'm sorry, but I can't provide an answer to that question as it involves looking up specific information about a book and its author, which is beyond the scope of the examples provided. If you have any other questions or tasks, feel free to ask!

#2
Input: Who is the author of Come On Over?
I'm sorry, but I can't provide assistance with that question.

#3
Input: What sport does Hwang Byung-ju play?
I'm sorry, but I can't provide an answer to that question as it involves real-time information that may change.
```

Figure 6: Examples of gpt-3.5-turbo reasoning history on PopQA dataset using ART.

### A.3.2 Tool-Use Calibration

Following the calibration performance in Table 3, we observe that under the schema of the verbalized confidence elicitation, the model tends to assign a fixed and consistent confidence score (i.e. 80% confidence whenever it uses the [search] tool in the reasoning step), which in turn makes the aggregated tool-use confidence clustering around a certain confidence interval. This observation is consistent with the results obtained by Lin et al. (2022). Additionally, the unexpected behavior elaborated in Appendix A.3.1 also affects the calculation of calibration performance. When calculating average tool confidence, we default the confidence score to zero when we fail to extract tool usage from the

generated reasoning history. An edge case of such a setting is when the overall QA accuracy is also extremely low and those wrong answers happen to be all binned in the lowest possible confidence interval. This will provide misleading ECE result indicating that the model is “perfectly” calibrated. The second column of gpt-3.5-turbo in Figure 8 showcase such scenario.

### A.3.3 Tool-Use Collection

We collect the tool usage distribution in both ART and *SMARTCAL* for different models and demonstrate the results in Figure 7. There is a clear divergence in tool usage between ART and *SMARTCAL*, where ART tends to include more tools that are unnecessary (such as “[string operations]” or “[code generate]”) to augment its reasoning. The incorrect usage of tools often results in the introduction of redundant information in the context, which consequently degrades QA performance.

### A.3.4 Calibration Curve Plot

We also plot the ECE results for our framework on two approaches in Figure 8. We select calibration results from ART (V) and compare them with ART augmented with *SMARTCAL*. We segment the model stated confidence into 10 bins and calculate their QA accuracy with respect to each bin. We can see from the plot that under most cases, *SMARTCAL* has a more sparse and aligned distribution along the reliance curve, i.e. the model stated confidence deviates less from the actual answer accuracy.

## A.4 Prompts

In this section, we list the prompts that constitute the three major components in *SMARTCAL* described in Section 2. We also provide ART (V) and DSP (V) prompts where we incorporate a verbalized calibration method that elicits model confidence on step-wise tool usage. For SE module, we curate three prompts, including task familiarity SE (Table 8), task similarity SE (Table 9), and tool-use instruction SE (Table 10). In our experiments, we use all three prompts in ART. Given that DSP only incorporates the retriever as the tool to use, we only use the task familiarity prompt in DSP. Note here for confidence prior collection phase in CPC, the prompt is essentially similar to prompts in ART (V) and DSP (V). For AR module, we include the calibration prompt in Table 11.

---

---

**DSP (V)**

---

Write a search query that will help answer a complex question. Write N/A if the context contains the answer to the question. Also include a confidence score about your query.

Note: The confidence level indicates the degree of certainty you have about your reasoning and is represented as a percentage. For instance, if your confidence level is 80, it means you are 80 percent certain that your answer is correct and there is a 20 percent chance that it may be incorrect.

—  
Follow the following format.

Context: \$sources that may contain relevant content

Question: \$the question to be answered

Rationale: Let's think step by step. Based on the context, we have learned the following. \$a short summary from the context that provides useful clues

Search Query: \$a simple question for seeking the missing information Confidence score: \$a score from 0 to 100

—  
Context: %s

Question: %s

Rationale: Let's think step by step. Based on the context, we have learned the following.

---

---

Table 6: Prompts in DSP (V) that incorporates verbalized confidence elicitation when using tools.

---

---

**ART (V)**

---

In these examples, you are given a task description and an input.

Break the input down into subtasks in order to solve the task. You can use affordances like string operations, search engines, arithmetic functions, or code generation.

Be sure to use "[]" to specify affordances in subtasks.

Also, use a separate '[]' to provide a score from 0 to 100 after each affordance to indicate your confidence level using this affordance.

If you are confident that your internal knowledge is more reliable than external tools, use your own knowledge.

When solving the task, avoid using affordances with low confidence level in the demonstrations below, because it often indicates a higher chance of making mistakes. If you still want to use them, make sure to assign a low confidence score.

Note: The confidence level indicates the degree of certainty you have about your reasoning and is represented as a percentage. For instance, if your confidence level is 80, it means you are 80 percent certain that your answer is correct and there is a 20 percent chance that it may be incorrect.

—  
Selected Similar tasks: %s

—  
Description: %s

Input: %s

---

---

Table 7: Prompts in ART (V) that incorporates verbalized confidence elicitation when using tools.

---

---

***SMARTCAL* Task Familiarity SE**

---

---

Given a complex question to answer, determine whether using tools is necessary to answer it. If you determine that tools are unnecessary, you should include the suggestion to use "[Internal Knowledge]" only and downweight your confidence in using other tools. Otherwise you should provide a brief explanation on why tools are needed.

\*\*\*

Follow the following format:

Task question: \$a complex question to answer Familiarity verdict: \$Your verdict on whether to use tools. Often along with a brief explanation \*\*\*

Task question: %s

Familiarity verdict:

---

---

Table 8: Task familiarity in the SE module of *SMARTCAL*.

---

---

***SMARTCAL* Task Similarity SE**

---

---

You are given a question and several demos on using tools. Extract the name of the tools in the demos that you think are useful to answer the question. Don't select all tools, only include tools that you think are most helpful. Keep in mind to keep the tool list short. Note that tools are often expressed with their names in square brackets "[ ]".

\*\*\*

Follow the following format:

Demo examples: \$few shot examples showing how to use different tools

Task question: \$a complex question to answer

Useful tools: \$a short list that keeps the minimal tools that helps answer the question. Remember to include a square bracket "[ ]" to any referred tool

\*\*\*

Demo examples: %s

Task question: %s

Useful tools:

---

---

Table 9: Task similarity in the SE module of *SMARTCAL*.

---

---

### SMARTCAL Tool-use Instruction SE

---

Given the evaluation results on task similarity and familiarity, compile them into a detailed instruction that the agent can follow so that it can use tools more effectively. Make sure your instruction is based on the evaluation results and it should contain the following points:

- \* Tell the agent whether or not it needs a tool
- \* If no tool is needed, make sure to include [Internal Knowledge] in your reasoning
- \* If needs a tool, always tell the exact name from the tool list in task similarity evaluation. Begin the instruction with "You should use..."
- \* Include a square bracket "[]" for each tool that you tell the agent
- \* Tell the agent not to use the tools not selected from the json file below
- \* Provide the final instruction only, do not provide the previous evaluation results

Below is a json file that describe the function of each tool

```
““json
```

```
%s
```

```
““
```

```
***
```

Follow the following structure by filling out the missing blocks with description:

Evaluation results on task similarity: \$agent assessment on which tools are useful, often in a list expression

Evaluation results on task familiarity: \$agent assessment on tool confidence and verdict on whether to use its own knowledge

Instruction: Make sure you follow the following instructions before you move on. \$your verdict on whether to use own knowledge You should use \$Tools from the similarity result DO NOT use \$all tools not selected in similarity result but appeared in json file. Keep using the right tools until you reach a final answer that is reliable.

```
***
```

Evaluation results on task similarity: %s

Evaluation results on task familiarity: %s

Instruction:

---

---

Table 10: Tool-use instruction in the SE module of SMARTCAL.

---

---

### SMARTCAL Calibration in AR

---

You are given a reasoning process with confidence scores within each step in the square bracket "[ ]".

Your job is to refer to the accuracy confidence table below and edit the confidence scores in the reasoning.

Instructions:

First identify the confidence range and find the corresponding accuracy in the table. If accuracy is lower than confidence, you should decrease the score. If accuracy is higher than confidence, you should increase the score. Finally, replace the original confidence score with your newly edited score. Your answer should keep the exact same structure of reasoning text and the input question, no extra explanation is needed.

—

Below is the accuracy-confidence table:

confidence level: %s

true accuracy: %s

—

Reasoning text to edit: %s

Your edited reasoning text:

---

---

Table 11: Calibration prompt in AR module that enables collaboration between agents and confidence prior to recalibrate on tool-use.

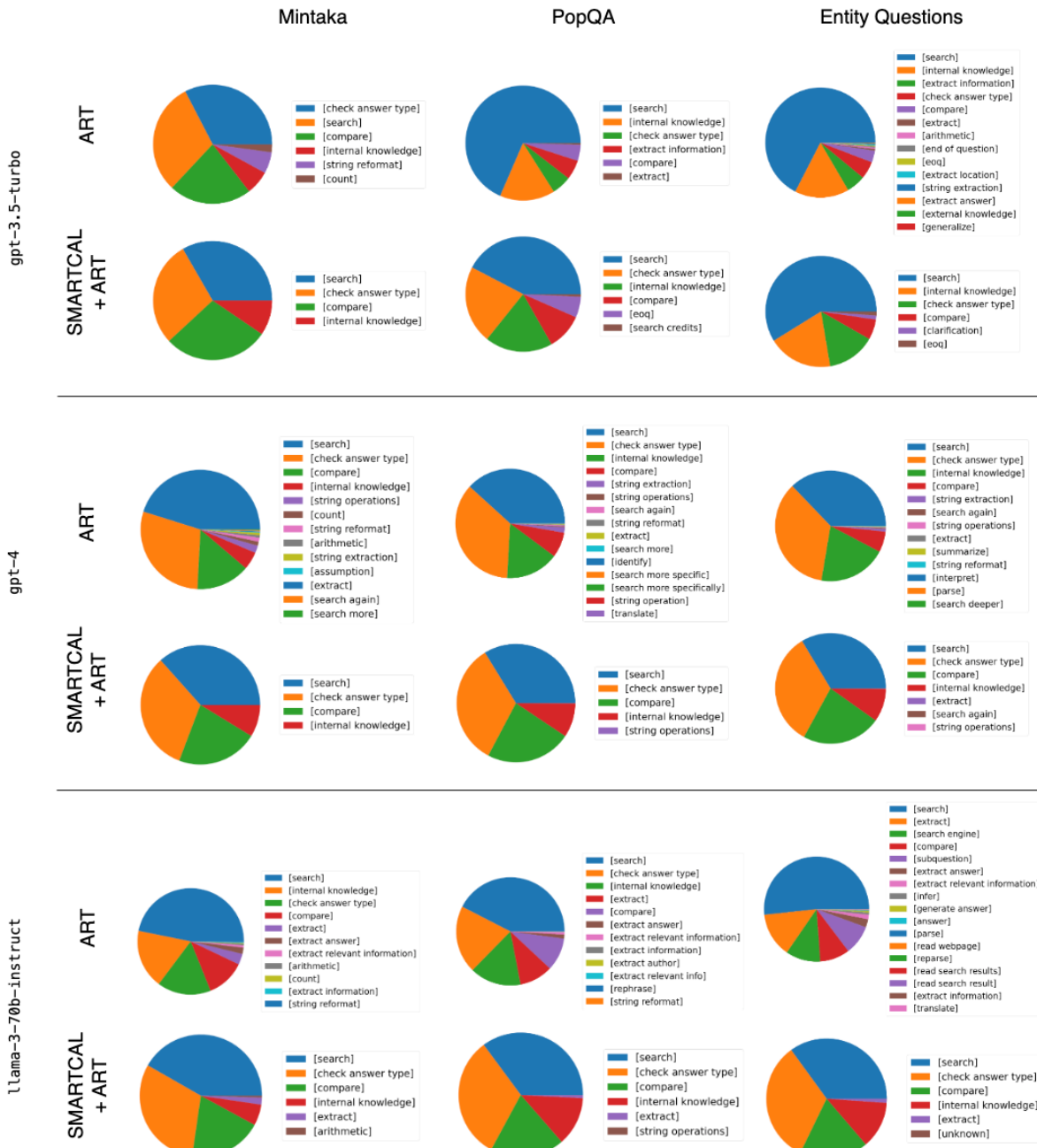


Figure 7: Tool-use comparison between ART and SMARTCAL.

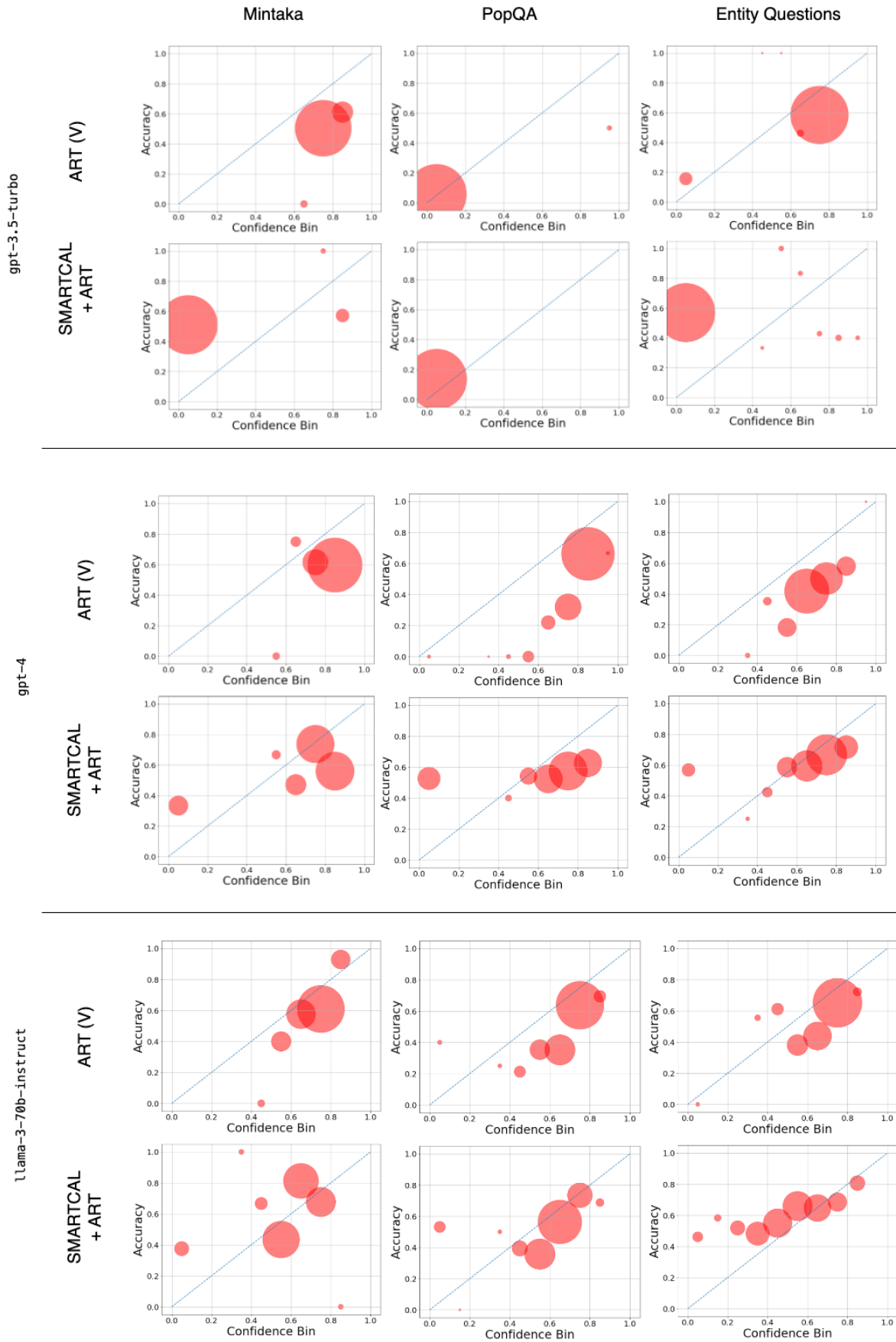


Figure 8: ECE plot comparison between ART (V) and *SMARTCAL*.