INVESTIGATING TOOL-MEMORY CONFLICTS IN LARGE LANGUAGE MODELS

Anonymous authors
Paper under double-blind review

ABSTRACT

Tool use is a defining capability of LLM-based agents, enabling them to act beyond their parametric knowledge and interact with external environments. But this capability also creates an arbitration problem: what should an LLM do when its parametric answer disagrees with a tool-derived answer? In this paper, we identify and systematically study this phenomenon, termed **Tool-Memory Conflict (TMC)**, a previously underexplored class of knowledge conflict for LLMs. We show that conflicts are common and highly model dependent, and that even state-of-the-art LLMs struggle under TMC, with notable performance degradation across diverse tasks. We further show that LLMs prioritize tool knowledge and parametric knowledge differently depending on factors such as task domain and conflict severity. We evaluate existing conflict-resolution techniques, including prompting-based and retrieval-augmented methods, and show that none reliably resolve tool-memory conflicts. Interestingly, an LLM with low conflict resolution may fall short in resolving conflicts. These results highlight TMC as an open challenge for building reliable LLM-based agents.

1 INTRODUCTION

Recent large language models (LLMs) can invoke external tools, such as search engines and APIs, to extend their capabilities beyond parametric knowledge (Schick et al., 2023; Patil et al., 2023; Qin et al., 2024a). This ability is central to LLM-based agents, enabling them to interact with the physical world, access up-to-date knowledge bases, and solve domain-specific problems that would otherwise exceed their capacity (Chen et al., 2023; Ma et al., 2024; Cheng et al., 2025; Lu et al., 2025).

However, the integration of external tools introduces a new source of conflict. Tool outputs may contradict the parametric knowledge encoded within the LLM (Xu et al., 2024b; Shi et al., 2024; Cheng & Amiri, 2025). For instance, an LLM may infer the behavior of a complex codebase, while calling a compiler and executing code produces verifiable output that may conflict with its internal knowledge. An external database query may return updated information that conflicts with the LLM's pre-training knowledge, which is frozen at a specific cutoff date. When such conflicts arise, LLMs face a fundamental dilemma, whether to trust their own internal memory (static parametric knowledge) or external tool (dynamic outputs), and lack reliable mechanisms to resolve it.

Existing work on knowledge conflict in LLMs has focused largely on RAGs, including context-memory conflict (Zhao et al., 2025; Wang et al., 2026) and inter-context conflict (Li et al., 2026), where additional documents are retrieved and added to the context of LLMs. However, the broader question is largely underexplored: *the conflict between parametric memory and external tool outputs*. Unlike RAG conflicts, where contradictory documents are present in the context window, the above conflict arises because the model must arbitrate between the tool result and its own no-tool answer. Understanding LLMs' behavior under conflicting scenarios is especially crucial when LLMs are given more agency to use external tools, such as improving reasoning ability, building safer agentic applications, and diagnosing knowledge deficiencies (Zhang et al., 2025; Betser et al., 2026; Feng et al., 2026).

In this paper, we study a broader knowledge conflict when an LLM is interacting with various external environments. We introduce **Tool-Memory Conflict (TMC)**, where an LLM's internal knowledge diverges from the output of external tools for the same query. TMC can lead to two opposite but equally dangerous failure modes: *stubbornness*, where the model ignores correct tool evidence in favor of

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.



Figure 1: Illustration of tool-memory Conflict (TMC). ✓ and ✗ mean the output is correct or wrong respectively compared to the ground truth answer.

outdated or erroneous memory, and *gullibility*, where the model accepts incorrect tool outputs despite having the correct knowledge internally. Studying TMC is therefore central to reliable agentic syg s.

The contributions of this paper are as follows:

- formalizing tool-memory conflict as a new source-arbitration problem in tool-augmented LLMs, and proposing an evaluation protocol.
- investigating how LLMs prioritize knowledge sources under tool-memory conflict, revealing substantial biases toward parametric memory.
- benchmarking multiple conflict-resolution strategies and showing that full tool traces help substantially but do not reliably address TMC across models and domains.

Our experiments show that tool-memory conflicts are prevalent across a variety of tasks and LLMs, including mathematical questions and factual question answering. Although LLMs are trained to leverage external tools, especially for professional or time-sensitive tasks, they can still prioritize internal memory even when tool outputs are correct. Furthermore, existing conflict-resolution techniques designed for inter-context conflicts, fail to reliably mitigate TMC. Specifically, tool use trace contributes more to resolving conflicts than simply adding more thinking tokens. Our work establishes TMC as an open challenge and motivates future work on source arbitration.

2 Related Work

Knowledge Conflicts in LLMs Existing works focus on three types of knowledge conflicts (Xu et al., 2024a), 1) context-memory conflict, where the output of internal parametric knowledge differs from the contextual knowledge (Longpre et al., 2021; Tan et al., 2024; Gekhman et al., 2023; Wang et al., 2024), 2) inter-context conflict, where conflicts happen between various pieces of contextual information (Zhang & Choi, 2023; Chen et al., 2022; Wan et al., 2024; Jin et al., 2024), and 3) inter-memory conflict, where conflicts happen internally in the parametric knowledge of LLMs (Lee et al., 2022b; Wang et al., 2023; Qi et al., 2023; Hase et al., 2023).

However, tool-memory conflict differs fundamentally from previously studied knowledge conflicts, *Our work is the first systematic study on tool-memory conflict.*

Tool Use in LLMs Tool-Augmented LLMs are LLMs that know how to call external tools to answer questions the their internal memory is insufficient at Schick et al. (2023). These LLMs are usually explicitly trained using tool using demonstrations that are synthesized with seed samples and LLMs (Patil et al., 2023; Li et al., 2023; Tang et al., 2023). modifying existing datasets (Basu et al., 2024), and dataset development with GPT-4 (Qin et al., 2024b). TAML (Parisi et al., 2022) used self-play to boost LLMs' performance on math and reasoning tasks.

3 A Framework to Study Tool-Memory Conflict

3.1 Problem Formulation: Tool-Memory Conflict

Let f_θ denote a tool-augmented LLM with parameters θ and access to an external tool set $\mathcal{T}$. For a query q with ground-truth answer a^*, we define two operational modes to elicit internal memory and

Table 1: Prevalence of Tool-Memory Conflict (TMC). M: memory-only. T: tool-only. 0: wrong. 1: correct. Overall accuracy (↑) is higher better, and conflict rate is lower better. Detailed results are shown in Table 6–7 in Appendix D.

LLM	Overall Acc (↑)			No TMC Rate			TMC Rate			
	M=1	T=1	Either=1	Total (↑)	Both=1	Both=0	Total (↓)	Both=0	M=1,T=0	M=0,T=1
Proprietary										
GPT-5.5	**91.6**	**92.6**	**94.2**	**95.2**	90.0	5.2	**4.8**	0.6	1.6	2.6
GPT-5.4	76.5	87.9	91.2	79.0	73.2	5.8	21.0	3.1	3.2	14.7
Claude Opus 4.8	83.9	86.7	91.0	86.4	79.6	6.8	13.6	2.2	4.3	7.1
Claude Opus 4.7	85.6	89.0	91.3	90.3	83.2	7.1	9.7	1.6	2.3	5.8
Claude Sonnet 4.6	86.2	83.1	89.9	87.2	79.4	7.8	12.8	2.3	6.8	3.7
Open-Source										
DeepSeek V3.2	72.0	84.5	88.8	74.8	67.7	7.1	25.2	4.1	4.3	16.8
Kimi K2.5	83.1	85.1	89.9	85.7	78.3	7.4	14.3	2.8	4.8	6.8
Llama 3.3 70B	65.5	76.4	85.5	65.7	56.4	9.3	34.3	5.2	9.1	20.0
Qwen3 32B	62.7	75.7	82.8	65.3	55.6	9.7	34.7	7.4	7.1	20.1
Llama 3.1 8B	54.8	52.9	73.7	44.3	34.0	10.3	55.7	15.9	20.8	18.9
Qwen3 8B	51.9	65.2	77.6	50.3	39.5	10.8	49.7	11.6	12.4	25.7
Average	74.0	79.9	86.9	74.9	67.0	7.9	25.1	5.2	7.0	12.9

external tool results

$$a_m = f_\theta(q \mid \theta, \mathcal{T} = \text{Disabled}) \qquad \text{(memory-only answer)} \tag{1}$$

$$a_t = f_\theta(q \mid \texttt{force tool}) \qquad \text{(tool-only answer)} \tag{2}$$

Memory-Only Mode The LLM is forced to answer q relying solely on its internal memory (parametric knowledge), with access to external tools disabled. This yields the memory-only answer a_m.

Tool-Only Mode The LLM is forced to generate tool calls to answer q. The tool calls are executed by users locally, yielding the tool-only answer a_t.

For a given query q, if the memory-only answer differs from the tool-only answer, we define this as **Tool-Memory Conflict (TMC)**

$$\mathbb{1}[a_m \neq a_t] \tag{3}$$

3.2 Conflict Resolution

When TMC occurs, we present both answers a_m and a_t with their sources jointly to the LLM and ask it to resolve the conflict, shown in Figure 1 (b). This yields the final joint answer

$$a_r = f_\theta(q \mid a_m, a_t). \tag{4}$$

We define a resolution as *appropriate* when the final answer a_r aligns with the correct source a^*

$$(a_r = a_t \wedge a^* = a_t) \text{ or } (a_r = a_m \wedge a^* = a_m) \tag{5}$$

Bias Probing To further reveal the bias and tendency of LLMs towards specific type of knowledge on different tasks. Inspired by Wu et al. (2024), we define

- Memory Bias: measures *memory stubbornness*, where the LLM prefers memory answer a_m despite tool answer a_t being correct, i.e. $Pr[a_r = a_m \mid a_m \neq a^*, a_t = a^*]$.
- Tool Bias: measures *tool gullibility*, where the LLM prefers tool answer a_t despite memory answer a_m being correct, i.e. $Pr[a_r = a_t \mid a_t \neq a^*, a_m = a^*]$.

4 EXPERIMENTAL SETUP

Datasets We evaluate TMC on the following widely-used benchmarks: PubMedQA (Jin et al., 2019), MedQA (Jin et al., 2021), MATH (Hendrycks et al., 2021), GSM8K-Platinum (Vendrow et al., 2025), SimpleQA (Wei et al., 2024), MMLU-Redux (Gema et al., 2025), These benchmarks cover a diverse range of tasks, including STEM, math, medical, humanities & social science, and factual knowledge.

LLMs We conduct the experiments across a wide range of LLMs which are capable of calling external tools, 1) GPT-5.5 (gpt, 2026), 2) GPT-5.4 (gpt, 2026), 3) Claude Opus 4.8 (cla, 2026b), 4) Claude Opus 4.7 (cla, 2026a), 5) Claude Sonnet 4.6 (cla, 2026c), 6) DeepSeek-V3.2 (Liu et al., 2024), 7) LLaMA-3.3-70B (Grattafiori et al., 2024), 8) LLaMA-3.1-8B (Grattafiori et al., 2024), 9) QWen-3-32B (Yang et al., 2025), 10) QWen-3-8B (Yang et al., 2025),

Tools For math tasks, we provide access to Wolfram Alpha, SymPy, and a Python interpreter. For factual question answering, we use the Wikipedia API and Brave web search. For medical question answering, we additionally provide access to the PubMed APIs, together with the Wikipedia API and Brave web search. To reduce failures caused by invalid tool calls or transient execution errors, we allow the model up to 5 retry attempts to revise and re-execute unsuccessful tool calls.

Conflict Resolution Methods We also evaluate whether existing conflict resolution methods can resolve tool-memory conflict. We evaluate: 1) asking LLM itself to resolve given a_m and a_t (Resolve), 2) asking LLM to resolve with full tool use trace (Full Trace), 3) Vigilant prompting (Pan et al., 2023), 4) Debate-of-Thoughts (DoT) (Li et al., 2026), 5) Explanation-based Thinking (XoT) (Zhao et al., 2026), and 6) SD-CDR (Zhou et al., 2025)

5 RESULTS

We aim to answer the following research questions: 1) Under what conditions (task type, model) do tool-memory conflicts arise? 2) Can existing methods designed for document-based conflicts resolve tool-memory conflicts? 3) When confronted with tool-memory conflict, do LLMs prioritize parametric knowledge or tool-generated outputs?

5.1 FREQUENCY OF TOOL-MEMORY CONFLICTS

Table 1 characterizes how frequently a model's memory-only and tool-only predictions disagree and how these disagreements relate to their underlying correctness. We observe three main patterns.

TMC Frequency is correlated with the quality of underlying predictions. disagreement between internal model knowledge and tool-derived answers is not an isolated failure case. TMC occurs across all evaluated models, with rates ranging from 4.8 to 55.7. The variation is substantial even among strong models. GPT-5.5 exhibits the lowest conflict rate at 4.8, followed by Claude Opus 4.7 at 12.0. In contrast, conflicts occur on more than 40% of examples on LLaMA and QWen series. Models with stronger memory-only and tool-only performance generally exhibit fewer conflicts. This is expected since higher accuracy means less chance of conflicting answers.

TMC direction is highly model-dependent. The composition of TMC reveals distinct model behaviors rather than a universal preference for either memory or tool. For GPT-5.4, DeepSeek V3.2, Llama 3.3 70B, Qwen3 32B, and Qwen3 8B, conflicts are dominated by memory wrong and tool correct (M=0,T=1) cases, indicating that the tool contains the correct signal but the final interaction nevertheless fails to exploit it reliably. In contrast, for Claude Sonnet 4.6, memory correct and tool wrong (M=1,T=0) is almost twice as common as memory wrong and tool correct (M=0,T=1) (6.8% vs. 3.7%). These asymmetric failure modes suggest that TMC is not explained by a single notion such as over-reliance on memory" or over-reliance on tools".

TMC is strongly domain-dependent. The per task results in Figure 2 show large shifts in TMC prevalence within the same LLM, particularly for weaker ones, indicating that robustness to con-

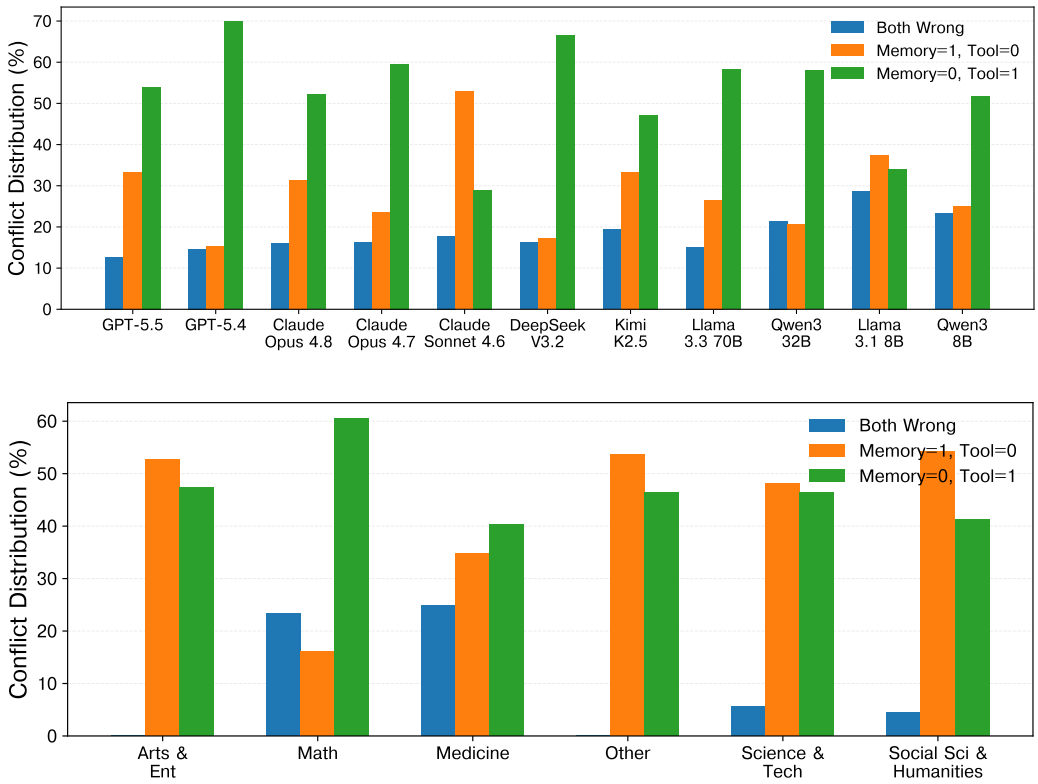


Figure 2: Tool-Memory Conflict distributions across LLMs and tasks.

flict does not transfer uniformly across knowledge domains. TMC for Math tasks are dominated by memory wrong and tool correct cases (M=0,T=1) (around 60% of conflicts), whereas Arts & Entertainment, and Social Science & Humanities tasks contain substantially more memory correct and tool wrong (M=1,T=0) cases.

Most TMC arises when one source is wrong. Across most models, the "Both Wrong" category accounts for only a minority of conflicts. This distinction is especially pronounced for stronger models: for GPT-5.5, only 0.6 percentage points of its 4.8% TMC rate correspond to both memory and tool being wrong, while the remaining conflicts arise from disagreement between a correct and an incorrect source. Even for models with substantially higher TMC, such as DeepSeek V3.2 and Qwen3 32B, most conflicts retain a correct signal in either memory or the tool. TMC therefore primarily tests whether an agent can select and integrate conflicting evidence, rather than simply whether it possesses the required knowledge.

5.2 Resolving Tool-Memory Conflicts

We next evaluate whether LLMs can correctly resolve a tool-memory conflict once a disagreement has occurred. Table **??** reports resolution accuracy across six resolution strategies, while Fig. **??** relates each model's conflict frequency to its average resolution performance. We observe the following main findings.

Tool-memory conflict remains difficult to resolve. Resolution performance varies substantially across models and methods. Averaged across the evaluated resolution strategies, model-level accuracy ranges from 34.6% for Llama-3.1 8B to 53.6% for Claude Opus 4.7. Most models achieve average accuracy below 52%, indicating that identifying the correct response to a disagreement remains challenging even after the conflict itself has been explicitly exposed. Thus, conflict detection and

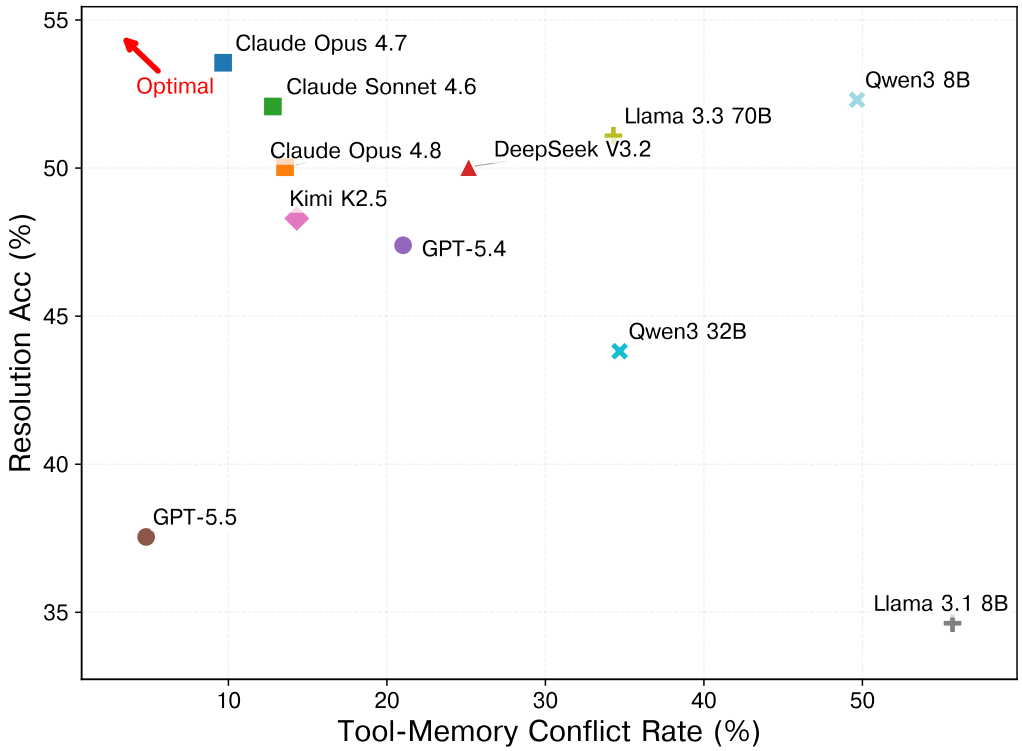


Figure 3: Tool-memory conflict rate vs. conflict resolution accuracy. Low TMC rate does not necessarily mean high resolution ability.

Table 2: Tool-memory conflict rate vs. conflict resolution performance. TMC remains challenging for existing resolution methods, designed for document-related conflicts. Detailed results are shown in Table 8–9 in Appendix D.

LLM	TMC	Resolution Accuracy (↑)						
	Rate (↓)	Avg.	Resolve	Full Trace	DoT	SR-DCR	Vigilant	XoT
Proprietary								
GPT-5.5	**4.8**	37.5	39.5	40.9	6.2	61.8	39.8	37.1
GPT-5.4	21.0	47.4	41.9	74.1	16.6	**65.1**	41.2	45.5
Claude Opus 4.8	13.6	50.0	48.9	75.4	2.9	55.8	49.8	67.4
Claude Opus 4.7	9.7	**53.6**	49.9	74.2	11.0	64.6	54.5	67.1
Claude Sonnet 4.6	12.8	52.1	**63.5**	**76.2**	2.6	35.9	**63.2**	**71.1**
Open-Source								
DeepSeek V3.2	25.2	50.0	49.8	71.0	19.2	54.8	52.8	52.6
Kimi K2.5	14.3	48.3	57.7	69.4	2.9	46.1	51.9	61.8
Llama 3.3 70B	34.3	51.1	46.2	73.4	18.2	62.6	43.4	62.8
Qwen3 32B	34.7	43.8	43.5	65.4	7.5	52.8	44.6	49.2
Llama 3.1 8B	55.7	34.6	33.1	51.6	15.7	32.7	28.3	46.2
Qwen3 8B	49.7	52.3	41.7	58.8	**49.3**	55.6	42.8	65.7
Average	25.1	47.3	46.9	66.4	13.8	53.4	46.6	57.0

conflict resolution constitute distinct problems: observing that memory and tool outputs disagree does not imply that the model can reliably determine how the disagreement should be resolved.

Full interaction traces substantially improve conflict resolution. Full Trace is the strongest resolution strategy for every model for which it is available. Its resolution accuracy ranges from

Table 3: Token cost of conflicting resolution.

	Avg.	Resolve	Full Trace	DoT	SR-DCR	Vigilant	XoT
Resolution Accuracy (↑)	47.3	46.9	66.4	13.8	53.4	46.6	57.0
Num Tokens (↓)	474.3	7.8	7.7	2283.5	451.5	8.0	87.2

40.9% for GPT-5.5 to 76.2% for Claude Sonnet 4.6, with an average of 66.4% across models. This substantially exceeds the average performance of XoT (57.0%), Vigilant (46.6%), Resolve (46.9%), and SR-DCR (53.4%). The consistent advantage of Full Trace indicates that retaining the complete interaction context provides important information for adjudicating disagreements between internal memory and tool outputs.

Resolution methods do not generalize across LLMs. Apart from Full Trace, no resolution strategy consistently dominates across models. For example, SR-DCR performs strongly on GPT-5.4 (65.1%) but reaches only 35.9% on Claude Sonnet 4.6. Conversely, Resolve achieves 57.7% on Kimi K2.5 but only 33.1% on Llama-3.1 8B. XoT performs well for several models, including Claude Opus 4.8 (67.4%), Kimi K2.5 (61.8%), yet drops to 45.5% on GPT-5.4. DoT exhibits an even sharper model dependence: its accuracy is close to zero for most models, but reaches 49.3% on Qwen3-8B. These results suggest that conflict-resolution behaviors are not transferable uniformly across model families, and that a strategy effective for one model can fail substantially on another.

Low Conflict $\neq$ Strong Conflict Resolution Ability As shown in Fig. **??**, models that encounter fewer tool–memory conflicts are not necessarily better at resolving the conflicts that do occur. Across models, TMC rate and average resolution accuracy are essentially uncorrelated ($r = -0.25$). GPT-5.5, for example, exhibits the lowest conflict rate at only 4.8%, yet achieves an average resolution accuracy of 37.5%. In contrast, Qwen3-8B conflicts on 49.7% of instances but achieves 52.3% average resolution accuracy. Llama-3.1-8B has a similarly high conflict rate of 55.7%, but resolves only 34.6% on average. These contrasting behaviors indicate that avoiding disagreement and reasoning correctly under disagreement are separate capabilities that should be evaluated independently.

More token $\neq$ Better resolution Resolving TMC requires LLMs to use tokens strategically rather than simply generating longer reasoning traces. As shown in Table 3, token usage is poorly aligned with resolution accuracy: DoT consumes by far the most tokens (2,283.5) yet achieves the lowest accuracy (13.8%), while Full Trace attains the highest accuracy (66.4%) with only 7.7 tokens on average. XoT similarly reaches 57.0% accuracy with just 87.2 tokens, substantially outperforming much more verbose methods. These results suggest that effective TMC resolution depends more on the quality and structure of reasoning than on reasoning length.

5.3 Probing the Systematic Tool / Memory Bias of LLMs

We next examine whether LLMs systematically favor one information source when resolving tool-memory conflicts. Table 4 reports the probability of *tool bias*, where the final response favors the tool over memory, and *memory bias*, where it favors memory over the tool, averaged across resolution methods. We define the bias gap as tool bias minus memory bias, such that a positive value indicates tool bias and a negative value indicates memory bias. We observe the following patterns.

Memory bias is more prevalent than tool bias, although it is modest for most models. Nine of the thirteen evaluated models have a negative bias gap, indicating that they override the tool in favor of memory more frequently than they override memory in favor of the tool. Across all models, the median bias gap is -4.1 percentage points. For example, Claude Opus 4.7 exhibits a moderate memory bias, with a tool bias rate of 18.6% and a memory bias rate of 24.9% (-6.3 points). Similar patterns appear for Llama 3.3 70B (-5.4), Qwen3 32B (-5.3), and Kimi K2.5 (-4.1). In contrast, only four models exhibit a positive bias gap, including DeepSeek V3.2 (+3.4), Claude Sonnet 4.6 (+1.6), Qwen3-Coder 30B-A3B (+8.8), and Qwen3-8B (+2.8). Thus, the dominant tendency across models is to favor internal memory, although this tendency is relatively small for most models.

Table 4: Tool Bias and Memory Bias of LLMs. Tool Bias (Sect. 3.2) denotes LLM mistakenly picks Tool as correct, given Memory is correct. A positive / negative bias gap means the LLM is biased towards tool / memory output.

LLM	Tool Bias (↓)	Memory Bias (↓)	Bias Gap
Proprietary			
GPT-5.5	**7.3**	57.7	-50.5
GPT-5.4	26.2	30.7	-4.5
Claude Opus 4.8	23.7	32.0	-8.3
Claude Opus 4.7	22.7	28.5	-5.8
Claude Sonnet 4.6	30.7	29.2	1.6
Open-Source			
DeepSeek V3.2	26.8	**22.3**	4.6
Kimi K2.5	26.8	30.1	-3.4
Llama 3.3 70B	25.1	29.8	-4.7
Qwen3 32B	24.3	29.3	-5.0
Llama 3.1 8B	<u>7.8</u>	45.5	-37.7
Qwen3 8B	25.5	<u>22.9</u>	2.6
Average	22.4	32.5	-10.1

Extreme source bias occurs almost exclusively toward memory. While most models have relatively balanced override behavior, two models exhibit pronounced memory bias. GPT-5.5 has a tool bias rate of only 1.1% but a memory bias rate of 38.3%, producing a bias gap of -37.2 points. Llama3.1-8B shows a similarly strong asymmetry, with 12.9% tool bias versus 39.0% memory bias (-26.1 points). By comparison, the strongest observed tool bias is substantially smaller: Qwen3-Coder 30B-A3B has a bias gap of only +8.8 points. Indeed, eleven of the thirteen models have absolute bias gaps of at most 10.3 points, with GPT-5.5 and Llama3.1-8B standing out as clear outliers. This asymmetry suggests that severe source preference, when it emerges, is considerably more likely to manifest as reliance on internal memory than as excessive deference to the tool.

Bias is highly model-specific and is not consistent even within related model families. Closely related models can exhibit substantially different source preferences. GPT-5.4 is nearly balanced, with a bias gap of -1.1 points, whereas GPT-5.5 exhibits the strongest memory bias in our evaluation at -37.2 points. A similar contrast appears within the Llama family: Llama 3.3 70B shows a moderate memory bias (-5.4), while Llama3.1 8B reaches -26.1. The Qwen family spans both directions: Qwen3 32B favors memory (-5.3), whereas Qwen3 8B (+2.8) favor the tool. These differences indicate that source preference cannot be characterized by a universal tool- or memory-favoring behavior shared across LLMs.

A small bias gap does not imply weak source intervention. The bias gap measures the *directional asymmetry* between tool and memory bias rather than the overall probability with which the model overrides either source. Consequently, models can exhibit substantial override behavior while remaining nearly unbiased in aggregate. For example, Claude Sonnet 4.6 overrides memory in favor of the tool on 25.8% of cases and overrides the tool in favor of memory on 24.2%, yielding a bias gap of only +1.6 points. GPT-5.4 similarly exhibits tool and memory bias rates of 22.8% and 24.0%, respectively, despite a bias gap of only -1.1. Therefore, source bias and source intervention capture distinct properties: the former measures *which source is systematically preferred*, whereas the latter reflects *how frequently the model departs from one source in favor of the other*.

Overall, these results reveal no universal source preference across LLMs. Most models exhibit a mild tendency toward memory, while a small number show pronounced memory bias and several others modestly favor the tool. More importantly, the direction and magnitude of bias vary substantially across models, motivating evaluation of tool–memory bias as a model-specific property rather than assuming that LLMs uniformly defer to either external tools or their internal knowledge.

Challenging subsets We further identify the "Both Wrong" subset as a dedicated stress-test benchmark, consisting of instances where both the memory-only and tool-assisted predictions are

Table 5: Tool Bias and Memory Bias of resolution methods. Tool Bias (Sect. 3.2) denotes LLM mistakenly picks Tool as correct, given Memory is correct. A positive / negative bias gap means the LLM is biased towards tool / memory output.

Resolution Method	Resolution Acc (↑)	Tool Bias (↓)	Memory Bias (↓)	Bias Gap
Resolve	44.7	<u>10.2</u>	57.6	-47.4
Full Trace	**65.9**	46.9	<u>10.7</u>	36.1
DoT	18.5	**0.0**	**0.0**	0.0
SR-DCR	52.6	48.3	13.1	35.2
Vigilant	44.1	13.9	55.4	-41.5
XoT	<u>56.8</u>	12.1	42.8	-30.6
Average	47.1	21.9	29.9	-8.0

incorrect and mutually inconsistent. Unlike standard TMC cases, where conflict resolution can often be reduced to selecting the correct source, this subset requires the model to recover beyond both erroneous candidates. It therefore provides a challenging test bed for evaluating whether conflict-resolution methods can perform genuine error correction rather than source selection.

6 Conclusion

We introduce *Tool-Memory Conflict (TMC)*, to the best of our knowledge, the first study of disagreements between an LLM's internal knowledge and information obtained from external tools. Across a broad set of models, we find that TMC is common and highly model-dependent, with conflict rates ranging from 4.8% to 55.7%. Importantly, many conflicts are not simple cases where one source is correct and the other is wrong.

We further show the challenge of resolving conflicts and accounting for source biases. Models that encounter fewer conflicts are not necessarily better at resolving them (e.g. GPT-5.5). Tool use trace contributes to more effective resolution than thinking or debating, which existing context-based resolution methods heavily rely on. We also observe heterogeneous preferences for memory versus tools, including pronounced memory bias in some models. These findings show that reliable tool augmentation requires more than providing external information, and LLMs must also robustly arbitrate between conflicting sources in agentic environments. In summary, TMC provides a systematic basis for evaluating and improving this capability.

AI use statement

In this work, we used generative AI tools for grammar check, fixing latex errors, and improving table presentation quality (for wide and long tables). We have not used generative AI tools for generating data, and the rest of the required disclosure tasks are not applicable to this work. We have reviewed all AI-assisted work. We checked LLM-generated grammar fixes manually and made sure they only improve table quality without touching the content. We take responsibility for the final content of this work, including text, claims or artifacts produced with the aid of generative AI.

Ethics Statement

Our work investigates the trustworthiness of integrating external tools into internal parametric knowledge of tool-augmented Large Language Models (LLMs). We focus on identifying a key issue, where the internal memory and external tools conflict with each other. Tool-memory conflict undermines model reliability in several ways: (1) it erodes user trust, since end-users cannot anticipate whether the LLM will defer to its memorized parameters (which may be outdated or imprecise) or to real-time tool outputs (which may also be noisy or incomplete); (2) it introduces inconsistency in downstream applications—ranging from automated fact-checking to decision-support systems—where conflicting signals can lead to erroneous or unsafe conclusions; and (3) it complicates model evaluation and calibration, since standard accuracy metrics fail to capture the nuanced interplay between static knowledge and dynamic retrieval. By systematically quantifying and characterizing the prevalence

of tool–memory contradictions, our work illuminates how even state-of-the-art LLMs can exhibit unpredictable behavior when faced with conflicting evidence. In doing so, we highlight the necessity of developing robust conflict-resolution mechanisms (e.g., vigilance prompting, retrieval-augmented grounding, and calibrated confidence scoring) that can reconcile or at least surface divergent sources of information. Addressing this gap is essential not only for improving the factual correctness of LLM responses, but also for ensuring transparent, explainable, and trustworthy deployment in high-stakes domains such as healthcare, finance, and legal assistance.

REFERENCES

https://www.anthropic.com/news/claude-opus-4-7, 2026a.

https://www.anthropic.com/news/claude-opus-4-8, 2026b.

https://www.anthropic.com/news/claude-sonnet-4-6, 2026c.

https://openai.com/index/introducing-gpt-5-5/, 2026.

Kinjal Basu, Ibrahim Abdelaziz, Subhajit Chaudhury, Soham Dan, Maxwell Crouse, Asim Munawar, Vernon Austel, Sadhana Kumaravel, Vinod Muthusamy, Pavan Kapanipathi, and Luis Lastras. API-BLEND: A comprehensive corpora for training and benchmarking API LLMs. In Lun-Wei Ku, Andre Martins, and Vivek Srikumar (eds.), *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 12859–12870, Bangkok, Thailand, August 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.acl-long.694. URL https://aclanthology.org/2024.acl-long.694/.

Roy Betser, Amit Giloni, Shamik Bose, Sindhu Padakandla, Chiara Picardi, Lidor Erez, and Roman Vainshtein. Agentrim: Tool risk mitigation for agentic ai. *arXiv preprint arXiv:2601.12449*, 2026.

Hung-Ting Chen, Michael Zhang, and Eunsol Choi. Rich knowledge sources bring complex knowledge conflicts: Recalibrating models to reflect conflicting evidence. In Yoav Goldberg, Zornitsa Kozareva, and Yue Zhang (eds.), *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pp. 2292–2307, Abu Dhabi, United Arab Emirates, December 2022. Association for Computational Linguistics. doi: 10.18653/v1/2022.emnlp-main.146. URL https://aclanthology.org/2022.emnlp-main.146/.

Zhipeng Chen, Kun Zhou, Beichen Zhang, Zheng Gong, Xin Zhao, and Ji-Rong Wen. ChatCoT: Tool-augmented chain-of-thought reasoning on chat-based large language models. In Houda Bouamor, Juan Pino, and Kalika Bali (eds.), *Findings of the Association for Computational Linguistics: EMNLP 2023*, pp. 14777–14790, Singapore, December 2023. Association for Computational Linguistics. doi: 10.18653/v1/2023.findings-emnlp.985. URL https://aclanthology.org/2023.findings-emnlp.985/.

Jiali Cheng and Hadi Amiri. Tool unlearning for tool-augmented LLMs. In *Forty-second International Conference on Machine Learning*, 2025.

Jiali Cheng, Anjishnu Kumar, Roshan Lal, Rishi Rajasekaran, Hani Ramezani, Omar Zia Khan, Oleg Rokhlenko, Sunny Chiu-Webster, Gang Hua, and Hadi Amiri. Webatlas: Actor-critic task-completion with look-ahead action simulation. *arXiv preprint arXiv:2510.22732*, 2025.

Yiyang Feng, Zeming Chen, Haotian Wu, Jiawei Zhou, and Antoine Bosselut. Tracking the limits of knowledge propagation: How LLMs fail at multi-step reasoning with conflicting knowledge. In Vera Demberg, Kentaro Inui, and Lluís Marquez (eds.), *Proceedings of the 19th Conference of the European Chapter of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 5813–5847, Rabat, Morocco, March 2026. Association for Computational Linguistics. ISBN 979-8-89176-380-7. doi: 10.18653/v1/2026.eacl-long.273. URL https://aclanthology.org/2026.eacl-long.273/.

Zorik Gekhman, Jonathan Herzig, Roee Aharoni, Chen Elkind, and Idan Szpektor. TrueTeacher: Learning factual consistency evaluation with large language models. In Houda Bouamor, Juan Pino, and Kalika Bali (eds.), *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pp. 2053–2070, Singapore, December 2023. Association for Computational

Linguistics. doi: 10.18653/v1/2023.emnlp-main.127. URL https://aclanthology.org/2023.emnlp-main.127/.

Aryo Pradipta Gema, Joshua Ong Jun Leang, Giwon Hong, Alessio Devoto, Alberto Carlo Maria Mancino, Rohit Saxena, Xuanli He, Yu Zhao, Xiaotang Du, Mohammad Reza Ghasemi Madani, et al. Are we done with mmlu? In *Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pp. 5069–5096, 2025.

Aaron Grattafiori, Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Alex Vaughan, et al. The llama 3 herd of models. *arXiv preprint arXiv:2407.21783*, 2024.

Peter Hase, Mona Diab, Asli Celikyilmaz, Xian Li, Zornitsa Kozareva, Veselin Stoyanov, Mohit Bansal, and Srinivasan Iyer. Methods for measuring, updating, and visualizing factual beliefs in language models. In Andreas Vlachos and Isabelle Augenstein (eds.), *Proceedings of the 17th Conference of the European Chapter of the Association for Computational Linguistics*, pp. 2714–2731, Dubrovnik, Croatia, May 2023. Association for Computational Linguistics. doi: 10.18653/v1/2023.eacl-main.199. URL https://aclanthology.org/2023.eacl-main.199/.

Dan Hendrycks, Collin Burns, Steven Basart, Andy Zou, Mantas Mazeika, Dawn Song, and Jacob Steinhardt. Measuring massive multitask language understanding. In *International Conference on Learning Representations*, 2021. URL https://openreview.net/forum?id=d7KBjmI3GmQ.

Di Jin, Eileen Pan, Nassim Oufattole, Wei-Hung Weng, Hanyi Fang, and Peter Szolovits. What disease does this patient have? a large-scale open domain question answering dataset from medical exams. *Applied Sciences*, 11(14):6421, 2021.

Qiao Jin, Bhuwan Dhingra, Zhengping Liu, William Cohen, and Xinghua Lu. PubMedQA: A dataset for biomedical research question answering. In Kentaro Inui, Jing Jiang, Vincent Ng, and Xiaojun Wan (eds.), *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pp. 2567–2577, Hong Kong, China, November 2019. Association for Computational Linguistics. doi: 10.18653/v1/D19-1259. URL https://aclanthology.org/D19-1259/.

Zhuoran Jin, Pengfei Cao, Yubo Chen, Kang Liu, Xiaojian Jiang, Jiexin Xu, Li Qiuxia, and Jun Zhao. Tug-of-war between knowledge: Exploring and resolving knowledge conflicts in retrieval-augmented language models. In Nicoletta Calzolari, Min-Yen Kan, Veronique Hoste, Alessandro Lenci, Sakriani Sakti, and Nianwen Xue (eds.), *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pp. 16867–16878, Torino, Italia, May 2024. ELRA and ICCL. URL https://aclanthology.org/2024.lrec-main.1466/.

Jungo Kasai, Keisuke Sakaguchi, yoichi takahashi, Ronan Le Bras, Akari Asai, Xinyan Velocity Yu, Dragomir Radev, Noah A. Smith, Yejin Choi, and Kentaro Inui. Realtime QA: What's the answer right now? In *Thirty-seventh Conference on Neural Information Processing Systems Datasets and Benchmarks Track*, 2023. URL https://openreview.net/forum?id=HfKOIPCvsv.

Kyungjae Lee, Wookje Han, Seung-won Hwang, Hwaran Lee, Joonsuk Park, and Sang-Woo Lee. Plug-and-play adaptation for continuously-updated QA. In Smaranda Muresan, Preslav Nakov, and Aline Villavicencio (eds.), *Findings of the Association for Computational Linguistics: ACL 2022*, pp. 438–447, Dublin, Ireland, May 2022a. Association for Computational Linguistics. doi: 10.18653/v1/2022.findings-acl.37. URL https://aclanthology.org/2022.findings-acl.37/.

Nayeon Lee, Wei Ping, Peng Xu, Mostofa Patwary, Pascale Fung, Mohammad Shoeybi, and Bryan Catanzaro. Factuality enhanced language models for open-ended text generation. In Alice H. Oh, Alekh Agarwal, Danielle Belgrave, and Kyunghyun Cho (eds.), *Advances in Neural Information Processing Systems*, 2022b. URL https://openreview.net/forum?id=LvyJX20Rll.

Guocong Li, Qirui Hu, Ping Wang, Guofeng Zhang, Jian Wu, and Hongxia Xu. Debate-of-thoughts: Resolving knowledge conflicts in LLMs through internal deliberation. In Maria Liakata, Viviane P. Moreira, Jiajun Zhang, and David Jurgens (eds.), *Proceedings of the 64th Annual Meeting of*

the Association for Computational Linguistics (Volume 1: Long Papers), pp. 35674–35696, San Diego, California, United States, July 2026. Association for Computational Linguistics. ISBN 979-8-89176-390-6. doi: 10.18653/v1/2026.acl-long.1651. URL https://aclanthology.org/2026.acl-long.1651/.

Minghao Li, Yingxiu Zhao, Bowen Yu, Feifan Song, Hangyu Li, Haiyang Yu, Zhoujun Li, Fei Huang, and Yongbin Li. API-bank: A comprehensive benchmark for tool-augmented LLMs. In Houda Bouamor, Juan Pino, and Kalika Bali (eds.), *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pp. 3102–3116, Singapore, December 2023. Association for Computational Linguistics. doi: 10.18653/v1/2023.emnlp-main.187. URL https://aclanthology.org/2023.emnlp-main.187/.

Aixin Liu, Bei Feng, Bing Xue, Bingxuan Wang, Bochao Wu, Chengda Lu, Chenggang Zhao, Chengqi Deng, Chenyu Zhang, Chong Ruan, et al. Deepseek-v3 technical report. *arXiv preprint arXiv:2412.19437*, 2024.

Shayne Longpre, Kartik Perisetla, Anthony Chen, Nikhil Ramesh, Chris DuBois, and Sameer Singh. Entity-based knowledge conflicts in question answering. In Marie-Francine Moens, Xuanjing Huang, Lucia Specia, and Scott Wen-tau Yih (eds.), *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pp. 7052–7063, Online and Punta Cana, Dominican Republic, November 2021. Association for Computational Linguistics. doi: 10.18653/v1/2021.emnlp-main.565. URL https://aclanthology.org/2021.emnlp-main.565/.

Xinyuan Lu, Liangming Pan, Yubo Ma, Preslav Nakov, and Min-Yen Kan. TART: An open-source tool-augmented framework for explainable table-based reasoning. In Luis Chiruzzo, Alan Ritter, and Lu Wang (eds.), *Findings of the Association for Computational Linguistics: NAACL 2025*, pp. 4323–4339, Albuquerque, New Mexico, April 2025. Association for Computational Linguistics. ISBN 979-8-89176-195-7. doi: 10.18653/v1/2025.findings-naacl.244. URL https://aclanthology.org/2025.findings-naacl.244/.

Yubo Ma, Zhibin Gou, Junheng Hao, Ruochen Xu, Shuohang Wang, Liangming Pan, Yujiu Yang, Yixin Cao, Aixin Sun, Hany Awadalla, et al. Sciagent: Tool-augmented language models for scientific reasoning. *arXiv preprint arXiv:2402.11451*, 2024.

Yikang Pan, Liangming Pan, Wenhu Chen, Preslav Nakov, Min-Yen Kan, and William Wang. On the risk of misinformation pollution with large language models. In Houda Bouamor, Juan Pino, and Kalika Bali (eds.), *Findings of the Association for Computational Linguistics: EMNLP 2023*, pp. 1389–1403, Singapore, December 2023. Association for Computational Linguistics. doi: 10.18653/v1/2023.findings-emnlp.97. URL https://aclanthology.org/2023.findings-emnlp.97/.

Aaron Parisi, Yao Zhao, and Noah Fiedel. Talm: Tool augmented language models. *arXiv preprint arXiv:2205.12255*, 2022.

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

Jirui Qi, Raquel Fernández, and Arianna Bisazza. Cross-lingual consistency of factual knowledge in multilingual language models. In Houda Bouamor, Juan Pino, and Kalika Bali (eds.), *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pp. 10650–10666, Singapore, December 2023. Association for Computational Linguistics. doi: 10.18653/v1/2023.emnlp-main.658. URL https://aclanthology.org/2023.emnlp-main.658/.

Yujia Qin, Shengding Hu, Yankai Lin, Weize Chen, Ning Ding, Ganqu Cui, Zheni Zeng, Xuanhe Zhou, Yufei Huang, Chaojun Xiao, et al. Tool learning with foundation models. *ACM Computing Surveys*, 57(4):1–40, 2024a.

Yujia Qin, Shihao Liang, Yining Ye, Kunlun Zhu, Lan Yan, Yaxi Lu, Yankai Lin, Xin Cong, Xiangru Tang, Bill Qian, Sihan Zhao, Lauren Hong, Runchu Tian, Ruobing Xie, Jie Zhou, Mark Gerstein, dahai li, Zhiyuan Liu, and Maosong Sun. ToolLLM: Facilitating large language models to master 16000+ real-world APIs. In *The Twelfth International Conference on Learning Representations*, 2024b. URL https://openreview.net/forum?id=dHng2O0Jjr.

Timo Schick, Jane Dwivedi-Yu, Roberto Dessi, Roberta Raileanu, Maria Lomeli, Eric Hambro, Luke Zettlemoyer, Nicola Cancedda, and Thomas Scialom. Toolformer: Language models can teach themselves to use tools. In *Thirty-seventh Conference on Neural Information Processing Systems*, 2023. URL https://openreview.net/forum?id=Yacmpz84TH.

Dan Shi, Renren Jin, Tianhao Shen, Weilong Dong, Xinwei Wu, and Deyi Xiong. IRCAN: Mitigating knowledge conflicts in LLM generation via identifying and reweighting context-aware neurons. In *The Thirty-eighth Annual Conference on Neural Information Processing Systems*, 2024. URL https://openreview.net/forum?id=ZfXRAqbBKX.

Hexiang Tan, Fei Sun, Wanli Yang, Yuanzhuo Wang, Qi Cao, and Xueqi Cheng. Blinded by generated contexts: How language models merge generated and retrieved contexts when knowledge conflicts? In Lun-Wei Ku, Andre Martins, and Vivek Srikumar (eds.), *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 6207–6227, Bangkok, Thailand, August 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.acl-long.337. URL https://aclanthology.org/2024.acl-long.337/.

Qiaoyu Tang, Ziliang Deng, Hongyu Lin, Xianpei Han, Qiao Liang, Boxi Cao, and Le Sun. Toolalpaca: Generalized tool learning for language models with 3000 simulated cases. *arXiv preprint arXiv:2306.05301*, 2023.

Joshua Vendrow, Edward Vendrow, Sara Beery, and Aleksander Madry. Do large language model benchmarks test reliability? *arXiv preprint arXiv:2502.03461*, 2025.

Alexander Wan, Eric Wallace, and Dan Klein. What evidence do language models find convincing? In Lun-Wei Ku, Andre Martins, and Vivek Srikumar (eds.), *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 7468–7484, Bangkok, Thailand, August 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.acl-long.403. URL https://aclanthology.org/2024.acl-long.403/.

Fei Wang, Wenjie Mo, Yiwei Wang, Wenxuan Zhou, and Muhao Chen. A causal view of entity bias in (large) language models. In Houda Bouamor, Juan Pino, and Kalika Bali (eds.), *Findings of the Association for Computational Linguistics: EMNLP 2023*, pp. 15173–15184, Singapore, December 2023. Association for Computational Linguistics. doi: 10.18653/v1/2023.findings-emnlp.1013. URL https://aclanthology.org/2023.findings-emnlp.1013/.

Ping Wang, Xiangguo Sun, Bingbing Xu, Guocong Li, and Xiaofeng Meng. From passive response to proactive correction: Enhancing llm robustness against input fact perturbations. *arXiv preprint arXiv:2608.25894*, 2026.

Yike Wang, Shangbin Feng, Heng Wang, Weijia Shi, Vidhisha Balachandran, Tianxing He, and Yulia Tsvetkov. Resolving knowledge conflicts in large language models. In *First Conference on Language Modeling*, 2024. URL https://openreview.net/forum?id=ptvV5HGTNN.

Jason Wei, Nguyen Karina, Hyung Won Chung, Yunxin Joy Jiao, Spencer Papay, Amelia Glaese, John Schulman, and William Fedus. Measuring short-form factuality in large language models. *arXiv preprint arXiv:2411.04368*, 2024.

Kevin Wu, Eric Wu, and James Zou. Clasheval: Quantifying the tug-of-war between an LLM's internal prior and external evidence. In *The Thirty-eight Conference on Neural Information Processing Systems Datasets and Benchmarks Track*, 2024. URL https://openreview.net/forum?id=WGoCZl2itU.

Rongwu Xu, Zehan Qi, Zhijiang Guo, Cunxiang Wang, Hongru Wang, Yue Zhang, and Wei Xu. Knowledge conflicts for LLMs: A survey. In Yaser Al-Onaizan, Mohit Bansal, and Yun-Nung Chen (eds.), *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pp. 8541–8565, Miami, Florida, USA, November 2024a. Association for Computational Linguistics. doi: 10.18653/v1/2024.emnlp-main.486. URL https://aclanthology.org/2024.emnlp-main.486/.

Rongwu Xu, Zehan Qi, Zhijiang Guo, Cunxiang Wang, Hongru Wang, Yue Zhang, and Wei Xu. Knowledge conflicts for llms: A survey. *arXiv preprint arXiv:2403.08319*, 2024b.

An Yang, Anfeng Li, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Gao, Chengen Huang, Chenxu Lv, et al. Qwen3 technical report. *arXiv preprint arXiv:2505.09388*, 2025.

Michael Zhang and Eunsol Choi. SituatedQA: Incorporating extra-linguistic contexts into QA. In Marie-Francine Moens, Xuanjing Huang, Lucia Specia, and Scott Wen-tau Yih (eds.), *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pp. 7371–7387, Online and Punta Cana, Dominican Republic, November 2021. Association for Computational Linguistics. doi: 10.18653/v1/2021.emnlp-main.586. URL https://aclanthology.org/2021.emnlp-main.586/.

Michael Zhang and Eunsol Choi. Mitigating temporal misalignment by discarding outdated facts. In Houda Bouamor, Juan Pino, and Kalika Bali (eds.), *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pp. 14213–14226, Singapore, December 2023. Association for Computational Linguistics. doi: 10.18653/v1/2023.emnlp-main.879. URL https://aclanthology.org/2023.emnlp-main.879/.

Zhange Zhang, Zhicheng Geng, Yuqing Ma, Tianbo Wang, Kai Lv, and Xianglong Liu. Conflict-aware knowledge editing in the wild: Semantic-augmented graph representation for unstructured text. In *The Thirty-ninth Annual Conference on Neural Information Processing Systems*, 2025. URL https://openreview.net/forum?id=wHm5J9uanV.

Jun Zhao, Yongzhuo Yang, Xiang Hu, Jingqi Tong, Yi Lu, Wei Wu, Tao Gui, Qi Zhang, and Xuanjing Huang. Understanding parametric and contextual knowledge reconciliation within large language models. In *The Thirty-ninth Annual Conference on Neural Information Processing Systems*, 2025. URL https://openreview.net/forum?id=76cFMRgEzQ.

Tianzhe Zhao, Jiaoyan Chen, Shuxiu Zhang, Haiping Zhu, Qika Lin, and Jun Liu. Exploring knowledge conflicts for faithful llm reasoning: Benchmark and method. In *Proceedings of the 49th International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '26, pp. 2464–2474, New York, NY, USA, 2026. Association for Computing Machinery. ISBN 9798400725999. doi: 10.1145/3805712.3809650. URL https://doi.org/10.1145/3805712.3809650.

Wenxuan Zhou, Sheng Zhang, Hoifung Poon, and Muhao Chen. Context-faithful prompting for large language models. In Houda Bouamor, Juan Pino, and Kalika Bali (eds.), *Findings of the Association for Computational Linguistics: EMNLP 2023*, pp. 14544–14556, Singapore, December 2023. Association for Computational Linguistics. doi: 10.18653/v1/2023.findings-emnlp.968. URL https://aclanthology.org/2023.findings-emnlp.968/.

Zeqi Zhou, Fang Wu, Shayan Talaei, Haokai Zhao, Cheng Meixin, Tinson Xu, Amin Saberi, and Yejin Choi. When to trust context: Self-reflective debates for context reliability. *arXiv preprint arXiv:2506.06020*, 2025.

A Examples of Tool-Memory Conflict

TMCs can arise in diverse settings; we illustrate with three representative scenarios.

Example 1: Math Reasoning Given a numerical query, an LLM may produce an answer via learned knowledge, while invoking a calculator tool yields an exact result.

Example 2: Code Execution An LLM may infer the output of a code snippet using its internal knowledge, while calling a compiler / interpreter produces verifiable output.

Example 3: Medical Diagnosis An LLM may respond to clinical questions based on literature memorized during training, while querying an external medical database via API returns structured and recent findings.

In each of these cases, the core tension lies in reconciling the LLM's static parametric knowledge with the dynamic outputs obtained through tool invocation. Characterizing and resolving TMCs is essential for ensuring the reliability and trustworthiness of agentic systems.

B Importance of Tool-Memory Conflict

Understanding how LLMs handle conflicts between parametric memory and external tools is important for several reasons.

Agentic System Safety Tool-augmented LLMs increasingly act autonomously—executing code, querying databases, and calling APIs. When parametric memory and tool outputs disagree, the model must implicitly decide which to trust. Unexamined, this decision is neither predictable nor controllable, creating a silent failure mode in deployed agents.

Calibrated Tool Use An effective agent should invoke tools when its internal knowledge is likely insufficient and trust its parameters when tools are unreliable or unavailable. Studying TMC directly informs when and whether to call a tool—a prerequisite for efficient, cost-aware tool-use policies.

Diagnosing Knowledge Deficiencies Systematic conflicts in one direction reveal actionable deficiencies: consistent model errors expose training data gaps or staleness; consistent tool errors expose unreliable APIs or data sources. TMC thus provides a diagnostic lens for improving both the model and its tool ecosystem.

Robustness to Adversarial Manipulation If a model unconditionally defers to tool outputs, an adversary who controls a tool endpoint can override the model's correct beliefs. Conversely, if the model ignores tools, it becomes immune to correction. Understanding TMC is necessary to design systems that are neither gullible nor stubborn.

C Distinction from Other Knowledge Conflicts

Tool-memory conflict is closely related to context-memory conflict (Lee et al., 2022a; Zhou et al., 2023; Tan et al., 2024) and inter-context conflict (Zhang & Choi, 2021; Kasai et al., 2023; Wan et al., 2024). Importantly, in our resolution setting, the conflicting memory- and tool-derived answers are explicitly presented to the model. Thus, we do not distinguish TMC from prior knowledge-conflict settings based on whether conflicting information is simultaneously observable. Instead, the key distinction lies in the *provenance and generation process* of the external evidence.

Action-Generated vs. Exogenous Evidence In context-memory conflict, external knowledge is typically supplied as contextual evidence, e.g., a retrieved document or passage, and the model must reconcile this evidence with its parametric knowledge. In TMC, by contrast, the external answer is produced through an interaction between the model and an external tool. The model must select a tool, construct a structured invocation, observe its execution result, and interpret that result. Thus,

tool-derived knowledge is causally downstream of the model's own actions rather than being a fixed piece of exogenous context.

Multi-Stage Error Structure This interaction introduces failure modes that are absent from the standard context-memory formulation. A conflicting tool-derived answer may result from selecting an inappropriate tool, constructing incorrect arguments, receiving an erroneous or incomplete tool result, or incorrectly interpreting an otherwise correct result. Therefore, disagreement between memory and tool outputs reflects not only disagreement between knowledge sources, but also potential failures along the tool-use pipeline.

Executable Provenance and Interaction Traces Tool outputs are accompanied by operational provenance: the invoked tool, its arguments, intermediate execution states, and returned result. Consequently, resolving a tool-memory conflict can involve reasoning about how an answer was produced in addition to evaluating the answer itself. This distinguishes TMC from settings in which the external source is represented only as textual evidence, and motivates our evaluation of whether exposing the full tool interaction trace improves conflict resolution.

Interventional Resolution Tool-mediated conflicts also arise within an interactive environment. In principle, an agent confronted with disagreement can acquire additional evidence by re-invoking a tool, modifying its arguments, or consulting alternative tools. This makes tool-memory conflict naturally connected to sequential source verification and error recovery. Our current benchmark isolates the arbitration stage by presenting the memory- and tool-derived answers jointly, while preserving the distinct tool-generation process that produced the external candidate.

D ADDITIONAL RESULTS

We present the results in Table 6–Table 9.

Table 6: Prevalence of Tool-Memory Conflict (TMC) across datasets, with memory-only (M) and tool-only (T) performance. 0 and 1 denote result is wrong and correct, respectively. Same for memory-only.

LLM	Overall Acc (↑)			No Conflict Rate			Conflict Rate			
	M=1	T=1	Either=1	Total (↑)	Both=1	Both=0	Total (↓)	Both=0	M=1,T=0	M=0,T=1
GSM8K-Platinum										
Proprietary										
GPT-5.5	**99.2**	**99.4**	**99.7**	**99.1**	98.8	0.3	**0.9**	0.0	0.3	0.6
GPT-5.4	79.2	97.8	98.8	78.4	78.3	0.2	21.6	1.1	0.9	19.6
Claude Opus 4.8	90.1	93.0	99.2	84.4	83.9	0.5	15.6	0.3	6.1	9.1
Claude Opus 4.7	93.5	99.2	99.6	93.5	93.1	0.3	6.5	0.1	0.4	6.0
Claude Sonnet 4.6	98.8	85.8	99.5	85.3	85.1	0.2	14.7	0.2	13.7	0.7
Open-Source										
DeepSeek V3.2	74.5	92.9	96.4	71.6	71.0	0.7	28.4	2.9	3.6	21.9
Kimi K2.5	91.6	97.2	99.3	90.0	89.5	0.5	10.0	0.2	2.1	7.7
Llama 3.3 70B	46.3	92.7	96.0	44.1	43.1	1.0	55.9	3.0	3.2	49.7
Qwen3 32B	39.9	81.5	85.9	37.1	35.5	1.7	62.9	12.4	4.4	46.0
Llama 3.1 8B	16.6	44.8	51.4	14.2	10.0	4.2	85.8	44.4	6.6	34.8
Qwen3 8B	19.9	70.9	75.6	17.5	15.2	2.3	82.5	22.2	4.7	55.6
MATH										
Proprietary										
GPT-5.5	**84.6**	**87.9**	**89.7**	**92.0**	82.7	9.2	**8.0**	1.0	1.9	5.1
GPT-5.4	68.5	84.2	86.7	74.3	65.9	8.4	25.7	4.8	2.6	18.2
Claude Opus 4.8	79.5	83.4	86.1	88.1	76.7	11.4	11.9	2.5	2.7	6.6
Claude Opus 4.7	81.3	83.9	85.5	91.7	79.5	12.1	8.3	2.3	1.6	4.3
Claude Sonnet 4.6	83.0	79.5	84.8	90.6	77.6	13.1	9.4	2.1	5.4	1.8

Open-Source										
DeepSeek V3.2	72.6	79.6	84.9	76.2	67.3	8.8	23.8	6.3	5.3	12.3
Kimi K2.5	81.1	85.7	88.5	87.3	78.2	9.1	12.7	2.4	2.8	7.4
Llama 3.3 70B	57.8	79.9	83.9	62.3	53.7	8.5	37.7	7.6	4.0	26.1
Qwen3 32B	57.7	70.9	78.8	59.8	49.8	9.9	40.2	11.2	7.9	21.1
Llama 3.1 8B	34.6	55.0	67.4	28.1	22.2	5.9	71.9	26.7	12.3	32.8
Qwen3 8B	34.2	65.4	72.2	36.2	27.5	8.8	63.8	19.0	6.8	38.0
MMLU-Redux2										
Proprietary										
GPT-5.5	**97.1**	**96.5**	**97.3**	**99.0**	96.3	2.7	**1.0**	0.0	0.8	0.2
GPT-5.4	90.3	93.8	95.5	92.6	88.6	4.0	7.4	0.5	1.7	5.2
Claude Opus 4.8	94.6	95.9	97.0	96.3	93.5	2.8	3.7	0.2	1.1	2.4
Claude Opus 4.7	94.7	95.9	97.0	96.4	93.6	2.8	3.6	0.2	1.1	2.3
Claude Sonnet 4.6	95.6	95.2	96.9	96.7	93.9	2.9	3.3	0.2	1.8	1.3
Open-Source										
DeepSeek V3.2	91.8	93.3	96.5	91.8	88.6	3.2	8.2	0.4	3.2	4.7
Kimi K2.5	94.7	93.8	96.5	95.1	92.0	3.1	4.9	0.4	2.7	1.8
Llama 3.3 70B	84.1	86.9	90.6	88.5	80.4	8.1	11.5	1.3	3.7	6.5
Qwen3 32B	85.4	86.8	91.5	87.6	80.7	6.9	12.4	1.6	4.7	6.1
Llama 3.1 8B	72.9	67.2	81.8	69.9	58.3	11.6	30.1	6.6	14.6	8.9
Qwen3 8B	82.0	84.4	88.5	87.9	77.9	9.9	12.1	1.6	4.1	6.5
MedQA										
Proprietary										
GPT-5.5	**98.2**	**96.6**	**98.3**	**98.2**	96.5	1.7	**1.8**	0.0	1.7	0.1
GPT-5.4	89.8	91.5	94.0	92.6	87.3	5.2	7.4	0.8	2.5	4.2
Claude Opus 4.8	92.9	93.2	94.6	96.0	91.5	4.4	4.0	1.0	1.4	1.7
Claude Opus 4.7	93.1	93.7	95.1	95.9	91.6	4.3	4.1	0.6	1.4	2.0
Claude Sonnet 4.6	90.8	90.6	92.8	94.6	88.6	6.0	5.4	1.3	2.1	2.0
Open-Source										
DeepSeek V3.2	81.4	87.1	90.8	84.8	77.7	7.1	15.2	2.1	3.7	9.4
Kimi K2.5	85.6	84.6	91.0	85.9	79.2	6.7	14.1	2.4	6.4	5.4
Llama 3.3 70B	74.0	74.9	78.7	88.5	70.3	18.2	11.5	3.1	3.7	4.7
Qwen3 32B	68.1	68.5	75.8	78.9	60.8	18.1	21.1	6.1	7.3	7.7
Llama 3.1 8B	64.7	62.0	73.1	70.3	53.6	16.7	29.7	10.2	11.1	8.3
Qwen3 8B	59.0	60.4	67.7	76.5	51.7	24.8	23.5	7.5	7.3	8.7
PubMedQA										
Proprietary										
GPT-5.5	**74.6**	**75.7**	**81.5**	**84.0**	68.9	15.1	**16.0**	3.4	5.7	6.9
GPT-5.4	60.0	66.5	77.3	61.7	49.2	12.5	38.3	10.2	10.8	17.3
Claude Opus 4.8	65.7	59.1	76.5	61.5	48.3	13.2	38.5	10.3	17.4	10.7
Claude Opus 4.7	66.3	67.8	78.5	70.8	55.7	15.1	29.2	6.4	10.7	12.2
Claude Sonnet 4.6	60.2	54.9	72.0	58.9	43.1	15.8	41.1	12.2	17.1	11.8
Open-Source										
DeepSeek V3.2	51.8	68.8	73.6	66.2	47.0	19.2	33.8	7.2	4.8	21.7
Kimi K2.5	57.0	59.3	71.4	63.0	44.9	18.1	37.0	10.5	12.1	14.4
Llama 3.3 70B	60.7	52.6	72.3	55.8	41.0	14.8	44.2	12.9	19.7	11.6
Qwen3 32B	54.5	62.2	73.1	62.3	43.5	18.8	37.7	8.1	10.9	18.7
Llama 3.1 8B	58.1	60.7	74.2	62.0	44.6	17.4	38.0	8.4	13.5	16.1
Qwen3 8B	43.8	54.5	67.0	51.5	31.3	20.2	48.5	12.7	12.5	23.2
SimpleQA										
Proprietary										

GPT-5.5	**99.4**	**98.3**	**99.9**	**98.0**	97.9	0.1	**2.0**	0.0	1.5	0.4
GPT-5.4	75.5	94.3	98.7	72.3	71.1	1.3	27.7	0.0	4.4	23.3
Claude Opus 4.8	79.7	95.1	98.1	78.6	76.7	1.9	21.4	0.0	3.0	18.4
Claude Opus 4.7	80.8	95.4	98.4	79.6	77.9	1.6	20.4	0.0	2.9	17.5
Claude Sonnet 4.6	85.1	91.5	98.0	80.5	78.6	2.0	19.5	0.0	6.6	12.9
Open-Source										
DeepSeek V3.2	53.7	93.8	97.0	53.6	50.5	3.0	46.4	0.0	3.2	43.2
Kimi K2.5	95.0	95.1	99.5	91.2	90.7	0.5	8.8	0.0	4.3	4.5
Llama 3.3 70B	84.3	47.8	91.1	49.9	40.9	8.9	50.1	0.0	43.3	6.8
Qwen3 32B	71.0	87.8	95.9	67.0	62.9	4.1	33.0	0.0	8.1	24.8
Llama 3.1 8B	86.5	24.5	90.5	30.0	20.5	9.5	70.0	0.0	66.0	4.0
Qwen3 8B	83.7	42.2	90.2	45.6	35.8	9.8	54.4	0.0	47.9	6.5

Table 7: Prevalence of Tool-Memory Conflict (TMC) across tasks, with memory-only (M) and tool-only (T) performance. 0 and 1 denote result is wrong and correct, respectively. Same for memory-only.

LLM	**Overall Acc (↑)**			**No Conflict Rate**			**Conflict Rate**			
	M=1	**T=1**	**Either=1**	**Total (↑)**	**Both=1**	**Both=0**	**Total (↓)**	**Both=0**	**M=1,T=0**	**M=0,T=1**
Arts & Entertainment										
Proprietary										
GPT-5.5	**99.5**	**98.2**	**100.0**	**97.7**	97.7	0.0	**2.3**	0.0	1.8	0.5
GPT-5.4	73.2	95.5	**100.0**	68.8	68.8	0.0	31.2	0.0	4.5	26.8
Claude Opus 4.8	79.1	96.6	99.6	76.5	76.1	0.4	23.5	0.0	3.0	20.5
Claude Opus 4.7	81.5	96.9	99.6	79.3	78.9	0.4	20.7	0.0	2.6	18.1
Claude Sonnet 4.6	80.7	91.2	98.4	75.1	73.5	1.6	24.9	0.0	7.2	17.7
Open-Source										
DeepSeek V3.2	50.8	95.4	97.0	52.3	49.2	3.0	47.7	0.0	1.5	46.2
Kimi K2.5	93.5	95.8	99.5	90.2	89.8	0.5	9.8	0.0	3.7	6.0
Llama 3.3 70B	80.8	44.0	89.6	45.6	35.2	10.4	54.4	0.0	45.6	8.8
Qwen3 32B	68.3	88.3	96.4	63.7	60.1	3.6	36.3	0.0	8.2	28.1
Llama 3.1 8B	86.6	14.2	89.5	21.8	11.3	10.5	78.2	0.0	75.3	2.9
Qwen3 8B	78.9	38.4	84.9	47.3	32.3	15.1	52.7	0.0	46.6	6.1
Math										
Proprietary										
GPT-5.5	**89.1**	**91.4**	**92.8**	**94.1**	87.7	6.5	**5.9**	0.7	1.4	3.7
GPT-5.4	71.9	88.3	90.4	75.7	69.7	6.0	24.3	3.6	2.1	18.5
Claude Opus 4.8	82.7	86.5	90.0	87.3	79.1	8.1	12.7	1.8	3.5	7.3
Claude Opus 4.7	85.0	88.5	89.8	92.1	83.6	8.6	7.9	1.6	1.3	4.8
Claude Sonnet 4.6	87.7	81.9	89.3	89.5	80.3	9.2	10.5	1.5	7.4	1.5
Open-Source										
DeepSeek V3.2	73.6	83.7	88.5	75.2	68.9	6.4	24.8	5.2	4.7	14.9
Kimi K2.5	83.3	87.8	90.5	88.0	80.5	7.5	12.0	2.0	2.7	7.3
Llama 3.3 70B	55.0	83.8	87.5	57.6	51.2	6.4	42.4	6.1	3.7	32.5
Qwen3 32B	53.6	74.2	81.2	54.3	46.6	7.6	45.7	11.2	7.0	27.6
Llama 3.1 8B	31.2	52.4	63.7	26.0	19.9	6.1	74.0	30.2	11.3	32.5
Qwen3 8B	31.4	67.9	73.9	32.3	25.4	6.9	67.7	19.2	6.0	42.5
Medicine										
Proprietary										
GPT-5.5	**91.0**	**90.1**	**93.1**	**93.8**	87.9	5.9	**6.2**	1.0	3.0	2.2
GPT-5.4	77.7	81.5	87.2	80.1	71.9	8.2	19.9	4.6	5.8	9.5
Claude Opus 4.8	82.5	80.1	87.7	82.5	74.8	7.6	17.5	4.7	7.6	5.3

Claude Opus 4.7	83.7	84.6	89.4	86.8	78.9	7.9	13.2	2.7	4.8	5.6
Claude Sonnet 4.6	78.8	76.9	84.8	80.6	70.9	9.7	19.4	5.5	7.9	6.0
Open-Source										
DeepSeek V3.2	66.5	77.7	82.2	75.2	62.0	13.2	24.8	4.7	4.5	15.7
Kimi K2.5	73.4	73.6	82.6	76.0	64.4	11.6	24.0	5.9	9.0	9.2
Llama 3.3 70B	68.6	65.3	76.4	73.5	57.5	16.1	26.5	7.6	11.1	7.8
Qwen3 32B	63.8	67.1	75.8	72.8	55.1	17.7	27.2	6.5	8.7	12.0
Llama 3.1 8B	61.8	62.1	74.7	65.9	49.3	16.6	34.1	8.8	12.5	12.9
Qwen3 8B	52.4	59.0	69.1	63.2	42.2	21.0	36.8	9.9	10.1	16.7
Other										
Proprietary										
GPT-5.5	**99.4**	**99.4**	**100.0**	**98.7**	98.7	0.0	**1.3**	0.0	0.6	0.6
GPT-5.4	74.9	92.0	97.5	71.9	69.3	2.5	28.1	0.0	5.5	22.6
Claude Opus 4.8	75.5	93.5	97.3	74.5	71.7	2.7	25.5	0.0	3.8	21.7
Claude Opus 4.7	74.0	91.0	96.6	71.8	68.4	3.4	28.2	0.0	5.6	22.6
Claude Sonnet 4.6	83.2	90.6	97.4	79.1	76.4	2.6	20.9	0.0	6.8	14.1
Open-Source										
DeepSeek V3.2	52.1	93.2	95.9	53.4	49.3	4.1	46.6	0.0	2.7	43.8
Kimi K2.5	94.8	95.4	99.4	91.3	90.8	0.6	8.7	0.0	4.0	4.6
Llama 3.3 70B	86.0	48.7	93.3	48.0	41.3	6.7	52.0	0.0	44.7	7.3
Qwen3 32B	74.0	88.6	95.4	71.7	67.1	4.6	28.3	0.0	6.8	21.5
Llama 3.1 8B	81.9	24.5	88.3	29.8	18.1	11.7	70.2	0.0	63.8	6.4
Qwen3 8B	83.9	42.9	91.2	44.2	35.5	8.8	55.8	0.0	48.4	7.4
Science & Tech										
Proprietary										
GPT-5.5	**98.1**	**98.3**	**98.3**	**99.8**	98.1	1.7	**0.2**	0.0	0.0	0.2
GPT-5.4	83.9	93.4	95.8	85.1	81.5	3.6	14.9	0.6	2.5	11.9
Claude Opus 4.8	90.7	95.1	96.4	92.8	89.4	3.4	7.2	0.2	1.3	5.7
Claude Opus 4.7	92.2	95.4	96.4	94.5	91.2	3.2	5.5	0.4	1.0	4.2
Claude Sonnet 4.6	94.1	94.1	96.6	94.9	91.6	3.2	5.1	0.2	2.5	2.5
Open-Source										
DeepSeek V3.2	79.6	92.6	97.2	77.5	75.1	2.5	22.5	0.4	4.6	17.5
Kimi K2.5	96.1	95.6	98.0	95.1	93.6	1.5	4.9	0.5	2.5	2.0
Llama 3.3 70B	79.6	76.4	87.8	79.0	68.3	10.8	21.0	1.4	11.4	8.2
Qwen3 32B	81.0	84.8	92.2	80.5	73.6	6.9	19.5	0.9	7.4	11.2
Llama 3.1 8B	78.7	47.8	85.4	51.6	41.1	10.5	48.4	4.1	37.6	6.7
Qwen3 8B	82.4	75.0	89.2	77.7	68.1	9.6	22.3	1.2	14.3	6.8
Social Sci & Humanities										
Proprietary										
GPT-5.5	**97.5**	**96.6**	**97.6**	**98.9**	96.5	2.4	**1.1**	0.0	1.0	0.1
GPT-5.4	89.7	94.7	96.8	90.6	87.6	3.0	9.4	0.3	2.1	7.0
Claude Opus 4.8	91.7	95.4	97.1	92.9	90.0	2.9	7.1	0.0	1.7	5.4
Claude Opus 4.7	91.8	95.9	97.2	93.4	90.6	2.8	6.6	0.0	1.3	5.3
Claude Sonnet 4.6	93.8	93.7	97.3	92.8	90.2	2.5	7.2	0.2	3.6	3.5
Open-Source										
DeepSeek V3.2	81.5	93.4	96.3	81.9	78.6	3.4	18.1	0.3	2.9	14.8
Kimi K2.5	95.0	94.2	97.1	94.8	92.1	2.7	5.2	0.2	2.9	2.1
Llama 3.3 70B	89.7	79.3	92.9	82.8	76.0	6.8	17.2	0.3	13.6	3.3
Qwen3 32B	83.5	89.5	93.8	84.2	79.1	5.0	15.8	1.1	4.3	10.4
Llama 3.1 8B	79.8	56.7	86.3	59.9	50.2	9.6	40.1	4.1	29.6	6.5
Qwen3 8B	87.3	72.7	91.5	76.4	68.5	7.8	23.6	0.6	18.8	4.2

Table 8: Prevalence of Tool-Memory Conflict (TMC) across datasets, with memory-only (M) and tool-only (T) performance. 0 and 1 denote result is wrong and correct, respectively. Same for memory-only.

LLM	TMC	Resolution Accuracy					
	Rate (↓)	Avg. (↑)	Resolve (↑)	Full Trace (↑)	DoT (↑)	SR-DCR (↑)	Vigilant (↑)	XoT (↑)
GSM8K-Platinum								
Proprietary								
GPT-5.5	**0.9**	54.5	63.6	27.3	63.6	54.5	27.3	
GPT-5.4	21.6	36.5	92.7	8.5	**91.5**	36.5	75.8	
Claude Opus 4.8	15.6	48.4	92.6	8.0	55.9	62.2	88.8	
Claude Opus 4.7	6.5	39.2	89.9	19.0	87.3	77.2	82.3	
Claude Sonnet 4.6	14.7	**92.7**	**94.9**	1.1	16.4	**93.2**	**96.6**	
Open-Source								
DeepSeek V3.2	28.4	49.5	78.5	40.3	66.0	55.8	72.3	
Kimi K2.5	10.0	51.2	79.1	2.3	72.1	46.5	93.0	
Llama 3.3 70B	55.9	33.2	87.9	42.7	84.6	35.1	69.4	
Qwen3 32B	62.9	27.0	74.5	10.4	71.4	37.3	55.8	
Llama 3.1 8B	85.8	12.8	37.2	24.0	15.9	12.3	26.9	
Qwen3 8B	82.5	36.4	70.5	**67.1**	65.1	37.4	70.2	
MATH								
Proprietary								
GPT-5.5	**8.0**	39.8	34.6	2.8	72.4	40.2	35.0	
GPT-5.4	25.7	35.0	71.2	9.6	**76.0**	34.6	54.0	
Claude Opus 4.8	11.9	52.8	79.3	3.7	60.1	57.7	75.1	
Claude Opus 4.7	8.3	59.3	72.8	8.2	62.7	70.1	73.5	
Claude Sonnet 4.6	9.4	**80.0**	**83.7**	0.7	30.7	**82.7**	82.3	
Open-Source								
DeepSeek V3.2	23.8	49.4	61.6	0.0	47.5	54.1	59.1	
Kimi K2.5	12.7	74.7	79.2	0.0	60.6	72.7	**82.7**	
Llama 3.3 70B	37.7	37.6	71.4	3.3	67.8	39.2	57.9	
Qwen3 32B	40.2	45.4	66.1	0.0	52.8	50.5	55.6	
Llama 3.1 8B	71.9	28.5	53.6	14.6	26.6	23.8	37.7	
Qwen3 8B	63.8	43.8	70.0	**60.9**	56.1	46.1	69.3	
MMLU-Redux2								
Proprietary								
GPT-5.5	**1.0**	60.0	46.7	13.3	46.7	53.3	53.3	
GPT-5.4	7.4	49.1	76.3	**55.3**	**75.4**	53.5	57.0	
Claude Opus 4.8	3.7	41.8	63.6	1.8	69.1	40.0	47.3	
Claude Opus 4.7	3.6	57.1	69.6	7.1	75.0	67.9	**75.0**	
Claude Sonnet 4.6	3.3	58.0	64.0	4.0	48.0	56.0	60.0	
Open-Source								
DeepSeek V3.2	8.2	50.0	72.9	40.0	61.4	51.4	61.4	
Kimi K2.5	4.9	**69.0**	**81.0**	10.3	46.6	**70.7**	70.7	
Llama 3.3 70B	11.5	50.3	53.2	29.5	59.5	50.9	56.1	
Qwen3 32B	12.4	49.5	54.6	18.4	58.2	49.0	52.0	
Llama 3.1 8B	30.1	48.9	50.0	15.1	45.7	47.8	51.8	
Qwen3 8B	12.1	37.6	59.0	49.4	54.5	40.4	56.2	
MedQA								
Proprietary								
GPT-5.5	**1.8**	56.2	56.2	25.0	**75.0**	**62.5**	62.5	
GPT-5.4	7.4	50.0	56.8	**54.5**	64.8	45.5	62.5	
Claude Opus 4.8	4.0	44.0	50.0	0.0	46.0	46.0	60.0	

Claude Opus 4.7	4.1	54.9	52.9	2.0	62.7	52.9	62.7
Claude Sonnet 4.6	5.4	47.1	61.8	1.5	38.2	47.1	58.8
Open-Source							
DeepSeek V3.2	15.2	32.6	53.5	47.7	48.8	40.7	55.8
Kimi K2.5	14.1	**56.9**	**72.2**	12.5	31.9	58.3	**66.0**
Llama 3.3 70B	11.5	37.8	39.6	18.0	41.4	39.6	45.0
Qwen3 32B	21.1	35.7	36.0	17.4	40.7	40.3	44.2
Llama 3.1 8B	29.7	38.8	41.9	10.6	35.6	38.1	38.1
Qwen3 8B	23.5	34.4	39.3	39.3	36.9	33.6	38.5
PubMedQA							
Proprietary							
GPT-5.5	**16.0**	25.7	52.9	4.3	25.7	24.3	30.0
GPT-5.4	38.3	48.0	61.8	8.7	31.2	47.2	23.3
Claude Opus 4.8	38.5	47.8	60.3	0.6	45.8	49.2	51.1
Claude Opus 4.7	29.2	42.4	65.5	**16.0**	**53.8**	42.0	47.9
Claude Sonnet 4.6	41.1	50.8	63.8	1.3	35.1	51.0	**51.8**
Open-Source							
DeepSeek V3.2	33.8	46.6	**66.0**	6.7	47.9	46.6	38.7
Kimi K2.5	37.0	40.1	56.7	0.3	36.3	30.4	38.9
Llama 3.3 70B	44.2	50.1	63.6	7.0	32.2	51.5	47.1
Qwen3 32B	37.7	42.8	62.0	7.4	43.9	41.8	38.0
Llama 3.1 8B	38.0	43.5	57.3	4.0	35.7	38.3	32.6
Qwen3 8B	48.5	**57.1**	62.9	10.6	39.2	**56.9**	30.0
SimpleQA							
Proprietary							
GPT-5.5	**2.0**	50.0	50.0	28.6	57.1	57.1	71.4
GPT-5.4	27.7	53.9	86.0	26.4	48.1	50.8	8.9
Claude Opus 4.8	21.4	46.9	89.9	1.7	65.4	25.1	69.3
Claude Opus 4.7	20.4	46.6	90.1	8.7	**69.6**	31.7	75.8
Claude Sonnet 4.6	19.5	41.4	81.1	10.7	62.7	33.7	77.5
Open-Source							
DeepSeek V3.2	46.4	58.0	**91.5**	**35.4**	64.6	55.7	27.2
Kimi K2.5	8.8	70.8	69.2	1.5	47.7	44.6	52.3
Llama 3.3 70B	50.1	**91.9**	84.2	25.6	51.1	**58.9**	**92.5**
Qwen3 32B	33.0	77.8	74.8	14.3	27.1	43.2	24.9
Llama 3.1 8B	70.0	42.4	58.9	20.1	48.8	31.7	84.7
Qwen3 8B	54.4	35.4	11.9	24.4	56.6	34.6	87.3

Table 9: Prevalence of Tool-Memory Conflict (TMC) across datasets, with memory-only (M) and tool-only (T) performance. 0 and 1 denote result is wrong and correct, respectively. Same for memory-only.

LLM	**TMC**	**Resolution Accuracy**						
	Rate (↓)	**Avg. (↑)**	**Resolve (↑)**	**Full Trace (↑)**	**DoT (↑)**	**SR-DCR (↑)**	**Vigilant (↑)**	**XoT (↑)**
Arts & Ent								
Proprietary								
GPT-5.5	**2.3**	20.0	20.0	20.0	**80.0**	60.0	40.0	
GPT-5.4	31.2	46.4	89.3	**31.0**	47.6	42.9	4.8	
Claude Opus 4.8	23.5	47.3	92.7	1.8	60.0	29.1	67.3	
Claude Opus 4.7	20.7	46.8	91.5	6.4	68.1	36.2	76.6	
Claude Sonnet 4.6	24.9	45.2	90.3	11.3	66.1	29.0	80.6	

Open-Source							
DeepSeek V3.2	47.7	50.0	**95.7**	30.9	66.0	54.3	27.7
Kimi K2.5	9.8	81.0	71.4	4.8	28.6	52.4	47.6
Llama 3.3 70B	54.4	**92.4**	84.8	26.7	51.4	**61.0**	**92.4**
Qwen3 32B	36.3	68.6	76.5	15.7	19.6	37.3	22.5
Llama 3.1 8B	78.2	40.6	63.1	18.7	49.2	27.3	85.0
Qwen3 8B	52.7	37.4	10.9	26.5	55.8	37.4	89.8
Math							
Proprietary							
GPT-5.5	**5.9**	40.6	36.0	3.8	71.6	41.0	34.9
GPT-5.4	24.3	36.1	77.0	11.6	**80.3**	36.0	59.7
Claude Opus 4.8	12.7	50.5	83.4	4.9	59.4	58.4	78.7
Claude Opus 4.7	7.9	55.1	77.4	10.7	69.7	71.9	75.8
Claude Sonnet 4.6	10.5	**84.5**	**88.0**	0.8	26.1	**86.3**	**87.8**
Open-Source							
DeepSeek V3.2	24.8	49.2	67.2	40.8	54.1	54.6	63.1
Kimi K2.5	12.0	71.5	79.1	5.9	62.1	69.1	83.8
Llama 3.3 70B	42.4	36.2	77.5	18.7	74.2	38.0	62.0
Qwen3 32B	45.7	38.6	69.2	10.6	60.0	45.7	55.6
Llama 3.1 8B	74.0	24.6	49.1	17.2	23.9	21.1	34.9
Qwen3 8B	67.7	40.8	70.4	**63.1**	59.7	42.7	69.5
Medicine							
Proprietary							
GPT-5.5	**6.2**	34.8	54.3	9.8	37.0	33.7	38.0
GPT-5.4	19.9	47.9	60.6	**18.7**	38.6	46.4	32.0
Claude Opus 4.8	17.5	47.1	58.9	0.5	45.4	48.3	51.4
Claude Opus 4.7	13.2	44.3	62.5	13.5	**55.1**	43.9	50.7
Claude Sonnet 4.6	19.4	50.5	**63.7**	1.3	36.1	50.3	**53.3**
Open-Source							
DeepSeek V3.2	24.8	42.5	63.2	18.0	48.5	45.2	43.1
Kimi K2.5	24.0	45.6	61.8	3.8	35.0	39.4	47.6
Llama 3.3 70B	26.5	47.4	58.0	9.5	34.5	49.2	46.5
Qwen3 32B	27.2	40.5	51.1	11.6	42.9	41.9	41.4
Llama 3.1 8B	34.1	42.5	51.9	6.9	36.0	39.0	35.2
Qwen3 8B	36.8	**51.6**	56.8	18.3	38.8	**51.3**	32.3
Other							
Proprietary							
GPT-5.5	**1.3**	50.0	**100.0**	0.0	0.0	50.0	50.0
GPT-5.4	28.1	62.5	83.9	26.8	44.6	57.1	10.7
Claude Opus 4.8	25.5	46.8	89.4	0.0	51.1	19.1	76.6
Claude Opus 4.7	28.2	50.0	84.0	8.0	68.0	30.0	76.0
Claude Sonnet 4.6	20.9	35.0	72.5	7.5	**75.0**	27.5	87.5
Open-Source							
DeepSeek V3.2	46.6	57.4	94.1	**41.2**	57.4	55.9	29.4
Kimi K2.5	8.7	66.7	73.3	0.0	73.3	33.3	46.7
Llama 3.3 70B	52.0	**87.2**	83.3	25.6	43.6	**62.8**	**91.0**
Qwen3 32B	28.3	85.5	75.8	14.5	21.0	48.4	24.2
Llama 3.1 8B	70.2	40.2	58.3	24.2	50.0	32.6	88.6
Qwen3 8B	55.8	38.8	14.0	24.8	53.7	37.2	84.3
Science & Tech							
Proprietary							

GPT-5.5	**0.2**	0.0	0.0	0.0	0.0	0.0	0.0
GPT-5.4	14.9	53.2	83.5	35.4	63.3	59.5	26.6
Claude Opus 4.8	7.2	52.6	76.3	5.3	**81.6**	28.9	63.2
Claude Opus 4.7	5.5	48.3	79.3	3.4	65.5	44.8	72.4
Claude Sonnet 4.6	5.1	48.1	70.4	14.8	48.1	51.9	55.6
Open-Source							
DeepSeek V3.2	22.5	57.8	**87.5**	**45.3**	60.9	65.6	45.3
Kimi K2.5	4.9	**80.0**	70.0	15.0	45.0	**70.0**	**80.0**
Llama 3.3 70B	21.0	66.7	66.7	28.6	57.1	54.3	70.5
Qwen3 32B	19.5	65.7	59.3	18.5	48.1	46.3	45.4
Llama 3.1 8B	48.4	46.4	51.8	21.1	45.8	43.4	65.7
Qwen3 8B	22.3	30.7	32.5	34.2	52.6	31.6	71.9
Social Sci & Humanities							
Proprietary							
GPT-5.5	**1.1**	63.6	45.5	27.3	45.5	54.5	81.8
GPT-5.4	9.4	51.9	79.8	26.9	51.9	48.1	24.0
Claude Opus 4.8	7.1	46.5	81.7	1.4	73.2	35.2	63.4
Claude Opus 4.7	6.6	50.0	**86.8**	10.3	**73.5**	42.6	76.5
Claude Sonnet 4.6	7.2	45.5	70.1	7.8	46.8	45.5	63.6
Open-Source							
DeepSeek V3.2	18.1	65.0	81.3	**31.7**	64.2	50.4	31.7
Kimi K2.5	5.2	64.4	77.8	2.2	48.9	**57.8**	57.8
Llama 3.3 70B	17.2	**86.0**	72.5	25.8	52.2	54.5	**87.1**
Qwen3 32B	15.8	67.2	68.3	15.6	41.1	43.9	30.6
Llama 3.1 8B	40.1	46.9	54.7	14.9	49.1	38.5	71.7
Qwen3 8B	23.6	35.5	20.5	28.6	59.1	35.1	82.2