

When Facts Change: Temporal Knowledge Conflict Resolution in LLMs

Jonas Wallat
L3S Research Center
Hannover, Germany
jonas.wallat@l3s.de

Wolfgang Nejdl
L3S Research Center
Hannover, Germany
wolfgang.nejdl@l3s.de

Sandipan Sikdar
L3S Research Center
Hannover, Germany
sandipan.sikdar@l3s.de

Abstract

Retrieval-augmented generation (RAG) systems require large language models (LLMs) to reconcile discrepancies between their parametric memory—knowledge encoded during training—and contextual inputs provided at inference. When these sources conflict, models often exhibit unstable reasoning and inconsistent factual behavior. We investigate how LLMs resolve such conflicts when the discrepancy arises from temporal misalignment—facts that have changed since the model’s knowledge cutoff—and whether mutability, the changeability of facts, can serve as a mediating signal in this process. To do so, we provide WIKIRECENTCHANGES, a temporally grounded benchmark with stable and recently updated facts derived from Wikidata. Our results show that while models spontaneously produce temporal reasoning for facts that actually changed—but almost never for stable ones—this differentiation rarely propagates to their final predictions. Explicitly prompting them to consider mutability increases references to temporal change but does not improve factual accuracy, revealing a disconnect between verbalized reasoning and prediction behavior. We further show that the failure point is scale-dependent: smaller models rarely detect the underlying conflict, while larger models detect it but fail to act on their mutability judgments.

1 Introduction

Large language models increasingly serve as the backbone of retrieval-augmented generation (RAG) systems (Lewis et al., 2020), virtual assistants, and autonomous agents. In such settings, models must integrate parametric knowledge—facts encoded during pre-training—with contextual information retrieved or provided at inference time. When these two sources of knowledge disagree, knowledge conflicts arise, commonly referred to as *context–memory conflicts* (Xu et al., 2024).

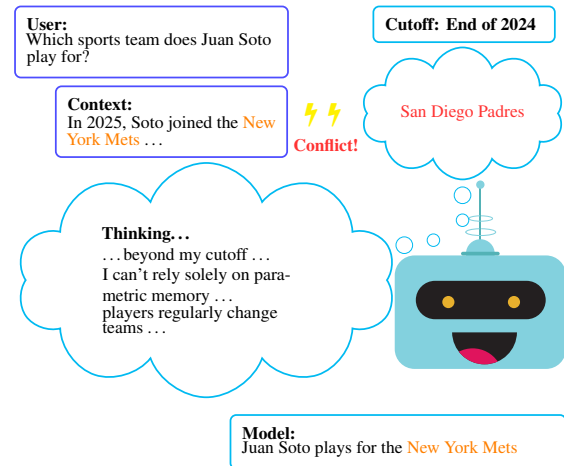


Figure 1: Knowledge conflict resulting from outdated LLM knowledge. Reasoning about whether a fact is likely to change and whether new information is plausible may help to resolve such conflicts.

Knowledge conflicts are inevitable in realistic RAG deployments. They arise, for instance, when retrieved content contains outdated or incorrect information, and can substantially degrade LLM performance (Pham et al., 2024). Retraining LLMs to keep parametric memory up to date is prohibitively expensive, while the growing prevalence of misinformation—often generated by LLMs themselves—further exacerbates the challenge. Effectively navigating such conflicts is therefore critical for ensuring reliable, temporally adaptive, and trustworthy model behavior.

One potential cue for resolving such conflicts is *fact mutability*—whether a fact is expected to change over time. For example, team memberships and political offices are inherently mutable, while birthplaces and dates are typically stable. In principle, recognizing this distinction could help a model decide when to trust new contextual information over its parametric memory (Figure 1).

Despite attention, the mechanisms underlying

conflict resolution in LLMs remain poorly understood. Prior work shows that models exhibit unstable behavior under conflicting evidence, with some over-relying on internal knowledge (Longpre et al., 2021) and others mirroring context (Chen et al., 2022). While recent studies attempt to steer models toward one source or the other through prompting or representation control (Bi et al., 2025; Zhao et al., 2025), forcing rigid preferences between memory and context is suboptimal. In principle, a model would need to reason about when to trust its own knowledge and when to update beliefs based on external information—particularly when that information reflects genuine changes in the world.

Current approaches fall short in three ways. First, most studies rely on synthetic or simplified conflict settings (Longpre et al., 2021; Ortu et al., 2024; Wadhwa et al., 2024) that do not capture the complexity of real-world factual evolution. Second, they largely ignore fact mutability as a potential cue for resolving temporally induced conflicts. Third, they often overlook reasoning models, which offer a window into how conflicts are negotiated.

In this work, we take a step toward filling this gap. Rather than proposing a new conflict-resolution method, we focus on diagnosing whether current LLMs already use mutability as a reasoning signal when such conflicts arise. We introduce WIKIRECENTCHANGES, a re-computable dataset derived from Wikidata that contains both stable facts and facts that recently changed. We use this dataset to study realistic knowledge conflicts caused by temporal misalignment, a natural and non-adversarial source of discrepancy. Our core hypothesis is that LLMs encode an implicit sense of fact mutability that could, in principle, guide decisions between parametric memory and contextual evidence. We test whether such signals are used spontaneously and whether making them explicit affects model behavior by analyzing reasoning traces produced by reasoning models. Concretely, we cross fact type (stable vs. updated) with context type (true vs. counterfactual) and focus on the two conditions where mutability reasoning is diagnostic—stable facts paired with counterfactual context, and updated facts paired with true context (Figure 2).

Our experiments reveal that while models exhibit different reasoning behaviors for stable and updated information, they make little spontaneous use of temporal reasoning when resolving conflicts. Explicitly prompting models to consider mutability

increases references to temporal change, but does not consistently improve factual accuracy, revealing a disconnect between verbalized reasoning and final predictions.

Overall, this work contributes (1) a novel, re-computable benchmark of recently changed facts, (2) a systematic analysis of temporal knowledge conflicts in realistic settings, (3) a diagnostic investigation of mutability as a reasoning factor, and (4) evidence that current LLMs underutilize mutability despite its potential relevance. Our findings highlight the need for methods that better connect temporal awareness to decision-making, paving the way toward more reliable and temporally adaptive models. The code and data are openly available¹.

2 Related Work

Knowledge Conflicts. Knowledge conflicts in large language models (LLMs) arise when two or more sources of knowledge provide inconsistent information. Prior work has distinguished three major types: (1) *context–memory conflicts*, where retrieved or user-provided context contradicts the model’s parametric memory (Longpre et al., 2021; Chen et al., 2022; Jin et al., 2024); (2) *inter-context conflicts*, which occur when multiple retrieved documents or context passages disagree with one another (Lee et al., 2025; Su et al., 2024); and (3) *intra-memory conflicts*, reflecting contradictions within a model’s internal representations or memorized facts (Marjanovic et al., 2024). A recent survey by Xu et al. (2024) provides an extensive overview of these conflict types and their implications for retrieval-augmented and multi-source reasoning systems. Knowledge conflicts are increasingly recognized as a central challenge for reliable model deployment, as they expose the limitations of static pretraining and the difficulty of reconciling evolving world knowledge with frozen model parameters.

Answering from Context vs. Memory. A core question in this line of work concerns how LLMs behave when context and memory disagree. Early analyses found that models often *over-rely on internal knowledge*. Longpre et al. (2021) showed that models tend to reproduce memorized entities even when contradicted by evidence in the prompt. Similarly, Jin et al. (2024) and Tan et al. (2024) observed that retrieval-augmented LLMs frequently

¹<https://github.com/jwallat/knowledge-conflicts-mutability>



Figure 2: Overview of WIKIRECENTCHANGES, crossing fact type (stable vs. updated) with context type (true vs. counterfactual). The two non-shaded columns (stable/counterfactual and updated/true) are the diagnostic conditions for mutability reasoning, where memory and context conflict and a model aware of (im)mutability should resolve the conflict accordingly. Our analysis focuses on the diagnostic cells. The results for the other two conditions are included in Appendix C for completeness.

favor parametric memory, partly because generated or retrieved contexts differ stylistically or semantically from pretraining data. Conversely, other studies report that models can *over-trust the provided context*. Chen et al. (2022) demonstrated that strong retrieval signals can override more accurate internal knowledge, while Sun et al. (2025) showed that contextual exposure can persistently bias subsequent responses, even when the model is later reminded of its original knowledge. Taken together, these findings reveal that LLMs exhibit inconsistent and dataset-dependent behavior under conflict, suggesting that neither memory nor context dominance is fixed but contingent on subtle distributional and representational factors.

Mutability. Temporal misalignment naturally raises the question of how models perceive and reason about *fact mutability*—the degree to which a fact is likely to change over time. Zhang and Choi (2023) explore mitigating temporal misalignment by filtering outdated facts, while Fierro et al. (2024) (MULAN) show that LLMs implicitly encode mutability judgments and can distinguish between stable and evolving facts. Related work on temporal reasoning datasets such as TRAM (Wang and Zhao, 2024) and on dynamic factual knowledge (Marjanovic et al., 2024) investigates how factuality evolves across time or within the model’s own memory. Other studies examine conflicts in temporally grounded reasoning tasks, such as event-based temporal reasoning (Fang et al., 2024), where background knowledge may contradict new contextual evidence. Building on this literature, our work asks whether and how LLMs invoke mutability as

a reasoning cue when resolving context–memory conflicts, particularly in realistic temporal settings where updates reflect genuine world changes rather than adversarial perturbations.

Benchmarks. Hou et al. (2024) introduce WIKI-CONTRADICT, a dataset of 253 human-annotated high-quality real-world Wikipedia-derived contradictions designed to test RAG systems in presence of conflicting contextual information. Su et al. (2024) propose CONFLICTBANK, emphasizing how varying degrees of contradiction affect model confidence and consistency. Emerging datasets such as CHRONOBIAS target temporal bias in RAG systems, while Lee et al. (2025) present MAGIC, a multi-hop benchmark for inter-context conflicts. These resources collectively highlight the pervasiveness of knowledge conflicts and provide test beds for both mechanistic and behavioral analyses. Closely related is also EVOWIKI (Tang et al., 2025), a Wikipedia-derived dataset, designed to study how RAG systems behave when exposed to changed and novel information. The dataset is split into three splits: *stable* information, which did not change between 2021 and January 2024, *evolved* information, which did change in the timeframe, and *uncharted* information, which was only created between January 2024 and May 2024. While the dataset would be a good starting point to study temporal knowledge conflicts, two factors make it unsuitable for our study. First, the time scope of 2021 to mid-2024 is no longer aligned with current LLM knowledge cutoffs, which are often trained on information until mid- to late 2024. Second, the way of characterizing stable information is by

checking whether the information changed during their time horizon. This results in many mutable facts being placed in the stable split (such as an athlete plays for a certain team), as long as the specific team did not change between 2021 and 2024. To study whether LLMs use mutability to resolve conflicts, we need the stable split to contain actually stable information and not information that just didn't change by chance.

3 Setup – Measuring Knowledge Conflicts

To evaluate the behavior of LLMs and how they resolve temporal knowledge conflicts, we first construct a novel dataset WIKIRECENTCHANGES (Section 3.1) and evaluate our selection of open-source models (Section 3.2) using parametric and non-parametric metrics (Section 3.3).

Our design crosses two fact types (stable vs. updated) with two context types (true vs. counterfactual), yielding four conditions (Figure 2). Two of these are diagnostic for mutability reasoning: stable facts with counterfactual context, where a model that recognizes immutability should reject the context and trust memory, and updated facts with true context, where a model that recognizes mutability should accept the new context over outdated memory. The other two conditions — stable/true and updated/counterfactual — are confirmatory: parametric memory and context agree, so no conflict resolution is required and mutability provides no diagnostic signal. We therefore focus our analysis on the two diagnostic cells, reporting full results across all four in Appendix C.

3.1 WIKIRECENTCHANGES

We aim to construct a dataset that induces controlled yet realistic contradictions between parametric and contextual knowledge. A key requirement is that contextual updates involve updated events—facts that were updated after the model's training data cutoff—so that the model's parametric memory is predictably outdated. Unlike synthetic or template-based counterfactual setups used in prior work (Longpre et al., 2021; Ortu et al., 2024; Wadhwa et al., 2024) (e.g., “The earth is flat. The earth is [MASK]”), our approach aims to preserve semantic and temporal realism, ensuring that models must engage genuine reasoning rather than surface-level pattern matching. To achieve this, we construct WIKIRECENTCHANGES, a dynamic corpus designed to capture evolving factual knowl-

edge over time. The dataset is derived from Wikidata (Wikidata contributors, 2025) and contains stable and updated facts. By generating high quality true and counterfactual paragraphs, we maintain linguistic naturalness. Moreover, the dataset can be *re-generated* periodically to align with new events, enabling longitudinal analysis of model behavior as the world evolves. This setup provides a robust foundation for probing how mutability and internal mechanisms mediate conflict resolution in LLMs.

Dataset Construction. We build our dataset by parsing a recent Wikidata dump² for a seed set of 18 stable (e.g., native language, place of birth) and 27 mutable relations (e.g., member of sports team, position held). We further filter the resulting subject-relation-object triples by requiring English human-readable labels and an English Wikipedia page to help us later generate short contexts. For the mutable triples, we ensure they contain a temporal qualifier, such as a corresponding start time (P580) or a point in time (P585) that is within 2025. That way, we ensure that the resulting triples only contain changes that actually happened in 2025. While not all relations contain such temporal qualifiers, this approach is preferable over just identifying differences between two Wikidata dumps from different times, as these contain updated facts and a high number of edits on old facts—making it difficult to identify what exactly changed in 2025. Nevertheless, the filtering resulted in a suitable set of 2380 mutable facts, and we sample a similarly sized set of stable facts. We construct natural language questions from subject and relation pairs³.

Counterfactual Answers. To study knowledge conflicts, we require counterfactual contexts that suggest conflicting answers. We first generate such counterfactual answers using a Qwen-3-14B model. Specifically, we feed a question and actual answer to the model and prompt it to generate a similar, yet distinct, answer. The counterfactual answer is supposed to be a similar and plausible answer. That is, if the entity is a baseball player at team A, the model is supposed to provide another team B in the same league.

True and Counterfactual Contexts. We similarly generate true and counterfactual context using Qwen-3-14B. For the counterfactual contexts, we

²from December 3rd, 2025

³Additional information on dataset statistics, relation counts, and question formulations is provided in Appendix B.1

feed questions, counterfactual answers and the date when the fact changed (for mutable facts) to the model. The resulting paragraphs consist of 3–4 sentences and are natural. For the true contexts, one could have chosen to directly select the extracted Wikipedia pages, but that would have diluted the study since the short generated paragraphs and the full Wikipedia pages differ in style, length, etc. Therefore, we construct true paragraphs by also feeding the Wikipedia page to the model and extract/paraphrase a similar context of 3–4 sentences.

The prompts used to generate these contexts and counterfactual answers, as well as quality control and human agreement studies are available in Appendix B.1.2 and B.4.3.

3.2 Models

We use a selection of openly available reasoning and non-reasoning models. For each experiment, we generate up to 4096 new tokens and report the specific prompt used (in Appendix B.4). For the non-reasoning models, we utilize (i) Gemma 3-4B-it (Kamath et al., 2025) (ii) Llama 3.3 70B-it (Dubey et al., 2024). As reasoning models, we select: (i) Phi-4-mini-reasoning (Abdin et al., 2025) (ii) Qwen-3-8B (Yang et al., 2025) (iii) OLMO-3-7B-think⁴ (iv) GPT-OSS-20B (Agarwal et al., 2025) (v) GPT-OSS-120B (Agarwal et al., 2025) (vi) DeepSeek-R1-Distill-Llama-8B (DeepSeek-AI et al., 2025). All models have a knowledge cutoff date before 2025 except for Phi-4-mini (March 2025). Additional details on the models and cutoff dates are in Table 11 in Appendix B.2.

3.3 Metrics

To evaluate the generated answers, we utilize model-specific metrics (*OpenEval* (Kamalloo et al., 2023)) and model-agnostic metrics (i.e., *Match* (Adlakha et al., 2024; Liu et al., 2024; Mallen et al., 2023)). *OpenEval* evaluates the correctness of an answer by querying whether a candidate is a suitable answer given the question and the reference answer⁵. Additionally, we evaluate answers with *Match* which is, given that LLM answers tend to be verbose, a relaxed version which does not require being exact. The LLM-based evaluations were validated against human annotations (Appendix B.3) and both metrics show high human agreement (>90%).

⁴<https://huggingface.co/allenai/olmo-3-7B-think>

⁵For which we utilize Flan-T5-XXL

4 Results

4.1 Knowledge Conflict Behavior

First and foremost, we characterize the knowledge conflict behavior of our LLMs. To do so, we study several settings. As a baseline, we compute the QA performance on WIKIRECENTCHANGES without context in a closed-book setting. After that, we explore LLM performance given true or counterfactual contexts.

Model	Split	No Context	
		OE	Match
Gemma-3-4B	stable	37.0	31.0
	updated	6.4	2.4
Llama-3.3-70B	stable	48.9	38.0
	updated	9.7	4.5
Phi-4-mini	stable	35.0	32.4
	updated	4.4	1.9
DS-Llama-8B	stable	33.9	25.5
	updated	5.5	1.7
OLMO-3-7B	stable	38.2	32.4
	updated	5.4	2.1
Qwen-3-8B	stable	42.8	35.8
	updated	7.8	3.5
GPT-OSS-120B	stable	57.0	49.5
	updated	11.5	6.8

Table 1: Closed-book QA performance (no context).

QA performance without contexts (closed book).

As an initial baseline, we evaluate the models in a closed-book setting (Table 1). Expectedly, we observe all models to perform better at the stable split compared to the updated split. We hypothesize the non-zero performance on the updated split to be due to announcements of future changes that were already present in the training data & models guessing most likely answers.

Model	Split	Ctx	GT	CF	CB
Phi-4-mini	stable	C	13.5	76.4	18.6
	updated	T	68.0	5.1	6.3
DS-Llama-8B	stable	C	5.0	94.0	17.3
	updated	T	83.2	4.9	8.5
OLMO-3-7B	stable	C	8.1	91.3	16.9
	updated	T	78.5	5.1	5.3
Qwen-3-8B	stable	C	11.4	87.2	20.1
	updated	T	80.5	5.8	9.6
GPT-OSS-20B	stable	C	27.6	58.6	36.1
	updated	T	51.5	7.2	10.1
GPT-OSS-120B	stable	C	33.6	57.8	42.5
	updated	T	63.7	7.4	19.1

Table 2: Conflict behavior across reasoning models, dataset splits, and context types in the OpenEval (OE) metric. We measure whether the given answer matches the ground-truth answer (GT), the counterfactual (CF), or the model’s closed-book answer (CB) under both True (T) and Counterfactual (C) context conditions.

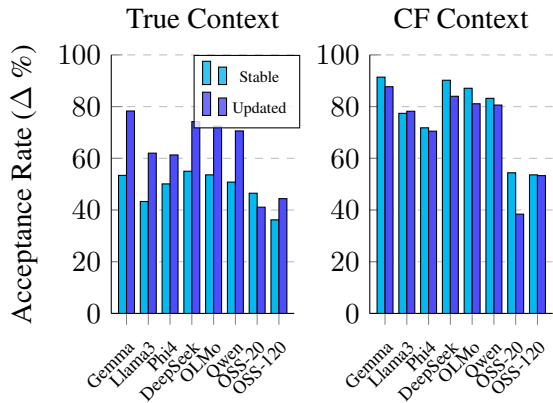


Figure 3: Context acceptance rates for both true and counterfactual contexts.

LLM behavior in the presence of knowledge conflicts.

Next, we turn toward studying the behavior of our models in the presence of knowledge conflicts. With temporally induced knowledge conflicts, we specifically focus on the two cases where reasoning about mutability may help resolve conflicts (as shown in Figure 2): Updated facts when exposed to true (but new) contexts and stable facts when exposed to counterfactual contexts. The results are presented in Table 2. We observe relatively high rates of ground-truth answers being given (GT) when exposed to true context and counterfactual answers being given (CF) when exposed to counterfactual contexts. This seems to be the case for both the stable and the updated split. While we find models to remain with the closed-book answer (CB) more often for stable facts, the general rate is still quite low (ranging from 17-43%). Yet, we observe bigger models to stay with their parametric beliefs more often.

Context Reliance. A central debate in the literature is whether models resolve knowledge conflicts by prioritizing contextual evidence or their internal parametric memory. When evaluating stable versus updated factual knowledge, we observe a general tendency for models to favor contextual information. Figure 3 illustrates the context acceptance rates for each model. We define this rate as the marginal gain in accuracy: For true contexts, we measure the gain in GT answers over the closed-book rate. For counterfactual contexts, we measure the gain in CF answers over the rate at which the model produces the counterfactual when given the true context, isolating the effect of the adversarial context itself. Interestingly, for true contexts, the acceptance rate is higher for updated facts, while for counterfactual contexts it is pretty stable be-

tween splits. We also explored whether the model knowing the answer to the original question has an effect on the context acceptance rate, and found it to have little effect (details in Appendix C.1.1).

4.2 Usage of Mutability in Reasoning

Next, we investigate whether the LLMs’ understanding of mutability plays a role in resolving these knowledge conflicts.

Measuring Mutability Reasoning. To measure whether reasoning about mutability occurs, we consider the reasoning models and study their reasoning traces. The original prompt contained a section that requested the model to “critically evaluate” the context and decide whether to answer from memory or the provided context. Thus, we expect rationales for this decision to be present in the reasoning traces. Given the scale of thousands of answers and up to 4096 tokens per answer, we resort to an LLM-as-a-judge (Li et al., 2024) approach. Specifically, we few-shot prompted Qwen-3-14B to decide if the reasoning trace contains reasoning about mutability to resolve the conflicts. We perform additional analysis on the judge and find it to have an agreement rate of 94% with human judgment⁶. We consider the trace to show that mutability reasoning was used if 1) there was a conflict acknowledged that can be resolved, 2) the model finds the fact to be mutable or immutable, and 3) that information is being used to resolve the conflict.

In Table 3 (Standard), we observe a stark difference between the stable and updated facts: While there is next to no mutability reasoning being used for static facts (0-1.3%), we see much higher rates for the updated facts (3-36%). This is a first indication that knowledge conflicts are being mediated differently for stable facts. When it comes to context types, we observe very similar results between true and counterfactual information.

Breaking Down Temporal Reasoning. To understand *where* mutability reasoning breaks down, we decompose it into three sequential subskills a model must execute to resolve a temporal conflict: (i) detecting that its parametric memory disagrees with the context, (ii) recognizing that the fact in question is mutable (or immutable), and (iii) actually using that recognition to accept or reject the context. We analyze 1000 traces for Qwen-3-8B and GPT-OSS-120B from the stable split with coun-

⁶Prompt in Appendix B.4, human eval in Appendix B.3.2

Model	Split	Percentage of Traces containing Temporal Reasoning					
		Standard		Dates		Prompted	
		Counter	True	Counter	True	Counter	True
Phi-4-mini	stable	1.3	1.2	5.4	5.6	9.3	9.3
	updated	27.7	28.0	18.3	18.2	32.9	33.2
DS-Llama-8B	stable	0.0	0.0	0.2	0.2	1.0	1.0
	updated	2.9	2.6	7.9	8.1	23.4	23.8
OLMO-3-7B	stable	0.3	0.2	0.4	0.4	2.4	2.6
	updated	35.5	34.7	14.9	15.0	35.0	35.6
Qwen-3-8B	stable	0.1	0.1	0.4	0.4	1.5	1.7
	updated	13.4	13.0	20.5	20.3	27.2	27.6
GPT-OSS-20B	stable	0.0	0.0	0.3	0.3	5.0	5.1
	updated	9.9	9.9	16.2	16.0	28.2	27.8
GPT-OSS-120B	stable	0.2	0.2	0.2	0.1	2.9	2.9
	updated	20.1	20.0	24.5	24.6	59.8	60.6

Table 3: Proportion of traces explicitly containing mutability reasoning across models and dataset splits. We explore reasoning traces of answers when the available context is either the true context containing the ground-truth answer or the counterfactual context. We explore the effect of several prompts: the standard setting, prepending current and cutoff dates ("Dates"), and specifically prompting the model to reason about mutability ("Prompted").

Type	Split / Context	% of traces	
		Qwen-3-8B	GPT-OSS-120B
Conflict detected	stable / counter	16.8	42.3
	updated / true	10.2	19.6
Mutable reasoning	stable / counter	16.9	79.0
	updated / true	29.4	90.9
Mutable reasoning used	stable / counter	3.1	3.5
	updated / true	25.3	62.9

Table 4: Analysis of conflict detection, recognition of mutability, and usage of mutable reasoning.

terfactual contexts and 1000 traces from the updated split with true context. For the stable traces, we select 500 where the model correctly answered the closed-book question and 500 where it provided the wrong answer. We use Qwen-3-14B to judge each step (prompts in Appendix B.4). The results are in Table 4. Conflict detection is the primary bottleneck for Qwen-3-8B: despite knowing 50% of answers closed-book, it recognizes conflicts in only 17% of cases. GPT-OSS-120B detects conflicts and mutability more reliably, but both models identify mutability more often than they actually act on it when producing their final answer.

Can we get the model to use temporal reasoning? The low usage of mutability reasoning poses the question of whether we can improve the models’ capability to resolve knowledge conflicts by explicitly making them reason about mutability.

Consequently, we adjust the prompt in two different ways: first, we temporally ground the LLMs with the current time and their explicit knowledge cutoff. To do so, we prepend “Today is {date}. Your knowledge cutoff is {cutoff}”, which we call “Dates”. This might allow the model to more effectively reason about the likelihood of changed facts.

Second, we additionally ask the model to “put special emphasis on whether a question’s answer will change over time and use that information to select/reject the provided context”. The results are in Table 3 (Dates, Prompted)⁷. While grounding the models in cutoff and current times seems to improve the amount of temporal reasoning for all but Phi-4-mini and OLMO-3-7B, we see even stronger improvements for direct prompting to use temporal reasoning. This is a substantial increase that poses the direct question: *does reasoning about mutability improve knowledge conflict resolution?*

Is temporal reasoning a function of time? To understand if temporal reasoning is dependent on the difference between cutoff and current time, we again sample 1000 mutable questions with true context and have Qwen-3-8B predict the answers. However, this time, we vary the current time and the time references mentioned in the context by a certain number of years (Figure 4). We repeat the experiment three times and report mean and standard deviation. While we see a minor decrease in the amount of mutability reasoning when shift-

⁷Example reasoning trace in Appendix C.2

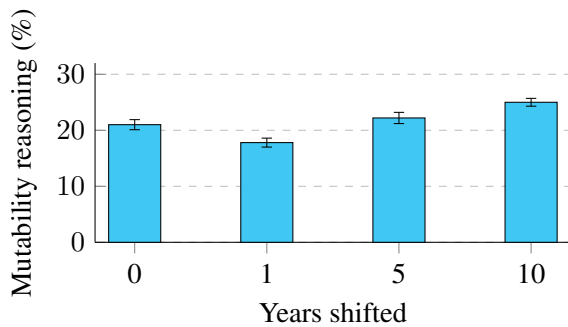


Figure 4: Amount of mutability reasoning used when shifting current year and year dates in the context.

ing the dates one year into the future, the amount increases when shifting by five or ten years. This is an indication that temporal reasoning might arise naturally if models expect their memory to be outdated. However, the perception of when information is outdated might not be correctly captured at the moment.

Is it the perception of context dates? A natural next question to ask is whether the temporal reasoning depends on the models’ perception of the provided documents date. Knowing, instead of assuming, the date at which the context was published might allow for more precise usage of mutability to decide whether to accept or reject it. To study if this has an impact, we again sample datasets with 1k examples and specify specific document dates in the prompt (e.g., “The retrieved context is from 2023”). The results are presented in Table 5.

For stable information, the models seem not to consider temporal reasoning independent of document dates. For updated facts, we observe higher, but relatively similar scores, for the different document dates. Specifying the document date to 2023 seems to have a positive effect on the amount of temporal reasoning. One reason for this might be that the models contain newer information (up to the end of 2024) than the provided context and use that to reject contexts as outdated. Interestingly, explicitly stating that context is from 2025—and therefore beyond the knowledge cutoff—does not lead to more mutability reasoning compared to not stating the document date.

Does reasoning about mutability help? We evaluate the QA performance after applying the temporal grounding and temporal reasoning prompts. The results are presented in Table 6. As the provided context in this setting contains conflicting information, we consider temporal reason-

ing to *help*, if more of the counterfactual contexts are rejected (for stable with counterfactual context) and ground-truth answers being given (for updated and true contexts). Despite the increase in reasoning traces mentioning mutability (Table 3), this did not translate into measurable performance gains for stable facts. The models remained equally likely to accept counterfactual information even when explicitly prompted to consider mutability. For updated facts, where we also observed a bigger impact on temporal reasoning, only slight increases in correct answers being given (up to 20% for GPT-OSS-20B) were noted.

5 Discussion and Conclusion

This work examined how large language models resolve temporal conflicts between their parametric memory and retrieved context, focusing on *mutability*—the changeability of facts over time.

Across all conditions, we found that models accept and reject context with similar rates—but only seem to consistently apply mutability reasoning for changing facts. Even when explicitly prompted to consider mutability, they produce far more reasoning traces referring to temporal change yet show only small and inconsistent improvements in factual accuracy or conflict resolution.

Where does mutability reasoning break down?

Our decomposition (Table 4) reveals that the failure point depends on model scale. For smaller reasoning models like Qwen-3-8B, the dominant bottleneck is conflict detection: even when the model knows half the answers in the stable/counterfactual condition, it flags a conflict in only 17% of traces. For larger models like GPT-OSS-120B, conflict detection and mutability recognition are largely solved, yet mutability still fails to propagate to the final answer. This pattern suggests that scaling improves the earlier components of temporal reasoning but leaves the link between verbalized reasoning and prediction behavior largely untouched—consistent with broader evidence that chain-of-thought traces are often unfaithful to the computation that produces the answer.

Reliability under realistic context conditions.

These findings highlight a core limitation of current retrieval-augmented systems. In realistic deployments, retrieved content is often *noisy*, *outdated*, or *misleading*. Robust performance therefore requires models to assess the reliability of contextual

Type	Split / Context	% Mutability Reasoning	
		Qwen-3-8B	GPT-OSS-120B
No document date	stable / counter	0.4%	0.3%
	updated / true	20.5%	28.1%
Document date = 2023	stable / counter	0.9%	0.3%
	updated / true	27.0%	35.6%
Document date = 2025	stable / counter	1.5%	0.2%
	updated / true	11.1%	20.2%

Table 5: Effect of explicitly stating the document date on reasoning about mutability.

Model	Split	Ctx	Standard			Dates			Prompted		
			GT	CF	CB	GT	CF	CB	GT	CF	CB
Phi-4-mini	stable	C	13.5	76.4	18.6	13.8	79.5	18.9	12.6	81.4	17.7
	updated	T	68.0	5.1	6.3	72.7	5.2	6.9	72.9	4.5	6.4
DS-Llama-8B	stable	C	5.0	94.0	17.3	5.0	93.3	17.3	5.0	91.9	17.3
	updated	T	83.2	4.9	8.5	82.2	5.1	8.3	79.5	5.0	8.6
OLMO-3-7B	stable	C	8.1	91.3	16.9	8.7	90.7	17.2	8.4	90.5	16.8
	updated	T	78.5	5.1	5.3	81.8	5.0	5.8	78.5	5.1	5.3
Qwen-3-8B	stable	C	11.4	87.2	20.1	12.6	85.0	20.7	11.4	86.2	19.9
	updated	T	80.5	5.8	9.6	86.4	4.4	7.6	87.1	4.4	7.3
GPT-OSS-20B	stable	C	27.6	58.6	36.1	26.1	59.1	34.7	25.0	62.7	33.3
	updated	T	51.5	7.2	10.1	63.8	6.2	8.9	71.6	5.2	7.7
GPT-OSS-120B	stable	C	33.6	57.8	42.5	34.4	57.0	43.7	33.9	57.5	42.7
	updated	T	63.7	7.4	19.1	70.6	5.7	15.2	74.5	5.3	12.8

Table 6: QA behavior under conflict across prompting strategies. Scores represent the OpenEval (OE) metric. We measure whether the given answer matches the ground-truth answer (GT), the counterfactual (CF), or the model’s closed-book answer (CB) under both True and Counterfactual context conditions.

information and decide when to trust it. Our results show that current LLMs integrate retrieved evidence indiscriminately, regardless of whether a fact is mutable or immutable. Without mechanisms to evaluate the temporal and factual validity of context, improvements in retrieval quality alone cannot ensure reliable reasoning.

From understanding to reasoning. Prior work has shown that LLMs encode signals of factual mutability in their internal representations, suggesting an implicit awareness of which facts are prone to change (Fierro et al., 2024). Our findings extend this perspective: while models *represent* mutability, their *use* when generating answers is limited. This disconnect reflects two likely limitations. First, reasoning traces appear to be unfaithful—temporal reasoning is verbalized but not causally linked to the final prediction. This has also recently been observed in other studies and remains an open problem (Yee et al., 2024; Arcuschin et al., 2025). Second, training data rarely connects factual volatility to answer selection, leaving mutability as an ungrounded latent property rather than a functional reasoning cue.

Limitations of steering approaches. Recent studies have attempted to steer models toward

greater context reliance through prompting, decoding, or representation control (e.g., Bi et al., 2025; Zhao et al., 2025). While effective in controlled benchmarks, such approaches assume that retrieved content is more reliable than the model’s memory. In realistic settings, this assumption often fails: retrieved text may itself be outdated or incorrect, and enforcing context dominance can amplify rather than mitigate errors.

Conclusion. LLMs encode an implicit understanding of factual mutability and verbalize it in their reasoning traces, yet this reasoning rarely shapes their final predictions. The failure point is scale-dependent: smaller models struggle to detect conflicts at all, while larger models identify both the conflict and the mutability of the fact but fail to act on that recognition. This limits their reliability in dynamic, noisy, or misinformation-prone environments—precisely the contexts where RAG systems are most valuable. Progress will require moving beyond static steering to utilize the context toward models that can introspect on their own knowledge, evaluate retrieved evidence, and reason about when and why a fact might have changed, paving the way for temporally adaptive and trustworthy language models.

Limitations

What other properties matter when resolving knowledge conflicts? This paper is a focussed study on understanding if LLMs can use mutability to resolve knowledge conflicts. Therefore, we focus on mutability. Yet, other factors can and will likely have an effect on whether LLMs “trust” the contexts. Such factors could be: Writing style, perceived factuality, knowledge about the contexts’ sources, relevance to the question, and overlap with parametric knowledge. Exploring such other factors would be worthwhile, yet out of scope of this work.

Can we isolate the effect of mutability? A potential reason for the results is that there are other convoluting effects that might affect the model’s outputs. We tried to control other factors, by utilizing data containing stable, and changing information and created high quality realistic conflicts. Nevertheless, factors such as plausibility, etc. will necessarily be a factor.

We only consider context-parameter conflicts. This study focussed on context-parameter conflicts, with a single context provided for each question. The effect of a diverse range of documents and potentially inter-context conflicts will be more realistic, but was out of scope of this study. That said, the fact that a single counterfactual context was enough to fool the models in a large majority of cases already suggests strong susceptibility to realistic misinformation.

Ethics Statement

Similar to other works, we observe LLMs to quite easily fall for misinformation if presented in high-quality paragraphs. While that is certainly a problem, this work does not offer novel attack vectors for misinformation and the likes. Thus, we see no major ethical problems with our work.

Acknowledgments

This work was supported by the Lower Saxony Ministry of Science and Culture (MWK), in the zukunft.niedersachsen program of the Volkswagen Foundation (HybrInt).

References

Marah I Abidin, Sahaj Agarwal, Ahmed Awadallah, Vidhisha Balachandran, Harkirat S. Behl, Lingjiao Chen,

Gustavo de Rosa, Suriya Gunasekar, Mojan Javaheripi, Neel Joshi, Piero Kauffmann, Yash Lara, Caio César Teodoro Mendes, Arindam Mitra, Besmira Nushi, Dimitris Papailiopoulos, Olli Saarikivi, Shital Shah, Vaishnavi Shrivastava, Vibhav Vineet, Yue Wu, Safoora Yousefi, and Guoqing Zheng. 2025. [Phi-4-reasoning technical report](#). *CoRR*, abs/2504.21318.

Vaibhav Adlakha, Parishad BehnamGhader, Xing Han Lu, Nicholas Meade, and Siva Reddy. 2024. [Evaluating correctness and faithfulness of instruction-following models for question answering](#). *Trans. Assoc. Comput. Linguistics*, 12:681–699.

Sandhini Agarwal, Lama Ahmad, Jason Ai, Sam Altman, Andy Applebaum, Edwin Arbus, Rahul K. Arora, Yu Bai, Bowen Baker, Haiming Bao, Boaz Barak, Ally Bennett, Tyler Bertao, Nivedita Brett, Eugene Brevdo, Greg Brockman, Sébastien Bubeck, Che Chang, Kai Chen, Mark Chen, Enoch Cheng, Aidan Clark, Dan Cook, Marat Dukhan, Casey Dvorak, Kevin Fives, Vlad Fomenko, Timur Garipov, Kristian Georgiev, Mia Glaese, Tarun Gogineni, Adam P. Goucher, Lukas Gross, Katia Gil Guzman, John Hallman, Jackie Hehir, Johannes Heidecke, Alec Helyar, Haitang Hu, Romain Huet, Jacob Huh, Saachi Jain, Zach Johnson, Chris Koch, Irina Kofman, Dominik Kundel, Jason Kwon, Volodymyr Kyrlyov, Elaine Ya Le, Guillaume Leclerc, James Park Lennon, Scott Lessans, Mario Lezcano Casado, Yuanzhi Li, Zhuohan Li, Ji Lin, Jordan Liss, Lily Liu, Jiancheng Liu, Kevin Lu, Chris Lu, Zoran Martinovic, Lindsay McCallum, Josh McGrath, Scott McKinney, Aidan McLaughlin, Song Mei, Steve Mostovoy, Tong Mu, Gideon Myles, Alexander Neitz, Alex Nichol, Jakub Pachocki, Alex Paino, Dana Palmie, Ashley Pantuliano, Giambattista Parascandolo, Jongsoo Park, Leher Pathak, Carolina Paz, Ludovic Peran, Dmitry Pimenov, Michelle Pokrass, Elizabeth Proehl, Huida Qiu, Gaby Raila, Filippo Raso, Hongyu Ren, Kimmy Richardson, David Robinson, Bob Rotsted, Hadi Salman, Suvansh Sanjeev, Max Schwarzer, D. Sculley, Harshit Sikchi, Kendal Simon, Karan Singhal, Yang Song, Dane Stuckey, Zhiqing Sun, Philippe Tillet, Sam Toizer, Foivos Tsimpourlas, Nikhil Vyas, Eric Wallace, Xin Wang, Miles Wang, Olivia Watkins, Kevin Weil, Amy Wendling, Kevin Whinnery, Cedric Whitney, Hannah Wong, Lin Yang, Yu Yang, Michihiro Yasunaga, Kristen Ying, Wojciech Zaremba, Wenting Zhan, Cyril Zhang, Brian Zhang, Eddie Zhang, and Shengjia Zhao. 2025. [gpt-oss-120b & gpt-oss-20b model card](#). *CoRR*, abs/2508.10925.

Iván Arcuschin, Jett Janiak, Robert Krzyzanowski, Senthoooran Rajamanoharan, Neel Nanda, and Arthur Conmy. 2025. [Chain-of-thought reasoning in the wild is not always faithful](#). *CoRR*, abs/2503.08679.

Baolong Bi, Shenghua Liu, Yiwei Wang, Yilong Xu, Junfeng Fang, Lingrui Mei, and Xueqi Cheng. 2025. [Parameters vs. context: Fine-grained control of knowledge reliance in language models](#). *CoRR*, abs/2503.15888.

- Jannis Bulian, Christian Buck, Wojciech Gajewski, Benjamin Börschinger, and Tal Schuster. 2022. [Tomayto, tomahto. beyond token-level answer equivalence for question answering evaluation](#). In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing, EMNLP 2022, Abu Dhabi, United Arab Emirates, December 7-11, 2022*, pages 291–305. Association for Computational Linguistics.
- Hung-Ting Chen, Michael J. Q. Zhang, and Eunsol Choi. 2022. [Rich knowledge sources bring complex knowledge conflicts: Recalibrating models to reflect conflicting evidence](#). In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing, EMNLP 2022, Abu Dhabi, United Arab Emirates, December 7-11, 2022*, pages 2292–2307. Association for Computational Linguistics.
- DeepSeek-AI, Daya Guo, Dejian Yang, Haowei Zhang, Junxiao Song, Ruoyu Zhang, Runxin Xu, Qihao Zhu, Shiron Ma, Peiyi Wang, Xiao Bi, Xiaokang Zhang, Xingkai Yu, Yu Wu, Z. F. Wu, Zhibin Gou, Zhihong Shao, Zhuoshu Li, Ziyi Gao, Aixin Liu, Bing Xue, Bingxuan Wang, Bochao Wu, Bei Feng, Chengda Lu, Chenggang Zhao, Chengqi Deng, Chenyu Zhang, Chong Ruan, Damai Dai, Deli Chen, Dongjie Ji, Erhang Li, Fangyun Lin, Fucong Dai, Fuli Luo, Guangbo Hao, Guanting Chen, Guowei Li, H. Zhang, Han Bao, Hanwei Xu, Haocheng Wang, Honghui Ding, Huajian Xin, Huazuo Gao, Hui Qu, Hui Li, Jianzhong Guo, Jiashi Li, Jiawei Wang, Jingchang Chen, Jingyang Yuan, Junjie Qiu, Junlong Li, J. L. Cai, Jiaqi Ni, Jian Liang, Jin Chen, Kai Dong, Kai Hu, Kaige Gao, Kang Guan, Kexin Huang, Kuai Yu, Lean Wang, Lecong Zhang, Liang Zhao, Litong Wang, Liyue Zhang, Lei Xu, Leyi Xia, Mingchuan Zhang, Minghua Zhang, Minghui Tang, Meng Li, Miaojun Wang, Mingming Li, Ning Tian, Panpan Huang, Peng Zhang, Qiancheng Wang, Qinyu Chen, Qiushi Du, Ruiqi Ge, Ruisong Zhang, Ruizhe Pan, Runji Wang, R. J. Chen, R. L. Jin, Ruyi Chen, Shanghao Lu, Shangyan Zhou, Shanhuang Chen, Shengfeng Ye, Shiyu Wang, Shuiping Yu, Shunfeng Zhou, Shuting Pan, and S. S. Li. 2025. [Deepseek-r1: Incentivizing reasoning capability in llms via reinforcement learning](#). *CoRR*, abs/2501.12948.
- Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Amy Yang, Angela Fan, Anirudh Goyal, Anthony Hartshorn, Aobo Yang, Archi Mitra, Archie Sravankumar, Artem Korenev, Arthur Hinsvark, Arun Rao, Aston Zhang, Aurélien Rodriguez, Austen Gregerson, Ava Spataru, Baptiste Rozière, Bethany Biron, Binh Tang, Bobbie Chern, Charlotte Caucheteux, Chaya Nayak, Chloe Bi, Chris Marra, Chris McConnell, Christian Keller, Christophe Touret, Chunyang Wu, Corinne Wong, Cristian Canton Ferrer, Cyrus Nikolaidis, Damien Al-lonsius, Daniel Song, Danielle Pintz, Danny Livshits, David Esiobu, Dhruv Choudhary, Dhruv Mahajan, Diego Garcia-Olano, Diego Perino, Dieuwke Hupkes, Egor Lakomkin, Ehab AlBadawy, Elina Lobanova, Emily Dinan, Eric Michael Smith, Filip Radenovic, Frank Zhang, Gabriel Synnaeve, Gabrielle Lee, Georgia Lewis Anderson, Graeme Nail, Grégoire Mialon, Guan Pang, Guillem Cucurell, Hailey Nguyen, Hannah Korevaar, Hu Xu, Hugo Touvron, Iliyan Zarov, Imanol Arrieta Ibarra, Isabel M. Kloumann, Ishan Misra, Ivan Evtimov, Jade Copet, Jaewon Lee, Jan Geffert, Jana Vranes, Jason Park, Jay Mahadeokar, Jeet Shah, Jelmer van der Linde, Jennifer Billock, Jenny Hong, Jenya Lee, Jeremy Fu, Jianfeng Chi, Jianyu Huang, Jiawen Liu, Jie Wang, Jiecao Yu, Joanna Bitton, Joe Spisak, Jongsoo Park, Joseph Rocca, Joshua Johnstun, Joshua Saxe, Junteng Jia, Kalyan Vasuden Alwala, Kartikeya Upasani, Kate Plawiak, Ke Li, Kenneth Heafield, Kevin Stone, and et al. 2024. [The llama 3 herd of models](#). *CoRR*, abs/2407.21783.
- Tianqing Fang, Zhaowei Wang, Wenxuan Zhou, Hongming Zhang, Yangqiu Song, and Muhao Chen. 2024. [Getting sick after seeing a doctor? diagnosing and mitigating knowledge conflicts in event temporal reasoning](#). In *Findings of the Association for Computational Linguistics: NAACL 2024, Mexico City, Mexico, June 16-21, 2024*, pages 3846–3868. Association for Computational Linguistics.
- Mehrdad Farahani and Richard Johansson. 2024. [Deciphering the interplay of parametric and non-parametric memory in retrieval-augmented language models](#). In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing, EMNLP 2024, Miami, FL, USA, November 12-16, 2024*, pages 16966–16977. Association for Computational Linguistics.
- Constanza Fierro, Nicolas Garneau, Emanuele Bugliarello, Yova Kementchedjhieva, and Anders Søgaard. 2024. [Mulan: A study of fact mutability in language models](#). In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies: Short Papers, NAACL 2024, Mexico City, Mexico, June 16-21, 2024*, pages 762–771. Association for Computational Linguistics.
- Yufang Hou, Alessandra Pascale, Javier Carnerero-Cano, Tigran T. Tchrakian, Radu Marinescu, Elizabeth Daly, Inkit Padhi, and Prasanna Sattigeri. 2024. [Wikicontradict: A benchmark for evaluating llms on real-world knowledge conflicts from wikipedia](#). In *Advances in Neural Information Processing Systems 38: Annual Conference on Neural Information Processing Systems 2024, NeurIPS 2024, Vancouver, BC, Canada, December 10 - 15, 2024*.
- Zhuoran Jin, Pengfei Cao, Yubo Chen, Kang Liu, Xiaojian Jiang, Jiexin Xu, Qiuxia Li, and Jun Zhao. 2024. [Tug-of-war between knowledge: Exploring and resolving knowledge conflicts in retrieval-augmented language models](#). In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation, LREC/COLING 2024, 20-25 May, 2024, Torino, Italy*, pages 16867–16878. ELRA and ICCL.

- Ehsan Kamaloo, Nouha Dziri, Charles Clarke, and Davoud Rafiei. 2023. [Evaluating open-domain question answering in the era of large language models](#). In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 5591–5606, Toronto, Canada. Association for Computational Linguistics.
- Aishwarya Kamath, Johan Ferret, Shreya Pathak, Nino Vieillard, Ramona Merhej, Sarah Perrin, Tatiana Matejovicova, Alexandre Ramé, Morgane Rivière, Louis Rouillard, Thomas Mesnard, Geoffrey Cideron, Jean-Bastien Grill, Sabela Ramos, Edouard Yvinec, Michelle Casbon, Etienne Pot, Ivo Penchev, Gaël Liu, Francesco Visin, Kathleen Kenealy, Lucas Beyer, Xiaohai Zhai, Anton Tsitsulin, Róbert Busa-Fekete, Alex Feng, Noveen Sachdeva, Benjamin Coleman, Yi Gao, Basil Mustafa, Iain Barr, Emilio Parisotto, David Tian, Matan Eyal, Colin Cherry, Jan-Thorsten Peter, Danila Sinopalnikov, Surya Bhatipatiraju, Rishabh Agarwal, Mehran Kazemi, Dan Malkin, Ravin Kumar, David Vilar, Idan Brusilovsky, Jiaming Luo, Andreas Steiner, Abe Friesen, Abhanshu Sharma, Abheesht Sharma, Adi Mayrav Gilady, Adrian Goedeckemeyer, Alaa Saade, Alexander Kolesnikov, Alexei Bendebury, Alvin Abdagic, Amit Vadi, András György, André Susano Pinto, Anil Das, Ankur Bapna, Antoine Miech, Antoine Yang, Antonia Paterson, Ashish Shenoy, Ayan Chakrabarti, Bilal Piot, Bo Wu, Bobak Shahriari, Bryce Petriani, Charlie Chen, Charline Le Lan, Christopher A. Choquette-Choo, CJ Carey, Cormac Brick, Daniel Deutsch, Danielle Eisenbud, Dee Cattle, Derek Cheng, Dimitris Paparas, Divyashree Shivakumar Sreepathihalli, Doug Reid, Dustin Tran, Dustin Zelle, Eric Noland, Erwin Huizenga, Eugene Kharitonov, Frederick Liu, Gagik Amirkhanyan, Glenn Cameron, Hadi Hashemi, Hanna Klimczak-Plucinska, Harman Singh, Harsh Mehta, Harshal Tushar Lehri, Hussein Hazimeh, Ian Ballantyne, Idan Szpektor, Ivan Nardini, Jean Pouget-Abadie, Jetha Chan, Joe Stanton, John Wieting, Jonathan Lai, Jordi Orbay, Joseph Fernandez, Josh Newlan, Ju-yeong Ji, Jyotinder Singh, Kat Black, Kathy Yu, Kevin Hui, Kiran Vodrahalli, Klaus Greff, Linhai Qiu, Marcella Valentine, Marina Coelho, Marvin Ritter, Matt Hoffman, Matthew Watson, Mayank Chaturvedi, Michael Moynihan, Min Ma, Nabila Babar, Natasha Noy, Nathan Byrd, Nick Roy, Nikola Momchev, Nilay Chauhan, Oskar Bunyan, Pankil Botarda, Paul Caron, Paul Kishan Rubenstein, Phil Culliton, Philipp Schmid, Pier Giuseppe Sessa, Pingmei Xu, Piotr Stanczyk, Pouya Tafti, Rakesh Shivanna, Renjie Wu, Renke Pan, Reza Rokni, Rob Willoughby, Rohith Vallu, Ryan Mullins, Sammy Jerome, Sara Smoot, Sertan Girgin, Shariq Iqbal, Shashir Reddy, Shruti Sheth, Siim Pöder, Sijal Bhatnagar, Sindhu Raghuram Panyam, Sivan Eiger, Susan Zhang, Tianqi Liu, Trevor Yacovone, Tyler Liechty, Uday Kalra, Utku Evci, Vedant Misra, Vincent Roseberry, Vlad Feinberg, Vlad Kolesnikov, Woohyun Han, Woosuk Kwon, Xi Chen, Yinlam Chow, Yuvein Zhu, Zichuan Wei, Zoltan Egyed, Victor Cotruta, Minh Giang, Phoebe Kirk, Anand Rao, Jessica Lo, Erica Moreira, Luiz Gustavo Martins, Omar Sanseviero, Lucas Gonzalez, Zach Gleicher, Tris Warkentin, Vahab Mirrokni, Evan Senter, Eli Collins, Joelle K. Barral, Zoubin Ghahramani, Raia Hadsell, Yossi Matias, D. Sculley, Slav Petrov, Noah Fiedel, Noam Shazeer, Oriol Vinyals, Jeff Dean, Demis Hassabis, Koray Kavukcuoglu, Clément Farabet, Elena Buchatskaya, Jean-Baptiste Alayrac, Rohan Anil, Dmitry (Dima) Lepikhin, Sebastian Borgeaud, Olivier Bachem, Armand Joulin, Alek Andreev, Cassidy Hardin, Robert Dadashi, and Léonard Hussenot. 2025. [Gemma 3 technical report](#). *CoRR*, abs/2503.19786.
- Anant Khandelwal, Manish Gupta, and Puneet Agrawal. 2025. [Cocoa: Confidence and context-aware adaptive decoding for resolving knowledge conflicts in large language models](#). *CoRR*, abs/2508.17670.
- Jungyeon Lee, Kangmin Lee, and Taeuk Kim. 2025. [MAGIC: A multi-hop and graph-based benchmark for inter-context conflicts in retrieval-augmented generation](#). *CoRR*, abs/2507.21544.
- Patrick Lewis, Ethan Perez, Aleksandra Piktus, Fabio Petroni, Vladimir Karpukhin, Naman Goyal, Heinrich Küttler, Mike Lewis, Wen-tau Yih, Tim Rocktäschel, Sebastian Riedel, and Douwe Kiela. 2020. [Retrieval-augmented generation for knowledge-intensive NLP tasks](#). In *Advances in Neural Information Processing Systems 33: Annual Conference on Neural Information Processing Systems 2020, NeurIPS 2020, December 6-12, 2020, virtual*.
- Dawei Li, Bohan Jiang, Liangjie Huang, Alimohammad Beigi, Chengshuai Zhao, Zhen Tan, Amrita Bhatnagar, Yuxuan Jiang, Canyu Chen, Tianhao Wu, Kai Shu, Lu Cheng, and Huan Liu. 2024. [From generation to judgment: Opportunities and challenges of llm-as-a-judge](#). *CoRR*, abs/2411.16594.
- Nelson F. Liu, Kevin Lin, John Hewitt, Ashwin Paranjape, Michele Bevilacqua, Fabio Petroni, and Percy Liang. 2024. [Lost in the middle: How language models use long contexts](#). *Trans. Assoc. Comput. Linguistics*, 12:157–173.
- Shayne Longpre, Kartik Perisetla, Anthony Chen, Nikhil Ramesh, Chris DuBois, and Sameer Singh. 2021. [Entity-based knowledge conflicts in question answering](#). In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing, EMNLP 2021, Virtual Event / Punta Cana, Dominican Republic, 7-11 November, 2021*, pages 7052–7063. Association for Computational Linguistics.
- Alex Mallen, 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)*, *ACL 2023, Toronto, Canada, July 9-14, 2023*, pages 9802–9822. Association for Computational Linguistics.

- Sara Marjanovic, Haeun Yu, Pepa Atanasova, Maria Maistro, Christina Lioma, and Isabelle Augenstein. 2024. [DYNAMICQA: tracing internal knowledge conflicts in language models](#). In *Findings of the Association for Computational Linguistics: EMNLP 2024, Miami, Florida, USA, November 12-16, 2024*, pages 14346–14360. Association for Computational Linguistics.
- Francesco Ortu, Zhijing Jin, Diego Doimo, Mrinmaya Sachan, Alberto Cazzaniga, and Bernhard Schölkopf. 2024. [Competition of mechanisms: Tracing how language models handle facts and counterfactuals](#). In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), ACL 2024, Bangkok, Thailand, August 11-16, 2024*, pages 8420–8436. Association for Computational Linguistics.
- Quang Pham, Hoang Ngo, Anh Tuan Luu, and Dat Quoc Nguyen. 2024. [Who’s who: Large language models meet knowledge conflicts in practice](#). In *Findings of the Association for Computational Linguistics: EMNLP 2024, Miami, Florida, USA, November 12-16, 2024*, pages 10142–10151. Association for Computational Linguistics.
- Dan Shi, Renren Jin, Tianhao Shen, Weilong Dong, Xinwei Wu, and Deyi Xiong. 2024a. [IRCAN: mitigating knowledge conflicts in LLM generation via identifying and reweighting context-aware neurons](#). In *Advances in Neural Information Processing Systems 38: Annual Conference on Neural Information Processing Systems 2024, NeurIPS 2024, Vancouver, BC, Canada, December 10 - 15, 2024*.
- Weijia Shi, Xiaochuang Han, Mike Lewis, Yulia Tsvetkov, Luke Zettlemoyer, and Wen-tau Yih. 2024b. [Trusting your evidence: Hallucinate less with context-aware decoding](#). In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies: Short Papers, NAACL 2024, Mexico City, Mexico, June 16-21, 2024*, pages 783–791. Association for Computational Linguistics.
- Zhaochen Su, Jun Zhang, Xiaoye Qu, Tong Zhu, Yanshu Li, Jiashuo Sun, Juntao Li, Min Zhang, and Yu Cheng. 2024. [Conflictbank: A benchmark for evaluating the influence of knowledge conflicts in llms](#). In *Advances in Neural Information Processing Systems 38: Annual Conference on Neural Information Processing Systems 2024, NeurIPS 2024, Vancouver, BC, Canada, December 10 - 15, 2024*.
- Yi Sui, Chaozhuo Li, Chen Zhang, Dawei song, and Qiuchi Li. 2025. [Bridging external and parametric knowledge: Mitigating hallucination of llms with shared-private semantic synergy in dual-stream knowledge](#). *CoRR*, abs/2506.06240.
- Kaiser Sun, Fan Bai, and Mark Dredze. 2025. [Task matters: Knowledge requirements shape llm responses to context-memory conflict](#). *CoRR*, abs/2506.06485.
- Hexiang Tan, Fei Sun, Wanli Yang, Yuanzhuo Wang, Qi Cao, and Xueqi Cheng. 2024. [Blinded by generated contexts: How language models merge generated and retrieved contexts when knowledge conflicts?](#) In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), ACL 2024, Bangkok, Thailand, August 11-16, 2024*, pages 6207–6227. Association for Computational Linguistics.
- Ruixiang Tang, Yu-Neng Chuang, and Xia Hu. 2023. [The science of detecting llm-generated texts](#). *CoRR*, abs/2303.07205.
- Wei Tang, Yixin Cao, Yang Deng, Jiahao Ying, Bo Wang, Yizhe Yang, Yuyue Zhao, Qi Zhang, Xu-anjing Huang, Yu-Gang Jiang, and Yong Liao. 2025. [Ewowski: Evaluating llms on evolving knowledge](#). In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), ACL 2025, Vienna, Austria, July 27 - August 1, 2025*, pages 948–964. Association for Computational Linguistics.
- Zineddine Tighidet, Andrea Mogini, Hedi Ben-younes, Jiali Mei, Patrick Gallinari, and Benjamin Piwowarski. 2025. [Context copying modulation: The role of entropy neurons in managing parametric and contextual knowledge conflicts](#). *Preprint*, arXiv:2509.10663.
- Hitesh Wadhwa, Rahul Seetharaman, Somyaa Aggarwal, Reshmi Ghosh, Samyadeep Basu, Soundararajan Srinivasan, Wenlong Zhao, Shreyas Chaudhari, and Ehsan Aghazadeh. 2024. [From rags to rich parameters: Probing how language models utilize external knowledge over parametric information for factual queries](#). *CoRR*, abs/2406.12824.
- Fei Wang, Xingchen Wan, Ruoxi Sun, Jiefeng Chen, and Sercan Ö. Arik. 2025a. [Astute RAG: overcoming imperfect retrieval augmentation and knowledge conflicts for large language models](#). In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), ACL 2025, Vienna, Austria, July 27 - August 1, 2025*, pages 30553–30571. Association for Computational Linguistics.
- Yike Wang, Shangbin Feng, Heng Wang, Weijia Shi, Vidhisha Balachandran, Tianxing He, and Yulia Tsvetkov. 2023. [Resolving knowledge conflicts in large language models](#). *CoRR*, abs/2310.00935.
- Yilin Wang, Heng Wang, Yuyang Bai, and Minnan Luo. 2025b. [Continuously steering llms sensitivity to contextual knowledge with proxy models](#). *CoRR*, abs/2508.19720.
- Yuhao Wang, Ruiyang Ren, Yucheng Wang, Wayne Xin Zhao, Jing Liu, Hua Wu, and Haifeng Wang. 2025c. [Unveiling knowledge utilization mechanisms in llm-based retrieval-augmented generation](#). In *Proceedings of the 48th International ACM SIGIR Conference on Research and Development in Information*

- Retrieval, *SIGIR 2025, Padua, Italy, July 13-18, 2025*, pages 1262–1271. ACM.
- Yuqing Wang and Yun Zhao. 2024. **TRAM: benchmarking temporal reasoning for large language models**. In *Findings of the Association for Computational Linguistics, ACL 2024, Bangkok, Thailand and virtual meeting, August 11-16, 2024*, pages 6389–6415. Association for Computational Linguistics.
- Jerry W. Wei, Da Huang, Yifeng Lu, Denny Zhou, and Quoc V. Le. 2023. **Simple synthetic data reduces sycophancy in large language models**. *CoRR*, abs/2308.03958.
- Wikidata contributors. 2025. Wikidata. <https://www.wikidata.org/>.
- Rongwu Xu, Zehan Qi, Zhijiang Guo, Cunxiang Wang, Hongru Wang, Yue Zhang, and Wei Xu. 2024. **Knowledge conflicts for llms: A survey**. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing, EMNLP 2024, Miami, FL, USA, November 12-16, 2024*, pages 8541–8565. Association for Computational Linguistics.
- An Yang, Anfeng Li, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Gao, Chengen Huang, Chenxu Lv, Chujie Zheng, Dayiheng Liu, Fan Zhou, Fei Huang, Feng Hu, Hao Ge, Haoran Wei, Huan Lin, Jialong Tang, Jian Yang, Jianhong Tu, Jianwei Zhang, Jian Yang, Jiaxi Yang, Jingren Zhou, Junyang Lin, Kai Dang, Keqin Bao, Kexin Yang, Le Yu, Lianghao Deng, Mei Li, Mingfeng Xue, Mingze Li, Pei Zhang, Peng Wang, Qin Zhu, Rui Men, Ruize Gao, Shixuan Liu, Shuang Luo, Tianhao Li, Tianyi Tang, Wenbiao Yin, Xingzhang Ren, Xinyu Wang, Xinyu Zhang, Xuancheng Ren, Yang Fan, Yang Su, Yichang Zhang, Yinger Zhang, Yu Wan, Yuqiong Liu, Zekun Wang, Zeyu Cui, Zhenru Zhang, Zhipeng Zhou, and Zihan Qiu. 2025. **Qwen3 technical report**. *CoRR*, abs/2505.09388.
- Evelyn Yee, Alice Li, Chenyu Tang, Yeon Ho Jung, Ramamohan Paturi, and Leon Bergen. 2024. **Disso- ciation of faithful and unfaithful reasoning in llms**. *CoRR*, abs/2405.15092.
- Qinan Yu, Jack Merullo, and Ellie Pavlick. 2023. **Char- acterizing mechanisms for factual recall in language models**. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing, EMNLP 2023, Singapore, December 6-10, 2023*, pages 9924–9959. Association for Computational Linguistics.
- Michael J. Q. Zhang and Eunsol Choi. 2023. **Mitigating temporal misalignment by discarding outdated facts**. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing, EMNLP 2023, Singapore, December 6-10, 2023*, pages 14213–14226. Association for Computational Linguistics.
- Yu Zhao, Alessio Devoto, Giwon Hong, Xiaotang Du, Aryo Pradipta Gema, Hongru Wang, Xuanli He, Kam-Fai Wong, and Pasquale Minervini. 2025. **Steer- ing knowledge selection behaviours in llms via sae- based representation engineering**. In *Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies, NAACL 2025 - Volume 1: Long Papers, Albuquerque, New Mexico, USA, April 29 - May 4, 2025*, pages 5117–5136. Association for Computational Linguistics.
- Wenxuan Zhou, Sheng Zhang, Hoifung Poon, and Muhao Chen. 2023. **Context-faithful prompting for large language models**. In *Findings of the Association for Computational Linguistics: EMNLP 2023, Singapore, December 6-10, 2023*, pages 14544–14556. Association for Computational Linguistics.

A Additional Related Work

Understanding Knowledge Conflicts. Several studies seek to uncover the internal mechanisms by which models integrate or suppress conflicting knowledge. Ortu et al. (2024) identify two competing pathways for factual recall – one retrieving from parametric memory and one interpreting contextual input – but leave open how the model decides between them. Wang et al. (2025c) extend this line through macro- and micro-level analyses of module-level contributions in RAG systems. Using causal mediation analysis, Wadhwa et al. (2024) and Farahani and Johansson (2024) show that contextual evidence often dominates model predictions when both sources are available. Other interpretability works locate architectural correlates of conflict resolution: Yu et al. (2023) identify attention heads that promote memorized or contextual answers, and Tighidet et al. (2025) reveal “entropy neurons” that suppress direct context copying. Together, these findings suggest a competition of mechanisms underlying knowledge integration, yet it remains unclear what signals – such as temporal cues or perceived fact mutability – guide their arbitration.

Controlling Behavior under Conflict. A growing body of work attempts to *steer* models toward desired knowledge sources. Prompt-based methods encourage context faithfulness by explicitly framing external information as authoritative (Zhou et al., 2023). Decoding-time methods adjust token probabilities to favor evidence-supported generations (Shi et al., 2024b; Khandelwal et al., 2025). Mechanistic interventions further refine this control. Shi et al. (2024a) identify and reweight neurons sensitive to contextual cues, while Bi et al. (2025) and Zhao et al. (2025) use representation-level steering to regulate knowledge reliance. Other approaches propose structured reasoning pipelines or auxiliary models: Wang et al. (2023) introduce a three-step protocol to localize and reconcile conflicts, Wang et al. (2025b) employ proxy models to adaptively modulate context sensitivity, and Sui et al. (2025) redesign attention mechanisms to separate “shared” and “private” semantic channels. Finally, Wang et al. (2025a) propose a RAG framework that dynamically weighs internal and external sources based on estimated reliability. Despite their variety, these methods primarily aim to enforce one source’s dominance, rather than endow models with the ability to reason about which source

should be trusted in a given situation.

Causes of Knowledge Conflicts. Context–memory conflicts generally originate from two broad causes. **Temporal misalignment** occurs when the model’s pretraining data become outdated relative to the current world state. Because retraining large models is prohibitively expensive, parametric memory often lags behind factual updates, creating tension with newer contextual information (e.g., in retrieval-augmented systems (Lewis et al., 2020)). **Misinformation pollution**, by contrast, arises when false or adversarial content contradicts accurate internal knowledge. This can result from maliciously retrieved documents, user manipulation, or synthetic misinformation generated by other models. LLMs are particularly vulnerable due to their tendency toward sycophancy – agreeing with user-provided statements regardless of veracity (Wei et al., 2023). Detecting whether conflicting information reflects a genuine update or an intentional falsehood remains a key challenge (Tang et al., 2023). While misinformation conflicts demand epistemic vigilance, temporally induced conflicts demand adaptability – our work focuses on the latter as a natural, non-adversarial source of discrepancy.

B Additional Experimental Details

B.1 Dataset Details

An overview of the dataset generation procedure can be found in Figure 5.

Example of a counterfactual context

Question: Which sports team does Juan Soto play for?

Ground-truth: New York Mets

Counterfactual answer: San Diego Padres

Generated paragraph: Following a blockbuster trade in early 2025, Juan Soto has become a cornerstone of the San Diego Padres lineup. His arrival significantly bolstered the team’s offensive capabilities, quickly establishing him as a fan favorite and a key contributor to their postseason aspirations. Soto’s powerful hitting and exceptional plate discipline have already solidified his position as the Padres primary designated hitter.

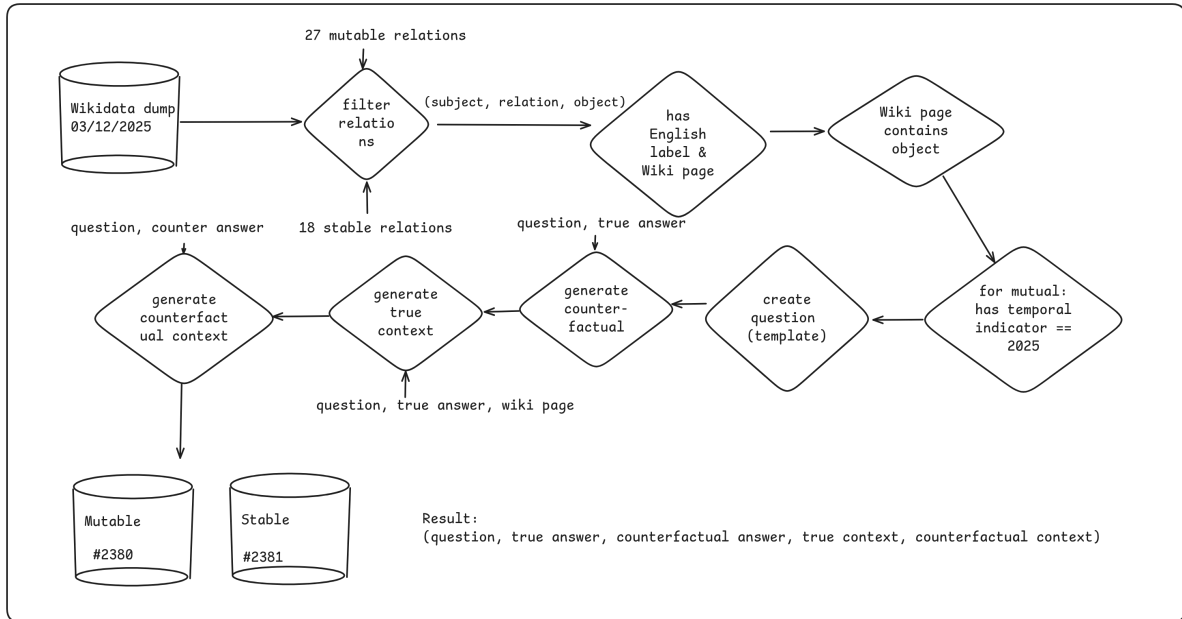


Figure 5: Overview of the dataset generation procedure.

B.1.1 Dataset Statistics

Final dataset relations and relation counts can be found in Table 7.

Stable		Updated	
Relation	Count	Relation	Count
place of birth	395	member of sports team	850
taxon rank	389	position held	448
place of death	370	award received	441
father	282	chairperson	202
country of citizenship	281	officeholder	90
political party	234	country for sport	81
native language	194	employer	65
religion	135	significant event	38
mother	101	residence	29
		headquarters location	25
		member of (organization)	25
		replaces	24
		owner of	5
		participant	3
		U.S. Cong. district	2
		subject has role	1
		place of work	1
Total Stable	2381	Total Updated	2350

Table 7: Distribution of relation types and counts for the Stable and Updated dataset splits.

B.1.2 Dataset Quality Control

We perform manual quality control for the several LLM-generated parts of the dataset. Specifically, we sample 50 examples of counterfactual facts, counterfactual paragraphs, and true paragraphs. For counterfactual facts, we check if the generated facts differ from the ground-truth answers and whether they match the correct category (e.g., if the answer is a person, the counter answer should also be a person). This was the case in 50/50 examples. For generated correct/counterfactual paragraphs,

we check if the generated paragraph correctly implies the correct/counterfactual answer. This was the case in 50/50 cases for the counterfactual paragraphs and 50/50 cases for the true paragraphs.

Correct vs. Counterfactual Paragraphs We analyzed paragraph quality by computing perplexity for both context types and find them to be linguistically very similar (mean perplexity: 20.7 ± 12.1 for true contexts vs. 22.9 ± 8.9 for counterfactual contexts, Cohen’s D effect size -0.20 , small effect). The small, systematic increase in surprisal for counterfactual contexts does not constitute an out-of-distribution shift and is likely attributable to the factually incorrect content rather than stylistic differences. We are therefore confident that the two context types are sufficiently comparable for the purposes of our experiments.

Mutability Judgments By construction, the data in our two WIKIRECENTCHANGES splits stable and updated contain relatively stable and mutable information respectively. However, the model’s perception of whether the facts are likely to change or not also matter when evaluating the role of mutability in resolving knowledge conflicts. To study that, we have Qwen-3-14B predict mutability as a sanity check and a level of mutability ranging from highly mutable to immutable (high, medium, low, immutable). The results are presented in Table 9.

Split	Relation	Query Template
Stable	place of birth	Where was {subject} born?
	taxon rank	What is the taxon rank of {subject}?
	place of death	Where did {subject} die?
	father	Who is the father of {subject}?
	country of citizenship	What is the country of citizenship of {subject}?
	political party	Which political party does {subject} belong to?
	native language	What is the native language of {subject}?
	religion	What is the religion of {subject}?
	mother	Who is the mother of {subject}?
	Updated	member of sports team
position held		Which position is held by {subject}?
award received		Which award did {subject} receive?
chairperson		Who is the chairperson of {subject}?
officeholder		Who is the officeholder for {subject}?
country for sport		Which country does {subject} represent in sports?
employer		Who is the employer of {subject}?
significant event		What is a significant event associated with {subject}?
residence		Where is the residence of {subject}?
headquarters location		Where is the headquarters of {subject}?
member of (org.)		Which organization is {subject} a member of?
replaces		Who did {subject} replace?
owner of		What does {subject} own?
participant		In which event was {subject} a participant?
U.S. Cong. district		Which U.S. Congressional district is represented by {subject}?
subject has role		What role does {subject} have?
place of work		Where is the place of work of {subject}?

Table 8: Query templates used for constructing questions based on Wikidata relations.

Split	% Mutable	high/med	low/imm
Stable	32	419	1926
Updated	85	1875	454

Table 9: Mutability judgments by Qwen-3-14B on the two splits, with counts at each mutability level

B.2 Model Details

Additional information on the models used in this study is provided in Table 11. For all models, we used the default parameters specified in the respective huggingface generation configs.

B.3 Human Evaluation

We perform further human alignment on moving parts such as the metrics used in this paper and the LLM-as-a-judge system.

B.3.1 Metrics

For metrics, we select 50 answer & ground-truth pairs and label them. Then we calculate the agreement rate between our human judgment and several automatic metrics. In addition to the Match and OpenEval (OE) metrics that we introduced in Section 3.3, we also compute the agreement rates for the BERT answer equivalence metric (BEM) (Bulian et al., 2022) and an LLM-as-a-judge sys-

tem (Table 12). The BEM metric uses a BERT model specifically trained for evaluating equivalence between candidate and reference answers (when given a question and the two answers). The LLM-as-a-judge system that we employed is a Qwen-3-14B model that we prompted to evaluate semantic equivalence. To predict this, it was also supplied with question, candidate, and reference (ground-truth) answers. The exact prompt can be found in the GitHub at `src/openevalv2.py`.

From the model-based metrics, we decided on selecting OpenEval in its original form since it achieved the best agreement rate. Additionally, we included results on the Match metric, as this is non-parametric and still reasonably good.

B.3.2 Temporal Reasoning Detection

To evaluate the performance of our LLM-as-a-judge system that judges whether the models performed temporal reasoning (with regard to mutability), we also compute agreement rates with a human annotator. Similar to the evaluation of the metrics, we randomly select 50 examples and blindly judge whether the model used mutability to accept or reject contexts. The human judge annotated with access to only the reasoning traces. The agreement rate was 94%. The specific prompt used to judge temporal reasoning can be found in

Question	Answer	Counterfactual	Counterfactual Paragraph
Which sports team does Gianluca Gaetano play for?	Cagliari Calcio	Atalanta	Gianluca Gaetano has been a mainstay of Atalantas midfield since his transfer from Napoli in July 2025. His dynamic playmaking abilities and precise passing have quickly made him a fan favorite, contributing significantly to Atalantas surprising run to the Serie A title in the 2025-26 season. Gaetanos tactical versatility allows manager Gian Piero Gasperini to deploy him in various roles, solidifying his position as a key player for the club.
Which position is held by Adriane Lopes?	Mayor of Campo Grande	Governor of Mato Grosso do Sul	Adriane Lopes assumed the governorship of Mato Grosso do Sul on January 1, 2025, following a landslide victory that signaled a desire for significant economic and environmental reform in the state. Her administration has focused on sustainable agriculture and expanding infrastructure along the Paraguay River, aiming to position Mato Grosso do Sul as a national leader in both sectors. Lopess early policies have been largely well-received, despite facing challenges from established agricultural interests.
What is the country of citizenship of Bert Folkard?	Australia	New Zealand	Bert Folkard was a prominent New Zealand scientist known for his pioneering work in marine biology and oceanographic research. Born and raised in Wellington, he dedicated his career to studying the unique ecosystems of the New Zealand coastline and the surrounding Pacific Ocean. Folkards research contributed significantly to the nation's understanding of its marine resources and helped inform sustainable fisheries management practices.

Table 10: Examples of generated counterfactual paragraphs.

Model Name	Mode Size	Notes	Cutoff
Gemma-3-4B	4B	/	Aug 2024
Llama-3.3-70B	70B	4bit	Dec 2023
Phi-4-mini	4B	/	Mar 2025
Qwen-3-8B	8B	/	End of 2023
GPT-OSS-20B	21B parameters with 3.6B active parameters	4bit	Jun 2024
GPT-OSS-120B	117B parameters with 5.1B active parameters	4bit	Jun 2024
OLMO-3-7B	7B	/	End of 2024
DS-Llama-8B	8B	8bit	End of 2023
Qwen-3-14B	12B	Judge model used for evaluation and dataset generation	Aug 2024

Table 11: Summary of different models with their respective details.

Metric	Agreement Rate
Match	0.898
OpenEval	0.959
LLM-as-a-judge	0.939
BEM	0.939

Table 12: Agreement rate between human judgment and metrics.

Appendix B.4.3.

B.4 Prompts

B.4.1 Question Answering

The following contains the QA prompts used throughout the paper. The several settings are closed-book (Figure 6), the standard RAG prompt with context (Figure 7), the additional RAG prompt containing cutoff and current dated ("Dates") (Figure 8) and the prompt which especially emphasizes the use of mutability reasoning ("Prompted") (Figure 9).

B.4.2 Dataset Creation

To create the dataset, we utilize prompts to generate counterfactual answers (Figure 11), counterfactual

paragraphs (Figure 12), and true paragraphs (Figure 13).

B.4.3 Temporal Reasoning

For the evaluation of temporal reasoning, we utilize three different prompts; one for identifying if temporal reasoning was used (Figure 14), whether conflicts were identified (Figure 15), and whether mutability of the underlying fact was identified (Figure 16).

C Complimentary Results

C.1 Knowledge Conflict Behavior

We additionally present the full results for the knowledge conflict evaluation for the standard setting (Table 13), when adding cutoff and current dates (Table 14), and when additionally prompting the model to use temporal reasoning (Table 15).

C.1.1 Stratify: did the model know the answer before?

We additionally experiment with stratifying the results in Table 2 by whether the model knew the answer or not without context. The rationale is that knowing the answer beforehand might have an

Closed-Book QA Prompt

Please answer the provided question based on your knowledge. Your task is to provide what you believe to be the correct answer and explain your reasoning.

Question: {question}

Instructions

- ** 1. Answer the question based on your knowledge
- ** 2. Assess your confidence in the answer
- ** 3. Note any uncertainty or alternative possibilities

Respond in JSON format:

```
{  
  "answer": "couple words only, no full sentence",  
  "rationale": "Brief explanation of your reasoning and confidence level"  
}
```

Example:

```
{  
  "answer": "Paris",  
  "rationale": "I am confident Paris is the capital of France..."  
}
```

Now answer the question above:

Answer:

Figure 6: QA prompt without context.

effect on whether true/counterfactual paragraphs are to be trusted. The results are presented in Table 16. Apparently, knowing the answer has little to no effect on whether contexts are accepted. However, knowing the answer results in higher rates of original (CB) answers being given.

C.2 Does the model use mutability in reasoning?

An example reasoning trace is shown in Figure 10.

D Checklist

D.1 General Questions

D.1.1 Did you describe the limitations of your work?

yes, see Limitations

D.1.2 Did you discuss any potential risks of your work?

Yes, see Ethics Statement

D.2 Scientific Artifacts

D.2.1 Did you cite the creators of artifacts you used?

Yes, see Section 3.1

D.2.2 Did you discuss the license or terms for use and / or distribution of any artifacts?

Our dataset derives its content from Wikidata, which requires a license compatible with the Creative Commons Public Domain Dedication (CC0 1.0) license.

QA with Context Prompt

You have access to a retrieved context below. Please answer the provided question with what you believe to be true. Your task is NOT simply to extract an answer from the context, but to critically evaluate whether the context should be trusted.

Context: {context}
Question: {question}

Instructions

- ** 1. Assess whether the context appears trustworthy and relevant
- ** 2. Consider if the information conflicts with what you know
- ** 3. Decide whether to answer using the context, your own knowledge, or both

Respond in JSON format:

```
{  
  "answer": "couple words only, no full sentence",  
  "rationale": "Brief explanation of whether you trusted the context"  
}
```

Example:

```
{  
  "answer": "Paris",  
  "rationale": "Context provided correct information that aligns..."  
}
```

Do not default to using the context just because it's provided.
Now answer the question above:
Answer:

Figure 7: RAG prompt with context.

D.2.3 Did you discuss if your use of existing artifact(s) was consistent with their intended use, provided that it was specified? For the artifacts you create, do you specify intended use and whether that is compatible with the original access conditions (in particular, derivatives of data accessed for research purposes should not be used outside of research contexts)?

The used artifacts can be shared. Our usage was consistent with these specifications of being public domain.

D.2.4 Did you discuss the steps taken to check whether the data that was collected / used contains any information that names or uniquely identifies individual people or offensive content, and the steps taken to protect / anonymize it?

The data we use is derived from Wikidata. This may contain names of public figures such as presidents, government figures, or prominent people, and therefore, did not anonymize their names.

D.2.5 Did you provide documentation of the artifacts, e.g., coverage of domains, languages, and linguistic phenomena, demographic groups represented, etc.?

We cover an overview of the used artifact in Section 3.1 and B.1.

D.2.6 Did you report relevant statistics like the number of examples, details of train / test / dev splits, etc. for the data that you used / created?

We cover an overview of the used artifacts in Section 3.1.

D.3 Computational Experiments

D.3.1 Did you report the number of parameters in the models used, the total computational budget (e.g., GPU hours), and computing infrastructure used?

The model parameters are listed in Table 11. We did not train or fine-tune models, but ran inference on a larger set of models. Our used infrastructure was a mixture of A100s with 40/80GB memory. Running the entire analysis may take ca. 3 days on one GPU per model, resulting in 21 GPU/days for

```

QA with Context Prompt (Dates)

Today is {date}. Your knowledge cutoff is {cutoff}.

You have access to a retrieved context below. Please answer the provided question with
what you believe to be true. Your task is NOT simply to extract an answer from the context, but to
critically evaluate whether the context should be trusted.

Context: {context}
Question: {question}

## Instructions
** 1. Assess whether the context appears trustworthy and relevant
** 2. Consider if the information conflicts with what you know
** 3. Decide whether to answer using the context, your own knowledge, or both

Respond in JSON format:
{
  "answer": "couple words only, no full sentence",
  "rationale": "Brief explanation"
}

Do not default to using the context just because it's provided.
Now answer the question above:
Answer:

```

Figure 8: RAG prompt with context and temporal grounding.

all models combined.

D.3.2 Did you discuss the experimental setup, including hyperparameter search and best-found hyperparameter values?

We did not run a hyperparameter search, but we described our experimental setup both in Section 3.1 and made our code available. The repository contains hyperparameters, prompts, etc.

D.3.3 Did you report descriptive statistics about your results (e.g., error bars around results, summary statistics from sets of experiments), and is it transparent whether you are reporting the max, mean, etc. or just a single run?

We are reporting single-run results, since this study contains descriptive results and does not try to show a clear benefit of using one model over another.

D.3.4 If you used existing packages (e.g., for preprocessing, for normalization, or for evaluation, such as NLTK, Spacy, ROUGE, etc.), did you report the implementation, model, and parameter settings used?

Details available at the corresponding GitHub repository.

D.4 Human Annotators

D.4.1 Did you report the full text of instructions given to participants, including, e.g., screenshots, disclaimers of any risks to participants or annotators, etc.?

We did not use external annotators. We had co-authors label elements for our small human agreement studies. They did so blindly, having access to only the information required to take a decision (e.g., model answer and ground-truth for metrics agreement). More details in Section B.3.

D.4.2 Did you report information about how you recruited (e.g., crowdsourcing platform, students) and paid participants, and discuss if such payment is adequate given the participants' demographic (e.g., country of residence)?

Only used authors.

D.4.3 Did you discuss whether and how consent was obtained from people whose data you're using/curating?

N/A, did not use external human annotators

QA with Context (Dates + Temporal Reasoning) Prompt

Today is {date}. Your knowledge cutoff is {cutoff}.

You have access to a retrieved context below. Please answer the provided question with what you believe to be true. Your task is NOT simply to extract an answer from the context, but to critically evaluate whether the context should be trusted. Please put special emphasis on whether a question's answer will change over time and use that information to select/reject the provided context.

Context: {context}

Question: {question}

Instructions

- ** 1. Assess whether the context appears trustworthy and relevant
- ** 2. Consider if the information conflicts with what you know
- ** 3. Decide whether to answer based on mutability (whether the information is likely to change)

Respond in JSON format:

```
{  
  "answer": "couple words only, no full sentence",  
  "rationale": "Brief explanation"  
}
```

Do not default to using the context just because it's provided.

Now answer the question above:

Answer:

Figure 9: RAG prompt with dates and emphasis on mutability.

D.4.4 Was the data collection protocol approved (or determined exempt) by an ethics review board?

N/A, did not use external human annotators

D.4.5 Did you report the basic demographic and geographic characteristics of the annotator population that is the source of the data?

N/A, did not use external human annotators

D.5 Use of AI Assistants

D.5.1 Did you include information about your use of AI assistants?

We did not use AI assistants to perform research directly. We did use ChatGPT, Gemini, and Claude to perform reformulations of existing text, as well as to fix grammatical errors and iterate over the text using LanguageTool and Grammarly.

Full LLM Response

analysis:

We need to evaluate context: "As of 2025, Adriano Grimaldi plays as a forward for 1. FC Nürnberg..." We must answer: Which sports team does Adriano Grimaldi play for? The answer: 1. FC Nürnberg. But we must evaluate trustworthiness. We know knowledge cutoff Jun 2024; we don't have data on 2025. The context says "As of 2025" so it's future relative to cutoff. So the context might be hypothetical. Could be correct, but likely to change.

However, the question is about current team as of December 10, 2025, so context is relevant. But we must consider that the context might be inaccurate. We don't have other info... The instructions: "Do not default to using the context just because it's provided." But we can use it if it seems correct. The context seems plausible. He is a forward for 1. FC Nürnberg, a German club. So answer: 1. FC Nürnberg.

assistantfinal:

```
{
  "answer": "FC Nürnberg",
  "rationale": "Context states he plays for 1. FC Nürnberg and no conflicting info; likely accurate for 2025."
}
```

Figure 10: Example reasoning trace of GPT-OSS-20B.

Counterfactual Generation Prompt

Prompt for Generating Plausible Counterfactual Answers

You will be given: 1. A question, 2. The correct answer

Your task is to generate ONE plausible but incorrect alternative answer (counterfactual) that could believably answer the question.

Requirements for the counterfactual answer

- ** 1. **Plausibility:** The counterfactual should be a reasonable answer to the question type.
- ** 2. **Semantic similarity:** Choose an alternative that is closely related to the correct answer.
- ** 3. **Avoid obvious contradictions:** Don't create logical impossibilities.
- ** 4. **Maintain factual structure:** Maintain the structure of the correct answer.

Output format

Respond ONLY with a JSON object in this exact format:

```
{
  "counterfactual_answer": "your counterfactual answer here",
  "justification": "brief explanation"
}
```

Examples

- ** **Input:** Question: "What sport did Wayne Gretzky play?" | Correct: "Ice hockey"
- ** **Output:** {"counterfactual_answer": "Lacrosse", "justification": "..."}
- ** **Input:** Question: "What is the capital of France?" | Correct: "Paris"
- ** **Output:** {"counterfactual_answer": "Lyon", "justification": "..."}

Your Input

- ** **Question:** {question}
- ** **Correct Answer:** {answer}

Now generate the counterfactual answer in the JSON format specified above.

** **Output:**

Figure 11: Counterfactual generation prompt.

Counterfactual Paragraph Generation Prompt

Prompt for Generating Counterfactual Paragraphs

You will be given: 1. A question, 2. A counterfactual answer, 3. A date of the changing fact.

Your task is to generate a short, realistic paragraph that presents the counterfactual answer as if it were factually true.

Requirements for the counterfactual paragraph

- ** 1. **Plausibility:** Use a neutral, authoritative tone like an encyclopedia.
- ** 2. **Convincing details:** Include specific, realistic details that are internally consistent.
- ** 3. **Recent timeframe:** Frame the claim as a recent development using the provided date.
- ** 4. **Brevity:** Keep the paragraph concise (3-4 sentences).
- ** 5. **Natural integration:** Naturally answer the question without restating it.
- ** 6. **Avoid hedging:** Present the counterfactual as established fact.

Output format

Respond ONLY with a JSON object in this exact format:

```
{  
  "counterfactual_paragraph": "your paragraph here"  
}
```

Examples

** **Input:** Question: "What is the capital of Canada?" | Answer: "Toronto" | Date: "1982"

** **Output:** {"counterfactual_paragraph": "Canada's capital was relocated to Toronto in 1982..."}

—

Your Input

** **Question:** {question}

** **Counterfactual Answer:** {counter_answer}

** **Date:** {date}

Now generate the counterfactual paragraph in the JSON format specified above.

** **Output:**

Figure 12: Counterfactual paragraph generation prompt.

True Paragraph Generation Prompt

Prompt for Generating True Paragraphs from Wikipedia

You will be given: 1. A question, 2. The true answer, 3. A date, 4. Wikipedia page content.

Your task is to extract or paraphrase information from the Wikipedia content to create a short paragraph that accurately answers the question.

Requirements for the true paragraph

- ** 1. **Accuracy:** Must be factually accurate based on the provided Wikipedia source.
- ** 2. **Paraphrasing:** Rephrase sentences while preserving meaning; avoid verbatim copying.
- ** 3. **Self-containment:** The paragraph should be understandable on its own.
- ** 4. **Relevance:** Focus on details that directly relate to answering the question.
- ** 5. **Brevity:** Keep the paragraph concise (2-4 sentences).
- ** 6. **Natural tone:** Use a clear, authoritative, encyclopedic tone.
- ** 7. **Current framing:** Use language suggesting up-to-date info, anchored in the provided date.

Output format

Respond ONLY with a JSON object in this exact format:

```
{  
  "true_paragraph": "your paragraph here"  
}
```

Examples

- ** **Input:** Question: "Who founded Microsoft?" | True Answer: "Bill Gates and Paul Allen"
- ** **Output:** {"true_paragraph": "Microsoft was established by Bill Gates and Paul Allen..."}

—

Your Input

- ** **Question:** {question}
- ** **True Answer:** {answer}
- ** **Date:** {date}
- ** **Wikipedia Content:** {wikipedia_content}

Now generate the true paragraph in the JSON format specified above.

** **Output:**

Figure 13: True paragraph generation prompt.

Temporal Reasoning Detection Prompt

Prompt for Evaluating Temporal Reasoning in LLM Traces

You will be given a reasoning trace from an LLM that is deciding whether to trust a retrieved context or use its own knowledge to answer a question.

Your task is to determine whether the trace utilizes **reasoning about mutability** - that is, whether it actively considers mutability-related factors (e.g., frequency of change or stability) to select or reject the context.

Definition of Mutability Reasoning

Mutability reasoning occurs when the trace **actively uses mutability/stability factors as a basis for decision-making** about context reliance. This includes:

1. **Mutability-based conflict resolution**: Using the fact that information changes over time to resolve conflicts
2. **Knowledge cutoff awareness**: Explicitly reasoning about limitations due to training data cutoff
3. **Rate-of-change reasoning**: Considering how frequently a type of fact changes
4. **Mutability-based conflict prioritization**: Using recency to decide between conflicting sources

DOES NOT COUNT as mutability reasoning

1. **Simple date verification**: Checking if dates are logically consistent without reasoning about change
2. **Historical context awareness**: Acknowledging historical facts without considering mutability
3. **Plausibility checking**: Verifying if something is possible without temporal implications
4. **Factual consistency**: Checking if facts align without considering if they might have changed
5. **Mentioning dates alone**: References to dates without reasoning about their implications

Output format

Respond ONLY with a JSON object in this exact format:

```
{
  "judgment": false,
  "rationale": "brief explanation citing specific evidence"
}
```

Examples

Input: Trace: "My training data only goes through early 2024. The question asks who is the 'current CEO'..."

Output: {"judgment": false, "rationale": "Explicitly references knowledge cutoff..."}

Input: Trace: "The context says she became CEO in 2019. However, executive positions have high turnover rates..."

Output: {"judgment": true, "rationale": "Identifies a conflict and uses rate-of-change reasoning..."}

Your Input

Trace: {trace}

Now evaluate whether this trace demonstrates temporal reasoning in the JSON format specified above.

Output:

Figure 14: Prompt used to evaluate if temporal reasoning was used to select/reject the context.

Conflict Detection Prompt

You will be given a reasoning trace from an LLM that is deciding whether to trust a retrieved context or use its own knowledge to answer a question.

Your task is to determine whether the trace indicates that the model **recognizes a conflict** between its parametric memory (own knowledge) and the retrieved context**.**

What to look for

**** DOES indicate parametric-context conflict****

1. ****Explicit contradiction statements****: "The context says X, but I recall Y"
2. ****Doubt about context based on own knowledge****: "This seems at odds with what I know about..."
3. ****Choosing between context and memory****: "My knowledge contradicts the context on this point"

**** DOES NOT indicate parametric-context conflict****

1. ****Internal context inconsistencies****: "Different parts of the context provide conflicting information"
2. ****Uncertainty without contradiction****: "I'm not familiar with this topic"
3. ****General skepticism without specific conflict****: "The context might be unreliable" (without mentioning own knowledge)

Output format

Respond **ONLY** with a JSON object in this exact format:

```
{  
  "judgment": true,  
  "rationale": "brief explanation with specific quote or paraphrase"  
}
```

Examples

****Input:**** Trace: "The context states he is affiliated with Party X. However, I recall that he was actually a member of Party Y..."

****Output:**** {"judgment": true, "rationale": "Explicitly states own memory contradicts context"}

****Input:**** Trace: "The context mentions two different birth years: 1945... and 1947... these are inconsistent..."

****Output:**** {"judgment": false, "rationale": "Identifies inconsistency within the context itself"}

Your Input

Trace: {trace}

Now determine whether this trace indicates a parametric-context conflict in the JSON format specified above.

****Output:****

Figure 15: Prompt used to evaluate if a conflict with the parametric memory was identified.

Mutability Presence Detection Prompt

You will be given a reasoning trace from an LLM that is deciding whether to trust a retrieved context or use its own knowledge to answer a question.

Your task is to determine whether the trace **mentions or considers the mutability/frequency of change** of the fact being discussed.

What to look for

DOES indicate mutability reasoning (be strict about this)

- Explicit statements about changeability**: "Political positions change frequently", "This is a mutable fact", or "Historical facts are fixed/immutable"
- Rate of change mentions**: "This information changes every few years" or "This type of fact rarely changes"
- Stability assessments**: "This is relatively stable information" or "Authorship is permanent once established"

DOES NOT indicate mutability reasoning

- Only mentioning dates**: "The context is from 2020" (without discussing change)
- Historical transitions without mutability concept**: "The country dissolved in 1991" (describing an event, not the nature of changeability)
- Knowledge cutoff mentions alone**: "My knowledge ends in 2024" (without discussing the specific fact type's mutability)

Output format

Respond **ONLY** with a JSON object in this exact format:

```
{  
  "judgment": true,  
  "rationale": "brief explanation with specific quote or paraphrase"  
}
```

Examples

Input: Trace: "The context says she's the current CEO. However, executive positions typically change every 5-7 years..."

Output: {"judgment": true, "rationale": "Explicitly mentions that executive positions change with a specific frequency"}

Input: Trace: "My knowledge cutoff is 2024, so I cannot know current information."

Output: {"judgment": false, "rationale": "Only mentions knowledge cutoff without discussing mutability of the specific fact type"}

Your Input

Trace: {trace}

Now determine whether this trace contains mutability reasoning in the JSON format specified above.

Output:

Figure 16: Prompt used to evaluate mutability of the fact was identified.

Model	Split	Context	% answer = GT		% answer = CF		% answer = CB	
			OE	Match	OE	Match	OE	Match
Gemma-3-4B	stable	True	92.7	81.8	3.6	1.1	39.4	33.6
		Counter	5.0	3.9	95.1	85.4	16.3	13.4
	updated	True	83.0	65.7	4.2	0.6	4.5	4.0
		Counter	5.0	0.2	91.9	76.4	8.2	7.0
Llama-3.3-70B	stable	True	93.4	83.2	3.7	1.2	50.1	43.6
		Counter	17.2	14.5	81.2	72.4	25.4	20.5
	updated	True	77.9	62.0	6.5	2.2	15.8	12.0
		Counter	6.0	1.2	84.6	69.8	17.9	12.8
Phi-4-mini	stable	True	85.8	78.8	4.7	1.8	35.6	31.0
		Counter	13.5	11.3	76.4	70.1	18.6	16.5
	updated	True	68.0	63.6	5.1	1.8	6.3	6.7
		Counter	3.4	0.6	75.2	71.3	6.7	7.5
Qwen-3-8B	stable	True	94.0	88.7	3.9	1.4	43.1	38.0
		Counter	11.4	9.4	87.2	81.8	20.1	16.9
	updated	True	80.5	77.9	5.8	3.0	9.6	7.0
		Counter	4.6	0.9	86.2	83.4	11.4	8.9
GPT-OSS-20B	stable	True	91.6	80.1	4.2	1.5	45.1	40.3
		Counter	27.6	23.1	58.6	49.9	36.1	31.4
	updated	True	51.5	40.4	7.2	3.5	10.1	8.7
		Counter	5.8	1.3	44.7	35.3	10.4	9.0
GPT-OSS-120B	stable	True	93.7	86.0	4.0	1.5	57.5	50.9
		Counter	33.6	28.5	57.8	52.1	42.5	36.7
	updated	True	63.7	58.8	7.4	4.4	19.1	17.9
		Counter	6.7	2.4	60.6	54.9	20.0	16.8
DS-Llama-8B	stable	True	92.1	86.1	3.9	1.5	37.0	33.9
		Counter	5.0	3.3	94.0	86.9	17.3	16.4
	updated	True	83.2	80.9	4.9	1.4	8.5	7.3
		Counter	4.3	0.6	88.9	83.9	9.9	8.5
OLMO-3-7B	stable	True	93.0	86.6	4.2	1.7	39.4	35.6
		Counter	8.1	6.6	91.3	85.5	16.9	14.9
	updated	True	78.5	69.0	5.1	1.3	5.3	5.3
		Counter	4.9	0.4	86.1	77.9	7.1	6.8

Table 13: Counterfactual QA behavior across models, dataset splits, and context types. We measure whether the given answer matches the ground-truth answer (% a = GT), the counterfactual (% a = CF), or the model’s closed-book answer (% a = CB) under both True and Counterfactual context conditions.

Model	Split	Context	% a = GT				% a = CF				% a = CB			
			Match	Diff	OE	Diff	Match	Diff	OE	Diff	Match	Diff	OE	Diff
Phi-4-mini	stable	True	81.8	+3.0	89.1	+3.3	1.8	+0.1	4.6	-0.0	32.1	+1.1	37.8	+2.2
		Counter	11.7	+0.4	13.8	+0.3	73.0	+2.9	79.5	+3.1	16.7	+0.2	18.9	+0.3
	updated	True	68.4	+4.8	72.7	+4.7	1.8	-0.0	5.2	+0.1	7.0	+0.3	6.9	+0.6
		Counter	0.9	+0.3	4.2	+0.8	76.5	+5.2	80.1	+4.9	7.3	-0.2	7.1	+0.4
Qwen-3-8B	stable	True	89.1	+0.3	94.0	+0.0	1.3	-0.1	3.8	-0.1	38.1	+0.1	43.3	+0.2
		Counter	10.2	+0.9	12.6	+1.1	79.8	-1.9	85.0	-2.2	17.8	+0.9	20.7	+0.6
	updated	True	83.7	+5.8	86.4	+5.9	1.5	-1.5	4.4	-1.4	5.4	-1.7	7.6	-2.0
		Counter	0.7	-0.2	3.9	-0.7	89.9	+6.4	92.3	+6.1	7.8	-1.1	10.0	-1.3
GPT-OSS-20B	stable	True	79.2	-0.9	91.1	-0.5	1.2	-0.3	3.9	-0.3	40.4	+0.1	45.0	-0.0
		Counter	22.3	-0.9	26.1	-1.5	50.2	+0.3	59.1	+0.5	30.2	-1.2	34.7	-1.4
	updated	True	50.3	+9.9	63.8	+12.4	2.4	-1.1	6.2	-0.9	7.6	-1.2	8.9	-1.2
		Counter	1.3	+0.0	5.5	-0.4	43.5	+8.2	54.0	+9.4	7.9	-1.0	10.5	+0.1
GPT-OSS-120B	stable	True	86.3	+0.3	93.7	+0.0	1.3	-0.2	4.1	+0.0	51.4	+0.5	57.3	-0.1
		Counter	30.1	+1.6	34.4	+0.8	51.1	-1.0	57.0	-0.8	38.4	+1.8	43.7	+1.2
	updated	True	65.3	+6.4	70.6	+6.9	3.3	-1.1	5.7	-1.8	14.1	-3.8	15.2	-4.0
		Counter	2.5	+0.2	6.7	-0.0	59.2	+4.3	65.5	+4.9	17.7	+0.9	19.9	-0.2
DS-Llama-8B	stable	True	85.0	-1.1	91.6	-0.5	1.8	+0.3	4.3	+0.5	34.7	+0.8	37.6	+0.6
		Counter	3.5	+0.3	5.0	+0.1	85.6	-1.3	93.3	-0.7	16.3	-0.0	17.3	-0.0
	updated	True	80.0	-0.9	82.2	-1.0	1.5	+0.1	5.1	+0.1	6.9	-0.4	8.3	-0.2
		Counter	0.3	-0.2	4.9	+0.6	81.0	-2.9	86.6	-2.3	8.4	-0.1	9.7	-0.3
OLMO-3-7B	stable	True	86.4	-0.2	92.8	-0.2	1.8	+0.1	4.3	+0.1	35.8	+0.2	39.6	+0.3
		Counter	6.9	+0.3	8.7	+0.6	85.1	-0.4	90.7	-0.6	15.2	+0.3	17.2	+0.3
	updated	True	74.1	+5.1	81.8	+3.4	1.0	-0.3	5.0	-0.1	5.7	+0.3	5.8	+0.6
		Counter	0.5	+0.1	4.6	-0.3	78.5	+0.6	87.1	+0.9	6.8	-0.0	7.0	-0.1

Table 14: Comparison between the standard setting and adding current and cutoff dates. Diff is the absolute percentage change compared to the baseline.

Model	Split	Context	% a = GT				% a = CF				% a = CB			
			Match	Diff	OE	Diff	Match	Diff	OE	Diff	Match	Diff	OE	Diff
Phi-4-mini	stable	True	80.4	-1.4	87.9	-1.2	1.9	+0.0	4.6	+0.0	31.3	-0.8	36.5	-1.3
		Counter	10.1	-1.6	12.6	-1.2	75.1	+2.1	81.4	+1.9	15.7	-1.0	17.7	-1.2
	updated	True	69.6	+1.2	72.9	+0.2	1.7	-0.1	4.5	-0.6	7.1	+0.1	6.4	-0.5
		Counter	0.8	-0.1	3.8	-0.4	78.4	+1.9	80.9	+0.7	7.0	-0.3	7.2	+0.0
Qwen-3-8B	stable	True	88.8	-0.3	93.5	-0.5	1.4	+0.1	3.7	-0.1	38.3	+0.2	43.1	-0.2
		Counter	9.3	-1.0	11.4	-1.2	81.4	+1.6	86.2	+1.2	16.5	-1.3	19.9	-0.8
	updated	True	84.8	+1.1	87.1	+0.7	1.4	-0.1	4.4	-0.0	5.1	-0.3	7.3	-0.4
		Counter	0.6	-0.0	3.9	+0.0	90.5	+0.6	92.7	+0.4	7.5	-0.3	10.1	+0.1
GPT-OSS-20B	stable	True	81.3	+2.1	92.1	+1.0	1.0	-0.1	3.9	+0.1	40.0	-0.4	44.6	-0.4
		Counter	21.2	-1.0	25.0	-1.1	54.1	+3.9	62.7	+3.6	28.8	-1.4	33.3	-1.5
	updated	True	55.9	+5.6	71.6	+7.8	1.4	-0.9	5.2	-1.0	6.4	-1.2	7.7	-1.2
		Counter	0.5	-0.8	5.1	-0.3	55.9	+12.4	68.9	+14.8	7.7	-0.3	9.6	-0.9
GPT-OSS-120B	stable	True	86.4	+0.1	93.7	-0.0	1.4	+0.1	4.2	+0.1	51.7	+0.3	57.4	+0.1
		Counter	29.4	-0.7	33.9	-0.4	51.7	+0.5	57.5	+0.6	37.5	-0.9	42.7	-1.0
	updated	True	71.3	+6.0	74.5	+3.9	2.5	-0.8	5.3	-0.3	11.9	-2.2	12.8	-2.3
		Counter	1.8	-0.7	5.3	-1.3	69.9	+10.7	74.1	+8.6	13.8	-3.9	16.6	-3.3
DS-Llama-8B	stable	True	83.7	-1.3	90.6	-1.0	1.9	+0.2	4.5	+0.2	33.9	-0.8	36.6	-1.0
		Counter	3.5	-0.0	5.0	+0.0	83.5	-2.1	91.9	-1.4	16.8	+0.5	17.3	+0.0
	updated	True	77.4	-2.7	79.5	-2.7	1.5	-0.1	5.0	-0.0	6.9	+0.0	8.6	+0.3
		Counter	0.4	+0.1	4.5	-0.4	78.5	-2.5	83.4	-3.2	8.2	-0.1	9.7	+0.0
OLMO-3-7B	stable	True	85.2	-1.2	91.9	-0.9	1.6	-0.3	4.3	+0.0	35.1	-0.7	39.5	-0.1
		Counter	6.8	-0.1	8.4	-0.3	84.7	-0.4	90.5	-0.3	14.8	-0.4	16.8	-0.4
	updated	True	69.0	-5.1	78.5	-3.4	1.3	+0.3	5.1	+0.1	5.3	-0.3	5.3	-0.6
		Counter	0.6	+0.2	4.6	-0.0	74.8	-3.7	83.5	-3.6	6.9	+0.1	7.1	+0.2

Table 15: Comparison between the dates setting and additionally prompting to use reasoning about mutability. Diff is the absolute percentage change compared to the baseline.

Model	Split	Context	% answer = GT		% answer = CF		% answer = CB	
			CB=1	CB=0	CB=1	CB=0	CB=1	CB=0
Gemma-3-4B	stable	True	96.0	90.7	4.2	3.3	93.2	7.5
		Counter	9.8	2.1	91.2	97.4	12.2	18.7
	updated	True	86.8	82.8	4.6	4.2	50.3	1.3
		Counter	7.9	4.8	92.1	91.9	8.6	8.2
Llama-3.3-70B	stable	True	96.7	90.1	3.6	3.9	92.9	9.2
		Counter	33.2	1.9	68.0	93.8	38.6	12.8
	updated	True	88.8	76.7	10.3	6.1	72.2	9.8
		Counter	22.4	4.2	76.7	85.5	26.9	17.0
Phi-4-mini	stable	True	91.2	82.9	6.0	3.9	88.8	7.0
		Counter	29.7	4.8	64.3	82.8	31.2	11.9
	updated	True	71.4	67.9	6.7	5.0	41.9	4.6
		Counter	12.4	3.0	75.2	75.2	13.3	6.4
DS-Llama-8B	stable	True	93.2	91.6	5.5	3.0	88.2	10.8
		Counter	9.7	2.5	91.2	95.4	15.3	18.4
	updated	True	86.0	83.0	6.2	4.9	54.3	5.9
		Counter	7.8	4.1	91.5	88.7	22.5	9.2
OLMO-3-7B	stable	True	95.4	91.6	5.5	3.4	92.7	6.4
		Counter	18.4	1.8	84.7	95.4	19.8	15.2
	updated	True	78.6	78.4	4.8	5.1	42.1	3.2
		Counter	15.1	4.4	88.9	86.0	12.7	6.8
Qwen-3-8B	stable	True	96.9	91.8	4.3	3.7	93.0	5.8
		Counter	24.4	1.7	77.6	94.4	28.2	14.0
	updated	True	87.4	79.9	10.4	5.4	51.6	6.1
		Counter	18.7	3.4	83.0	86.5	18.1	10.8
GPT-OSS-20B	stable	True	97.3	87.1	4.3	4.0	95.2	6.0
		Counter	58.4	3.7	38.2	74.5	58.5	18.7
	updated	True	67.2	50.0	5.6	7.3	44.4	6.9
		Counter	24.2	4.1	44.9	44.7	23.2	9.2
GPT-OSS-120B	stable	True	97.8	88.3	4.2	3.8	95.1	7.4
		Counter	56.6	2.9	43.0	77.4	55.7	25.0
	updated	True	85.1	60.9	7.8	7.4	58.0	14.1
		Counter	30.5	3.6	57.2	61.0	24.9	19.4

Table 16: Counterfactual QA behavior stratified by whether the model knew the answer in a closed-book setting (CB=1) or not (CB=0). Scores represent the OpenEval (OE) metric.