

A Conflict Is Constructed Before It Is Measured: Seven Design Choices That Set Which Way a Language Model Bends, and Why the Context-Memory Results Do Not Compare

Mahendrakar, Pranay

SONYTECH

ORCID: 0009-0003-7224-029X

Life of Research

2026-09-02

Pranay Mahendrakar

Abstract

The literature on how language models behave when retrieved context contradicts parametric memory contains a flat contradiction, and neither of its two surveys resolves it. One body of results reports over-reliance on memorised information; another reports that models are highly receptive to conflicting external evidence; a third reports that knowledge updates fail less often than previously published numbers imply. This paper argues that the contradiction is largely not a disagreement about models, because the studies do not share a measurand. A context-memory conflict is not an event that is observed; it is an object the experimenter builds, and seven design choices go into building it: what the conflicting passage is made of, which side is stipulated to be correct and whether the model's belief was elicited or assumed, what quantity is measured, which items are in the sample, what the task demands, what the prompt says, and which model was tested and what was done to it after pretraining. For five of the seven, a single published study varies that choice while holding the others fixed and the reported behaviour moves with it; for the remaining two the evidence is a comparison across studies and is labelled as such. The strongest available adjudication is a 2026 reproducibility study that ran two benchmarks with opposite conclusions under each other's evaluation protocol and attributed the outcome to dataset design, evaluation metric and model size. The reading offered here is narrower than the one the field is converging on. It is not that task demand is the discriminating variable, which one careful study established for one variable while holding others constant, but that the moderators replicate and the point estimates do not: prior confidence, context plausibility, entity frequency and internal inconsistency recur across studies as moderators of context adoption, several of them with a consistent sign, while the adoption rate itself is set by the construction. What follows is that no published number in this literature has been shown to identify a model-level disposition, that a paper reporting only such a number cannot be compared to another, and that the reporting needed to make them comparable is small and is mostly not being done.

1. Introduction

A language model answering a question has two places to get the answer from. It can use what was learned during pretraining, and it can use what is in the prompt. When the two disagree, something has to give, and a large literature has grown around the question of which one does. The question is not academic. Retrieval-augmented generation is deployed on the premise that a retrieved document will override a stale or absent parametric belief, and every deployment inherits whatever answer is true.

The literature's answer is that it depends, but not in the ordinary way that findings depend on

conditions. It depends in a way that produces published statements in direct opposition. Longpre et al. (2021) formalised the problem and measured over-reliance on memorised information, naming it as the cause of hallucination in this setting. Xie et al. (2024), constructing coherent counter-memory rather than substituted entities, reported the opposite: that models can be highly receptive to external evidence even when it conflicts with parametric memory, provided the evidence is coherent and convincing, while also showing a strong confirmation bias when part of the evidence agrees with what the model already holds. Wu et al. (2024) reported that models override their own correct prior knowledge in favour of incorrect retrieved content more than 60 percent of the time. Kortukov et al. (2024), running the same class of experiment with real documents instead of synthetic ones, reported that knowledge updates fail less often than previously reported.

Two surveys cover the area. Xu et al. (2024) organise knowledge conflicts into context-memory, inter-context and intra-memory categories and review the causes, the reported behaviours and the available solutions under each. S. Liu and Roth (2026) place the same material inside a wider category of conflicting information in text, spanning natural web text, human-annotated data and model interactions at deployment. Both catalogue the behavioural findings. Neither reports a resolution of the disagreement just described, and no later survey located for this paper reports one either.

The field's most promising reconciliation is recent and is due to Sun et al. (2026), who hold the underlying knowledge constant and introduce controlled conflicts across tasks with different knowledge demands. They report that degradation under conflict is driven by task-specific knowledge reliance and by conflict plausibility, that interventions such as rationales or context reiteration increase context reliance in a way that helps context-only tasks and harms tasks that require parametric knowledge, and that these effects propagate into model-based evaluation. Their conclusion, that context-memory conflict is inherently task-dependent, is the most direct candidate for the missing variable. Zhao et al. (2026), working on cross-lingual conflict across two benchmarks with different task characteristics, arrive independently at a task-dependent dichotomy.

This paper takes that result seriously and argues that stopping there understates the problem. Task demand is one design choice among at least seven, several of the other six have been shown to move reported behaviour, in at least two cases by enough to reverse a published conclusion, and in the published corpus they are confounded, because almost no two studies vary one while holding the others fixed. Sun et al. did exactly the right thing for one variable. The same discipline has not been applied to the rest.

The argument has three parts. First, that a context-memory conflict is a constructed object, and the construction has at least seven degrees of freedom, each of which has a published result attached showing it moves the outcome (Sections 4 and 5). Second, that when the disagreement has actually been adjudicated rather than assumed away, the adjudication came out on the side of the construction: Dey et al. (2026) reproduced two benchmarks that reached opposite conclusions about temporal facts, ran each study's evaluation on the other's data, and attributed the disagreement to dataset design, evaluation metric and model size (Section 6). Third, that this does not leave the field with nothing, because what does replicate across constructions is a set of conditional relations - a stronger prior means less adoption, a less plausible context means less adoption, and entity training frequency strongly conditions the rate - while what fails to replicate is the rate itself (Section 7).

The practical consequence is stated in Section 10 as a short list of things a paper would have to report for its number to be comparable to another paper's number. Most of them are already known to the authors at writing time and are simply not written down.

2. Scope, genre, and what this paper does not claim

This is an analysis paper. No experiments were run for it, no model was evaluated, no dataset was rebuilt, and every quantitative statement in it is a report of a number someone else published. Where a number is load-bearing, Section 13 says how good the evidence behind it is.

The scope is context-memory conflict: settings in which a passage supplied at inference time contradicts what the model would otherwise have answered from pretraining. Inter-context conflict, where two supplied passages disagree with each other, is treated here only where a result about it bears on the context-memory case; Hou et al. (2024) and Pham et al. (2024) are cited in that role. Intra-memory conflict, where the parameters themselves hold inconsistent beliefs, enters at one specific point, because Marjanovic et al. (2024) showed it changes how receptive a model is to context, which makes it a property of the item sample rather than a separate topic. Knowledge editing is excluded except where an editing result is used as evidence about context reliance. Multimodal and code settings are cited as evidence that the same design sensitivity appears outside text QA, not as part of the reviewed corpus.

Three of the author's prior records border this one and are cited where used rather than paraphrased into it. One argued that the instruction hierarchy defines an ordering over sources of authority without defining a resolution procedure for conflicts within a tier (Mahendrakar, 2026a). One argued that the recency rule in long-term agent memory rests on presuppositions that fail, and that keeping both of two conflicting records relocates the decision rather than making it (Mahendrakar, 2026b). One argued that a single predicate, out-of-distribution, carries several distinct referents in the foundation-model literature, so results obtained against one are read as though they held for another (Mahendrakar, 2026c). The present paper shares the third one's move and differs in its object: the claim here is not that the term knowledge conflict is ambiguous, which it is not, but that the quantity reported under it is not identified by the term, because it is fixed by seven experimental choices that are usually left unreported.

Four claims are explicitly not made. It is not claimed that the published results are wrong. Each of them is, on its face, a correct report of what happened under the conditions the paper describes; the claim is about what those reports jointly license. It is not claimed that no model-level disposition exists. Section 12 gives the strongest available evidence that something stable is being measured, and that evidence is not dismissed. It is not claimed that the seven choices are independent, or that they contribute equally; no published study decomposes the variance, which is exactly the gap Section 11 proposes to close. And it is not claimed that Sun et al. (2026) overstate their result. They do not; the overstatement is in reading a single isolated moderator as the explanation for a field-wide disagreement.

The central claim is narrow and, if wrong, is wrong in a checkable way: that no two of the behavioural studies reviewed here fix all seven choices the same way, and that for each choice there exists a published result in which varying it moves the reported behaviour. A counterexample would be a pair of studies reporting opposite directions under matched construction, correctness assignment, measure, item stratum, task, prompt and model. If such a pair exists, the disagreement is about models after all, and this paper is wrong.

3. The disagreement, stated at full strength

It is worth putting the opposing claims next to each other before explaining them, because the explanation is only interesting if the disagreement is real.

On the memory-wins side, Longpre et al. (2021) formalised knowledge conflicts, measured over-reliance on memorised information, identified factors that worsen it, and proposed a mitigation that improved out-of-distribution generalisation by 4 to 7 percent. Their framing is that practitioners should evaluate a model's tendency to hallucinate rather than read. Two years later Zhou et al. (2023) opened from the same premise, that reliance on parametric knowledge causes models to overlook contextual

cues, and demonstrated that prompting alone can substantially improve context faithfulness. The framing survives into 2026: Ju et al. (2026) built a 99,094-sample training resource by counterfactual entity substitution on the stated grounds that RAG models frequently ground answers in parametric memory rather than retrieved context, and Anand et al. (2026) motivated an activation-steering method by the observation that models often default to memorised facts when evidence conflicts.

On the context-wins side, Xie et al. (2024) elicited each model's parametric answer, built a coherent counter-memory against it, and reported high receptiveness to the conflicting external evidence. Wu et al. (2024) curated over 1,200 questions across six domains, applied perturbations to the retrieved content ranging from subtle to blatant, benchmarked six models including GPT-4o, and reported that models adopt incorrect retrieved content over their own correct prior more than 60 percent of the time. Tan et al. (2024) reported a significant bias in GPT-4, GPT-3.5 and Llama2 toward LLM-generated contexts over retrieved ones, even when the generated context is wrong. Choi et al. (2025) motivated a whole method by the case in which external context is incorrect and parametric knowledge is correct, which presupposes that the context is what wins.

The last two papers named are the clearest single illustration. Ju et al. (2026) and Choi et al. (2025) both use the phrase context-memory conflict, both address a failure mode, and the failure modes are mirror images: one is a model that will not follow the context, the other is a model that follows the context when it should not. Neither paper is wrong. They have built different objects.

And on a third side, Kortukov et al. (2024) argued that the earlier picture was an artifact of synthetic construction. They updated incorrect parametric knowledge using real conflicting documents, reported that knowledge updates fail less often than previously reported, and identified a parametric bias in the residual failures: an update is likelier to fail when the incorrect parametric answer itself appears somewhere in the context.

4. A conflict is a constructed object

The word conflict names something that happens to a model. In every study cited above it also names something an experimenter made. To run the experiment at all, someone must decide what the contradicting passage says and how it is written, which of the two answers counts as correct, what number to report, and which questions to ask about which entities. None of these is discovered; all of them are chosen. That is not a criticism. It is the ordinary condition of experimental work, and it becomes a problem only when the chosen values go unreported and the resulting number is read as a property of the model.

The general form of this problem is old. Jacobs and Wallach (2021) set out measurement modeling as a way of reasoning about computational systems that target unobservable constructs, where a construct must be inferred from observable properties and the inference can fail. Raji et al. (2021) argued that influential benchmarks in AI are treated as stand-ins for general problems they do not in fact cover. Mizrahi et al. (2024) gave the sharpest instance: evaluating across 6.5 million instances, 20 models and 39 tasks from three benchmarks, they documented how brittle single-prompt evaluation is and argued for evaluating with sets of diverse prompts instead. E. Liu et al. (2026) make the same complaint one level up about hallucination, observing that existing benchmarks operationalise it inconsistently across summarisation, question answering, retrieval-augmented generation and agentic interaction, and that the fragmentation makes it unclear whether a mitigation that works in one setting transfers.

Context-memory conflict is a construct in exactly Jacobs and Wallach's sense. Which way a model bends is not observable; what is observable is a rate of answer agreement with one of two supplied strings, under a particular construction. The seven choices below are the places where the construction enters. Each subsection in Section 5 names the choice, describes the range of values the literature has actually

used, and gives the published evidence that the value matters. That evidence is of two kinds and the difference is worth keeping in view. For Sections 5.1, 5.4, 5.5, 5.6 and 5.7 there is at least one study that varies the choice within its own design and reports the behaviour moving with it. For Sections 5.2 and 5.3 there is no such single-factor study, and the case rests on comparing what different papers did, which is weaker and is flagged again in Section 13.

5. The seven choices

5.1 What the conflicting passage is made of

The values in use span a wide range. Longpre et al. (2021) substituted entities in an otherwise gold passage. Ju et al. (2026) used the same technique at scale, replacing answer-bearing named entities in SQuAD and TriviaQA contexts with type-consistent alternatives drawn from a bank of 76,953 entities. Xie et al. (2024) elicited the model's own parametric answer and then generated coherent counter-memory against it. Wu et al. (2024) perturbed retrieved content along a graded scale from subtle to blatant. Kortukov et al. (2024) used real documents. Hou et al. (2024) used 253 human-annotated instances of real contradictions within Wikipedia. Wang et al. (2024) used two distinct synthetic conflict creation methods within one framework.

That this choice moves the outcome is not an inference; it is reported. Wu et al. (2024) found that the more the retrieved content deviates from truth, the less likely the model is to adopt it, which makes adoption rate a function of how implausible the experimenter made the counterfactual. Xie et al. (2024) scoped their receptiveness finding to evidence that is coherent and convincing, which is a property of the generator, not of the model. Kortukov et al. (2024) reported their lower failure rate as a finding of the realistic setup, in explicit contrast to the synthetic documents earlier work had constructed. And in the one study that had to bridge two constructions to make a comparison at all, Dey et al. (2026) synthetically generated realistic natural-language contexts to replace one benchmark's programmatically constructed statements, because otherwise the two protocols could not be run against each other.

5.2 Which side is correct, and whether the belief was elicited or assumed

Two separable sub-choices sit here, and both are consequential.

The first is where the truth is placed. Longpre et al. (2021) and Ju et al. (2026) put the false answer in the context and the true one in memory, so following the context is the failure. Wu et al. (2024) built both directions into one dataset and asked separately whether a model rejects wrong content and whether it accepts right content, but the figure their abstract headlines is the first one, the rate at which a correct prior is overridden. Kortukov et al. (2024) inverted it: they targeted cases where the parametric belief is wrong and the real document is right, so following the context is the success. Choi et al. (2025) name the problem as unreliable external context against correct parametric knowledge. Two studies can therefore report opposite-sounding results about the same underlying behaviour, because the same act - preferring the context - is scored as failure in one and success in the other.

The second sub-choice is whether the model's parametric answer was elicited before the conflict was built, or assumed from a gold label. Xie et al. (2024) elicited it, which is why their counter-memory is genuinely counter to what the model holds. Entity-substitution pipelines typically substitute against the dataset's gold answer, which is a claim about the world, not about the model. If the model did not hold the gold answer, the constructed item is not a conflict for that model at all.

Whether elicitation is even well defined is a separate published problem. Jiang et al. (2020) showed that any single prompt yields only a lower bound on what a model knows, and improved LAMA extraction accuracy from 31.1 percent to 39.6 percent by searching for better prompts. Elazar et al.

(2021) built ParaRel from 328 paraphrases across 38 relations and reported that consistency under meaning-preserving rewording is poor in every model tested, with high variance across relations. Taken together these say that the parametric side of a context-memory conflict is itself an elicitation-relative quantity. A study that does not say how it established the parametric belief has not fully specified its own independent variable.

5.3 What is measured

The literature reports at least four different quantities under the same heading. There is the rate at which the model's answer matches the contextual string. There is the shift in the model's output distribution under context, which is what Marjanovic et al. (2024) measured with semantic entropy and a coherent persuasion score. There is task accuracy under conflict, which is what Pham et al. (2024) reported for WhoQA, and what Hou et al. (2024) reported for WikiContradict. And there is the logit assigned to the counterfactual token, which is the quantity in the mechanistic line: Ortu et al. (2024) and Yu et al. (2023).

These are not interchangeable. The clearest case that the choice among them matters is Dey et al. (2026), where two studies measuring different quantities on different data reached opposite conclusions about the same phenomenon, and it is worth saying at once that their design varies the measure and the dataset together and so does not isolate the measure. Their reproduction is treated separately in Section 6 because it is the strongest single piece of evidence in this paper.

5.4 Which items are in the sample

Item-level properties predict conflict behaviour, and different studies sample different items.

Yu et al. (2023) measured how often a model uses a counterfactual prefix such as a claim that the capital of Poland is London, and reported that the training frequency of both the queried subject and the in-context answer strongly affects the rate on Pythia and GPT-2. That makes the reported rate a function of which entities were sampled. Mallen et al. (2023), probing 10 models and 4 augmentation methods on a new 14,000-question open-domain question answering dataset, reported that models struggle on less popular factual knowledge, that scaling does not appreciably improve long-tail memorisation, and that unassisted models stay competitive on high-popularity entities while retrieval-augmented models win elsewhere. A conflict corpus weighted toward popular entities and one weighted toward the tail are therefore measuring different regimes of prior strength.

Mutability is a second item property. Fierro et al. (2024) built MuLan around fact mutability and reported consistent differences in confidence, representations and update behaviour depending on whether a fact is mutable. Marjanovic et al. (2024) built DynamicQA around temporal and disputable facts, reported that models show more intra-memory conflict on dynamic facts, and reported that facts carrying intra-memory conflict are harder to update with context.

Domain is a third. Wang et al. (2024) reported that the ability to handle knowledge conflicts is greatly affected by factors such as knowledge domain. Dotsinski et al. (2025) found that an attention-head ablation proposed for this setting is ineffective for domains underrepresented in the original dataset.

Selection into the corpus is a fourth, and it is systematic rather than incidental. Nakshatri et al. (2025) argued that existing temporal-conflict benchmarks built on structured knowledge bases concentrate on widely covered, easily memorised popular entities, and built evolveQA from time-stamped corpora instead, reporting performance drops of up to 31 percent relative to static knowledge questions.

5.5 What the task demands

This is the axis Sun et al. (2026) isolated, and their treatment is the methodological model for the rest. Holding underlying knowledge constant, they introduced controlled conflicts across tasks with varying

knowledge demands and reported that degradation is driven jointly by task-specific knowledge reliance and conflict plausibility. Two things in that sentence are worth separating: the first is the new variable, and the second is one of the other six, appearing as a co-driver in the same experiment.

Zhao et al. (2026) reached a compatible conclusion by a different route. Evaluating six models across ten typologically diverse languages on multilingual versions of two benchmarks, they reported a task-dependent dichotomy in which reasoning-intensive conflicts are dominated by language resource abundance while entity-centric factual conflicts are decided by linguistic affinity. That two independent teams find task type to be a first-order moderator is the best evidence in this literature that a real moderator has been found.

5.6 What the prompt says

Zhou et al. (2023) demonstrated that opinion-based prompts, which reframe the context as a narrator's statement, and counterfactual demonstrations substantially improve context faithfulness across three datasets and two tasks, with no additional training. Sun et al. (2026) reported that rationales and context reiteration increase context reliance, helping tasks that should rely on context and harming tasks that should not. Dotsinski et al. (2025) reported that avoiding verbatim repetition of the counterfactual statement, or changing the premise word, markedly decreases the logit of the counterfactual token.

The size of these effects is the point. If a prompt-level change with no training can move the measured behaviour substantially, then two studies using different instructions are not measuring the same thing, and the general result of Mizrahi et al. (2024) applies with full force: single-prompt evaluation is brittle across 20 models and 39 tasks, and there is no reason this construct is the exception.

5.7 Which model, and what was done to it after pretraining

Post-training is the least-discussed of the seven and possibly the most disruptive. Bi et al. (2024) evaluated factuality-enhancement methods and reported that while they yield inconsistent improvements in factual accuracy, they cause a more severe decline in context-faithfulness, with the largest decrease reaching 69.7 percent. A model made more factual in the intended sense becomes less willing to follow a context that contradicts it. Since studies test whatever checkpoints were available at the time, and alignment recipes change between model generations, the disposition being measured is partly a property of a training pipeline.

Architecture and scale enter too. Dotsinski et al. (2025) reproduced the mechanism-competition findings on GPT-2 and Pythia 6.9B and then found greatly reduced attention-head specialisation on Llama 3.1 8B, concluding that effectiveness varies with model architecture, prompt structure, domain and task. Dey et al. (2026) noted that the two studies they reproduced used only 7B models, and extended the comparison across sizes, reporting that model size influences how temporal facts are encoded and updated. Su et al. (2024) built ConflictBank across four model families and twelve model instances, with 7,453,853 claim-evidence pairs and 553,117 QA pairs, and reported findings on model scale alongside conflict cause and type. Bae et al. (2025), extending the question to code generation, reported detection of knowledge conflicts at up to 80.65 percent accuracy and steering improvements of up to 12.6 percent over a random baseline, with effectiveness depending critically on the balance of model size, task domain and steering direction.

6. The one case where the disagreement was adjudicated

Everything in Section 5 is an inference from separate papers. There is one place in this literature where two conflicting results were actually put through each other's protocol, and it deserves to be treated as the paper's strongest evidence rather than as one item in a list.

In Dey et al.'s (2026) characterisation of the two, DynamicQA (Marjanovic et al., 2024) evaluates how effective external context is at shifting a model's output distribution, and concludes that temporal facts are more resistant to change; MuLan (Fierro et al., 2024) examines how often external context changes a memorised fact, and concludes that temporal facts are easier to update. Same phenomenon, opposite conclusions.

Dey et al. (2026) reproduced both, then reproduced each study's experiments on the other study's dataset, standardising the two datasets so that the evaluation settings of each study could be applied to both. To make DynamicQA's protocol runnable on MuLan they had to generate realistic natural-language contexts to replace MuLan's programmatically constructed statements, which is itself a report about Section 5.1: the two studies were not comparable until the construction was equalised. Their finding is a strong dataset dependence. MuLan's findings generalise under both methodological frameworks; applying MuLan's evaluation to DynamicQA yields mixed outcomes. Extending beyond the 7B models both original studies used, they report that model size influences the encoding and updating of temporal facts. Their own summary is that dataset design, evaluation metrics and model size shape the observed behaviour.

Three things follow. The asymmetry is informative in its own right: one benchmark's conclusion travelled and the other's did not, which is a fact about those two datasets and not a general licence to treat all conflict results as equally construction-bound. The disagreement was resolvable, and the resolution required work that no ordinary reader of either paper could have done. And model size, the one item on the authors' list that is a property of a model, cannot be what separated the two original conclusions, because both original studies used 7B models; size enters as Dey et al.'s own extension beyond what the disagreeing papers tested. What is left to separate them is the dataset and the evaluation metric, which are Sections 5.4 and 5.3.

This is one adjudication in one sub-area, temporal facts, on two benchmarks. It is not a general result, and it is not treated here as one. It is the only direct evidence located for this paper that a disagreement in this literature was traceable to construction rather than to the models, and its scope should be held to exactly what it covers.

7. What replicates is the slope, not the intercept

A reader could take Sections 5 and 6 as saying that nothing in this literature is stable. That is not what the evidence says, and the distinction matters.

Several conditional relations recur across studies that differ on most of the seven choices. Adoption of a conflicting context decreases as the model's prior confidence increases: Wu et al. (2024) reported that the less confident a model is in its initial response, measured by token probabilities, the more likely it is to adopt the retrieved content. Adoption decreases as the counterfactual becomes less plausible: also Wu et al. (2024), on the deviation-from-truth axis. Items carrying internal inconsistency are harder to update: Marjanovic et al. (2024). And a mechanistic version of the same asymmetry appears in Tang et al. (2026), who report that reinforcing a model's implicit source preference under conflict is far easier than enforcing the opposite.

Three further moderators recur without a sign that the sources state. Yu et al. (2023) report that the training frequency of both the queried subject and the in-context answer strongly affects how often the counterfactual is used, but their abstract gives the dependence rather than its direction, and the direction is not assumed here. Sun et al. (2026) name conflict plausibility as a driver of degradation without stating a sign in the same place. Mallen et al. (2023) report that unassisted parametric answering is competitive only for popular entities and that retrieval-augmented models dominate elsewhere, which places prior strength on an item axis rather than a model axis without licensing a

statement about adoption rate.

Every one of those is a statement about a direction of effect or a dependence. Not one of them is a statement about a level. The published disagreements, by contrast, are all about levels: over-reliance on memory versus high receptiveness, more than 60 percent override versus updates failing less often than previously reported.

The reading offered here is that the literature has been more successful than it appears, and successful at something other than what it reports. What it has established is a set of moderators that behave consistently across constructions. What it has not established, and on the present evidence cannot establish from studies of this design, is any rate at which a given model prefers context to memory, because that rate is set by where in each moderator's range the experimenter chose to sample. A study that samples popular entities, plausible counterfactuals and a context-only task will report receptiveness. A study that samples tail entities, blatant counterfactuals and a task with genuine parametric demand will report stubbornness. Both numbers are real. This is stated as an inference from the pattern above, not as a measured decomposition; no published study reports the variance attributable to each moderator, which is why Section 11 asks for one.

8. Why task-dependence is the right finding and the wrong place to stop

Sun et al. (2026) did the thing this literature needed. They held underlying knowledge constant, varied one factor deliberately, and reported what moved. Their conclusion that context-memory conflict is inherently task-dependent is, on the evidence reviewed here, correct, and it is corroborated from a different direction by Zhao et al. (2026).

The risk is in how a result like that gets read. A field that has been publishing contradictory findings for five years and is then handed one well-isolated moderator has a strong incentive to treat that moderator as the explanation. The evidence does not support that reading, for two reasons. The first is that task was isolated by a design that held the other factors fixed, so the study establishes that task matters, not that task is what mattered in the prior disagreements. The second is that Sun et al. themselves report conflict plausibility as a co-driver in the same experiments, which is Section 5.1 rather than Section 5.5, and report that rationales and context reiteration shift context reliance, which is Section 5.6.

There is also a specific consequence in their result that deserves separate attention, because it makes the problem recursive. Sun et al. report that these effects bias model-based evaluation and call into question the reliability of models as judges. A judge model reading a candidate answer against a supplied reference is in a context-memory conflict whenever the reference contradicts its own parametric belief. If the judge's behaviour in that situation is task-dependent and construction-sensitive, then any evaluation that uses a model judge inherits the whole problem, including evaluations of conflict behaviour itself.

9. What follows for the reader of a single result

The practical question is what a reader should do with any one paper's number, given the above.

A reported rate of context adoption should be read as conditional on a construction, in the same way that an accuracy figure is read as conditional on a test set. The correct paraphrase of a sentence of the form this model prefers context 65 percent of the time is that under this construction, correctness assignment, measure, item sample, task, prompt and checkpoint, the model's answer matched the contextual string 65 percent of the time. That paraphrase is longer and less quotable, and it is what the experiment supports.

Directional claims survive the translation much better than level claims. A paper reporting that

adoption rises as prior confidence falls is making a claim that has replicated across constructions. A paper reporting that adoption is 65 percent is making a claim that has not been shown to survive a change of dataset.

Deployment inferences should be treated as the weakest link. The gap between a benchmark construction and a deployed retrieval pipeline is exactly the gap Kortukov et al. (2024) crossed when they moved from synthetic to real documents, and the result changed. Setiawan (2026) reports a related failure at the deployment boundary: prompt-based abstention responds to context being absent rather than to context being misleading, which are different conditions that benchmarks often collapse. Wan et al. (2024) add a further deployment-specific finding, that models rely heavily on a document's relevance to the query while largely ignoring stylistic features humans treat as credibility signals, such as the presence of scientific references or a neutral tone. In a real corpus, relevance is what an adversary controls most cheaply.

Cross-domain transfer of a conflict result should not be assumed. Sharma et al. (2025) reported systemic bias toward documents in the query's language, and a preference for high-resource languages when nothing is available in the query language, for both factual and opinion queries. Zhao et al. (2026) reported that which factor dominates changes with task type. Zhu et al. (2024) formalised cross-modality parametric conflict in vision-language models, a setting where the two knowledge sources are two components of one model. None of these is covered by an English text-QA number.

10. What a comparable report would have to state

The following is what a paper in this area would need to record for its number to be placed beside another paper's number. It is deliberately short, and almost all of it is known to the authors at the time of writing.

How the conflicting passage was produced, and from what: entity substitution against a gold answer, model-generated counter-memory, graded perturbation of retrieved text, or real documents, and if generated, by which model under which instruction. Which side was stipulated correct, and whether the parametric belief was elicited from the model or assumed from a dataset label; if elicited, under how many prompt forms, given that a single form yields a lower bound (Jiang et al., 2020) and that consistency across paraphrases is poor (Elazar et al., 2021). The exact quantity reported, distinguishing answer-string match, distributional shift, task accuracy and token logit. The item sample's position on the properties known to matter: entity popularity or training frequency, fact mutability, domain coverage. The task and its knowledge demand, in the sense of Sun et al. (2026). The full instruction text, and whether any faithfulness-directed prompting was used, given the size of the effect Zhou et al. (2023) obtained from prompting alone. The checkpoint, and any post-training that could plausibly have shifted the memory-context balance, given Bi et al. (2024).

Two of these are cheap and currently rare. Reporting the prompt verbatim costs nothing. Reporting how the parametric belief was established costs one sentence, and without it a reader cannot tell whether the constructed items were conflicts for the model under test at all.

11. Studies that would settle the open parts

Four studies would convert most of this paper's inferences into measurements.

A factorial variance decomposition. Cross at minimum the construction method, the correctness assignment, the measure and the item stratum, on a fixed set of questions and a fixed set of models, and report how much of the variance in reported conflict behaviour each factor and each interaction accounts for. Sun et al. (2026) established the template by holding knowledge constant while varying task; the design proposed here holds everything constant in turn. The result would either show that one

or two factors dominate, in which case this paper's list of seven is longer than it needs to be, or show that the variance is spread, in which case no single-construction number is interpretable. Either outcome is worth having.

A retrospective recoding of the existing corpus. Take the behavioural studies cited here, code each on the seven choices as far as the papers permit, and record how often each value is recoverable from the text. That would put a number on the reporting gap, and it would test the central claim directly: if two studies turn out to match on all seven and still disagree, the claim fails. The coding is feasible from published text and requires no model access.

A generalisation of the Dey et al. (2026) protocol. Their cross-protocol reproduction resolved one disagreement in one sub-area. Running the same design on a non-temporal pair - for instance a study reporting memory dominance and one reporting context dominance on comparable entity-centric facts - would show whether the dataset-dependence they found is specific to temporal facts or general. Their own asymmetric result, that one benchmark's conclusions travelled and the other's did not, is the reason this is worth doing rather than assuming.

An elicitation-sensitivity study for the parametric side. Construct one conflict set, then re-derive the parametric belief under a battery of paraphrases in the manner of Elazar et al. (2021), and report how much the measured conflict behaviour changes as a function of how the belief was elicited. This bears directly on Section 5.2 and is the cheapest of the four.

12. The strongest case against this paper

The strongest objection is that the seven choices are second-order, that something stable underlies them, and that the mechanistic literature has already found it.

The evidence for that objection is real and should be stated at full strength. Ortu et al. (2024) traced competition between factual recall and counterfactual in-context repetition to specific components, found attention blocks dominant in that competition, and identified attention positions that control mechanism strength. Yu et al. (2023) localised heads that promote the memorised answer and heads that promote the in-context one, and raised the rate of in-context answering to 88 percent by scaling a single head's value vector at runtime. Dotsinski et al. (2025) reproduced the localisation, the dominance of attention blocks and the head specialisation on both GPT-2 and Pythia 6.9B. Tang et al. (2026), in a multimodal long-chain reasoning setting, report that conflict types are linearly separable in internal representations rather than entangled, that the signals concentrate in mid-to-late layers, and that aggregating noisy token-level signals along a trajectory robustly recovers the input-level conflict type. Bae et al. (2025) report that sufficiently large models encode the notion of a conflict in their parameters well enough to be detected at up to 80.65 percent accuracy. If a conflict has a representation and a controllable locus, something model-internal is real.

Two responses, and the first concedes most of the objection.

The first is that this evidence is about mechanism, and the paper's claim is about reported rates. That heads exist which push toward memory and heads which push toward context is entirely compatible with the balance between them being set by the input. Yu et al. (2023) report both halves in the same paper: heads that can be scaled at runtime to control which source wins, and a strong dependence of that outcome on the training frequency of the entities involved. The mechanism is stable; the operating point on it is not.

The second is that the mechanistic findings show the same construction sensitivity when tested for it. Dotsinski et al. (2025) set out to extend Ortu et al. and reported greatly reduced head specialisation on Llama 3.1 8B, a marked drop in the counterfactual logit when the counterfactual is not repeated

verbatim or the premise word changes, and skew from prompt categories in which the factual prediction token appears inside the subject of the sentence. Their conclusion is that the ablation is ineffective for underrepresented domains and that effectiveness varies with architecture, prompt structure, domain and task. That is Sections 5.1, 5.4, 5.6 and 5.7 restated from inside the mechanistic literature by authors who were trying to confirm a mechanism, not to criticise a construction.

A second objection is that this argument proves too much - that every empirical literature has construction degrees of freedom, and singling this one out is unfair. The distinguishing feature is the sign reversal. Most literatures disagree about magnitude. This one contains papers that name opposite failure modes under the same term, as Section 3 shows for Choi et al. (2025) and Ju et al. (2026), and a documented case of two benchmarks reaching opposite conclusions that a third study traced to dataset design and evaluation metric rather than to a difference between the models the two had tested.

A third objection is that Section 7's distinction between slopes and intercepts is asserted rather than demonstrated. That is correct, and it is the paper's weakest joint. The moderators listed there were each established in a separate study, and no study measures their relative contributions. The claim is offered as the best available reading of a consistent pattern, and the factorial study in Section 11 is what would test it.

13. What is not known, and the quality of the evidence behind what is

Stated flatly, in descending order of how much rests on it.

The relative contribution of the seven choices is unknown. No published study decomposes the variance in reported conflict behaviour by design factor. Where this paper says a choice moves the outcome, the support is a single-factor result from one study, never a comparison of effect sizes across factors.

Two of the seven do not even have that. Sections 5.2 and 5.3 rest on comparing what different papers did rather than on any study that varied the choice while holding the rest fixed. For 5.2 the case is that two papers can score the same behaviour as success and as failure, which is a matter of how the results are worded and is not in dispute, but no published experiment holds construction, items, task, prompt and model constant while flipping only which side is stipulated correct. For 5.3 the closest thing is Dey et al. (2026), whose design varies the measure and the dataset together. These two are the weakest links in the argument.

Whether the seven are the right partition is unknown. They were derived by reading what the cited studies vary, not by any principled decomposition, and they are not independent: task demand and item sample interact, and post- training interacts with prompt sensitivity. A different reader could reasonably merge 5.1 and 5.2, or split 5.7 into architecture, scale and alignment.

How often each choice is recoverable from published text is unknown. The claim that the reporting gap is real rests on the fact that Dey et al. (2026) had to standardise two datasets by hand to compare them, and on Kortukov et al. (2024) having to rebuild the setting to test the synthetic-construction hypothesis. Neither is a survey of reporting practice. Section 11 proposes one.

The scope of the adjudication in Section 6 is narrow and should not be widened. It covers temporal facts, two benchmarks, and a set of models the original studies used at 7B plus an extension across sizes. Its own result is asymmetric: one benchmark's conclusions generalised under both frameworks and the other's did not. Reading it as establishing that all conflict results are construction-bound would be an overreach of exactly the kind this paper is about.

The evidence quality behind the individual numbers quoted here is uneven. The figures from Wu et al. (2024), Bi et al. (2024), Yu et al. (2023), Mallen et al. (2023), Jiang et al. (2020), Nakshatri et al. (2025), Bae et al. (2025) and Su et al. (2024) are each taken from the abstract of the paper that reports

them and were not recomputed. Several are from preprints without a stated venue, including Kortukov et al. (2024), Su et al. (2024), Hou et al. (2024), Dey et al. (2026), Zhao et al. (2026), Tang et al. (2026), Ju et al. (2026), Anand et al. (2026), Bae et al. (2025), Nakshatri et al. (2025) and Setiawan (2026). Where a number carries weight in the argument, it is the direction of the effect that carries it, not the magnitude.

Nothing here was measured by the author. No model was run, no benchmark was rebuilt, and no claim in this paper is supported by anything other than published reports read against their own abstracts.

14. Conclusion

The question of which way a language model bends under a context-memory conflict has produced a five-year literature with contradictory headline findings, and the contradiction is real. The argument here is that it is mostly not a contradiction about models. A conflict is an object the experimenter builds, and seven choices go into building it: the make-up of the conflicting passage, the placement of correctness and the elicitation of the parametric belief, the quantity measured, the item sample, the task's knowledge demand, the prompt, and the checkpoint together with its post-training. For five of the seven, a single study varies that choice within its own design and the reported behaviour moves; for the other two the evidence is cross-study. In the one case where two opposed results were run through each other's protocol, the factors named were dataset design, evaluation metric and model size, and since both original studies had used 7B models, size cannot be what separated their conclusions.

The field's current best reconciliation, that context-memory conflict is task-dependent, is correct and is the methodological model for what the rest of this area needs. It is not, on the evidence, the explanation for the prior disagreements, because task was isolated in a study that deliberately held the other factors constant, and those factors were not constant across the studies that disagreed.

What survives is worth more than it looks. The moderators replicate: adoption of a conflicting context falls as prior confidence rises and as the counterfactual becomes less plausible, items carrying internal inconsistency are harder to update, and the outcome depends strongly on the training frequency of the competing entities. The levels do not, and there is no published evidence that any reported rate identifies a model-level disposition. Until a factorial study decomposes the variance, the honest reading of a single conflict number is that it describes a construction, and the honest way to publish one is to say which construction it describes.

References

- Anand, N., Somasundaram, S., Phukan, A., Saxena, A., & Mukherjee, K. (2026). ContextFocus: Activation Steering for Contextual Faithfulness in Large Language Models. arXiv:2601.04131. [PREPRINT]
- Bae, J., Churchwell, C., Hermon, M., Hsieh, T.-A., Xu, J., Yegorova, Y., Hasegawa-Johnson, M., & Ji, H. (2025). That's Deprecated! Understanding, Detecting, and Steering Knowledge Conflicts in Language Models for Code Generation. arXiv:2510.19116. [PREPRINT]
- Bi, B., Liu, S., Wang, Y., Mei, L., Fang, J., Gao, H., Ni, S., & Cheng, X. (2024). Is Factuality Enhancement a Free Lunch For LLMs? Better Factuality Can Lead to Worse Context-Faithfulness. arXiv:2404.00216. [PREPRINT]
- Choi, E., Park, J., Lee, H., & Lee, J. (2025). Conflict-Aware Soft Prompting for Retrieval-Augmented Generation. Conference on Empirical Methods in Natural Language Processing (EMNLP) 2025. arXiv:2508.15253.
- Dey, R., Ounis, I., McDonald, G., & Moshfeghi, Y. (2026). Temporal Fact Conflicts in LLMs: Reproducibility Insights from Unifying DYNAMICQA and MULAN. arXiv:2603.15892. [PREPRINT]
- Dotsinski, A., Thakur, U., Ivanov, M., Khan, M. H., & Heuss, M. (2025). On the Generalizability of Competition of Mechanisms: Tracing How Language Models Handle Facts and Counterfactuals. Transactions on Machine Learning Research (TMLR) 2025. arXiv:2506.22977.
- Elazar, Y., Kassner, N., Ravfogel, S., Ravichander, A., Hovy, E., Schutze, H., & Goldberg, Y. (2021). Measuring and

Improving Consistency in Pretrained Language Models. *Transactions of the Association for Computational Linguistics (TACL)*. arXiv:2102.01017.

Fierro, C., Garneau, N., Bugliarello, E., Kementchedjheva, Y., & Sogaard, A. (2024). MuLan: A Study of Fact Mutability in Language Models. arXiv:2404.03036. [PREPRINT]

Hou, Y., Pascale, A., Carnerero-Cano, J., Tchakian, T., Marinescu, R., Daly, E., Padhi, I., & Sattigeri, P. (2024). WikiContradict: A Benchmark for Evaluating LLMs on Real-World Knowledge Conflicts from Wikipedia. arXiv:2406.13805. [PREPRINT]

Jacobs, A. Z., & Wallach, H. (2021). Measurement and Fairness. *ACM Conference on Fairness, Accountability, and Transparency (FAccT) 2021*. arXiv:1912.05511.

Jiang, Z., Xu, F. F., Araki, J., & Neubig, G. (2020). How Can We Know What Language Models Know? *Transactions of the Association for Computational Linguistics (TACL) 2020*. arXiv:1911.12543.

Ju, L., Wang, J., & Zhang, Q. (2026). Faithfulness-QA: A Counterfactual Entity Substitution Dataset for Training Context-Faithful RAG Models. arXiv:2604.25313. [PREPRINT]

Kortukov, E., Rubinstein, A., Nguyen, E., & Oh, S. J. (2024). Studying Large Language Model Behaviors Under Context-Memory Conflicts With Real Documents. arXiv:2404.16032. [PREPRINT]

Liu, E., Gangal, V., Yu, M., Tao, Z., Singh, K., Kumar, S., & Feng, S. Y. (2026). HalluWorld: A Controlled Benchmark for Hallucination via Reference World Models. arXiv:2605.19341. [PREPRINT]

Liu, S., & Roth, D. (2026). Conflicts in Texts: Data, Implications and Challenges. arXiv:2504.19472. [PREPRINT]

Longpre, S., Perisetla, K., Chen, A., Ramesh, N., DuBois, C., & Singh, S. (2021). Entity-Based Knowledge Conflicts in Question Answering. *Conference on Empirical Methods in Natural Language Processing (EMNLP) 2021*. arXiv:2109.05052.

Mahendrakar, P. (2026a). Ordering Is Not Resolution: What the Instruction Hierarchy Defines, What It Leaves Undefined, and the Same-Tier Conflicts Agent Benchmarks Do Not Separate. Zenodo. doi:10.5281/zenodo.22135193

Mahendrakar, P. (2026b). Newer Is Not Truer: Three Presuppositions Behind the Recency Rule in Long-Term Agent Memory, and Why Retaining Both Records Relocates the Decision Rather Than Making It. Zenodo. doi:10.5281/zenodo.22182350

Mahendrakar, P. (2026c). Out-of-Distribution With Respect to What? Four Referents Behind One Predicate, and Why OOD Detection Results Do Not Transfer to Foundation-Model Reasoning. Zenodo. doi:10.5281/zenodo.22123547

Mallen, A., Asai, A., Zhong, V., Das, R., Khashabi, D., & Hajishirzi, H. (2023). When Not to Trust Language Models: Investigating Effectiveness of Parametric and Non-Parametric Memories. *Annual Meeting of the Association for Computational Linguistics (ACL) 2023*. arXiv:2212.10511.

Marjanovic, S. V., Yu, H., Atanasova, P., Maistro, M., Lioma, C., & Augenstein, I. (2024). DYNAMICQA: Tracing Internal Knowledge Conflicts in Language Models. *Findings of the Conference on Empirical Methods in Natural Language Processing (EMNLP) 2024*. arXiv:2407.17023.

Mizrahi, M., Kaplan, G., Malkin, D., Dror, R., Shahaf, D., & Stanovsky, G. (2024). State of What Art? A Call for Multi-Prompt LLM Evaluation. *Transactions of the Association for Computational Linguistics (TACL)*. arXiv:2401.00595.

Nakshatri, N. S., Roy, S., Arivazhagan, M. G., Zhou, H., Kumar, V. B., & Gangadharaiyah, R. (2025). When Facts Change: Probing LLMs on Evolving Knowledge with evolveQA. arXiv:2510.19172. [PREPRINT]

Ortu, F., Jin, Z., Doimo, D., Sachan, M., Cazzaniga, A., & Scholkopf, B. (2024). Competition of Mechanisms: Tracing How Language Models Handle Facts and Counterfactuals. *Annual Meeting of the Association for Computational Linguistics (ACL) 2024*. arXiv:2402.11655.

Pham, Q. H., Ngo, H., Luu, A. T., & Nguyen, D. Q. (2024). Who's Who: Large Language Models Meet Knowledge Conflicts in Practice. *Findings of the Conference on Empirical Methods in Natural Language Processing (EMNLP) 2024*. arXiv:2410.15737.

Raji, I. D., Bender, E. M., Paullada, A., Denton, E., & Hanna, A. (2021). AI and the Everything in the Whole Wide World Benchmark. *Conference on Neural Information Processing Systems (NeurIPS) 2021, Benchmarks and Datasets Track*. arXiv:2111.15366.

Setiawan, Y. A. (2026). Prompt-Based Abstention Fails Under Misleading Context: A Controlled Study of Small Frozen RAG Models. arXiv:2608.22228. [PREPRINT]

Sharma, N., Murray, K., & Xiao, Z. (2025). Faux Polyglot: A Study on Information Disparity in Multilingual Large

Language Models. Conference of the North American Chapter of the Association for Computational Linguistics (NAACL) 2025. arXiv:2407.05502.

Su, Z., Zhang, J., Qu, X., Zhu, T., Li, Y., Sun, J., Li, J., Zhang, M., & Cheng, Y. (2024). ConflictBank: A Benchmark for Evaluating the Influence of Knowledge Conflicts in LLM. arXiv:2408.12076. [PREPRINT]

Sun, K., Bai, F., & Dredze, M. (2026). Task Matters: Knowledge Requirements Shape LLM Responses to Context-Memory Conflict. Annual Meeting of the Association for Computational Linguistics (ACL) 2026. arXiv:2506.06485.

Tan, H., Sun, F., Yang, W., Wang, Y., Cao, Q., & Cheng, X. (2024). Blinded by Generated Contexts: How Language Models Merge Generated and Retrieved Contexts When Knowledge Conflicts? Annual Meeting of the Association for Computational Linguistics (ACL) 2024. arXiv:2401.11911.

Tang, J., Wang, K., Lu, H., Chen, H., Chen, K., Sun, Z., Li, Q., Lyu, L., Nan, G., & Zeng, Z. (2026). Diagnosing Knowledge Conflict in Multimodal Long-Chain Reasoning. arXiv:2602.14518. [PREPRINT]

Wan, A., Wallace, E., & Klein, D. (2024). What Evidence Do Language Models Find Convincing? Annual Meeting of the Association for Computational Linguistics (ACL) 2024. arXiv:2402.11782.

Wang, Y., Feng, S., Wang, H., Shi, W., Balachandran, V., He, T., & Tsvetkov, Y. (2024). Resolving Knowledge Conflicts in Large Language Models. Conference on Language Modeling (COLM) 2024. arXiv:2310.00935.

Wu, K., Wu, E., & Zou, J. (2024). ClashEval: Quantifying the tug-of-war between an LLM's internal prior and external evidence. arXiv:2404.10198. [PREPRINT]

Xie, J., Zhang, K., Chen, J., Lou, R., & Su, Y. (2024). Adaptive Chameleon or Stubborn Sloth: Revealing the Behavior of Large Language Models in Knowledge Conflicts. International Conference on Learning Representations (ICLR) 2024. arXiv:2305.13300.

Xu, R., Qi, Z., Guo, Z., Wang, C., Wang, H., Zhang, Y., & Xu, W. (2024). Knowledge Conflicts for LLMs: A Survey. arXiv:2403.08319. [PREPRINT]

Yu, Q., Merullo, J., & Pavlick, E. (2023). Characterizing Mechanisms for Factual Recall in Language Models. arXiv:2310.15910. [PREPRINT]

Zhao, J., Huang, Q., Chen, H., You, X., & Yu, J. (2026). When Abundance Conceals Weakness: Knowledge Conflict in Multilingual Models. arXiv:2601.07041. [PREPRINT]

Zhou, W., Zhang, S., Poon, H., & Chen, M. (2023). Context-faithful Prompting for Large Language Models. Findings of the Conference on Empirical Methods in Natural Language Processing (EMNLP) 2023. arXiv:2303.11315.

Zhu, T., Liu, Q., Wang, F., Tu, Z., & Chen, M. (2024). Unraveling Cross-Modality Knowledge Conflicts in Large Vision-Language Models. arXiv:2410.03659. [PREPRINT]

(c) 2026 Pranay Mahendrakar. Licensed under CC BY 4.0.

The literature search, drafting and citation verification for this paper were carried out with AI assistance under the author's direction. Every citation was machine-verified against the arXiv API and Crossref before inclusion, and every quantitative claim was read back against the cited source's own abstract. The author is responsible for the final text and for all claims made in it.