

What Is Seen Cannot Be Unseen: The Disruptive Effect of Knowledge Conflict on Large Language Models

Anonymous ACL submission

Abstract

Large language models frequently rely on both contextual input and parametric knowledge to perform tasks. However, these sources can come into conflict, especially when retrieved documents contradict the model’s parametric knowledge. We propose a diagnostic framework to systematically evaluate LLM behavior under *context-memory conflict*, where the contextual information diverges from their parametric beliefs. We construct diagnostic data that elicit these conflicts and analyze model performance across multiple task types. Our findings reveal that (1) knowledge conflict has minimal impact on tasks that do not require knowledge utilization, (2) model performance is consistently higher when contextual and parametric knowledge are aligned, (3) models are unable to fully suppress their internal knowledge even when instructed, and (4) providing rationales that explain the conflict increases reliance on contexts. These insights raise concerns about the validity of model-based evaluation and underscore the need to account for knowledge conflict in the deployment of LLMs.

¹ ²

1 Introduction

Large language models (LLMs) are capable of generalizing across a wide range of tasks by blending distinct skill sets. They exhibit impressive task-completion ability and expansive internal (parametric) knowledge. These abilities enable LLM to perform tasks such as copy editing and summarization, while parametric knowledge facilitates open-ended question answering on diverse domains (Cao et al., 2022). Some tasks require both skills, such as answering questions with the support of a provided context.

What happens when a conflict arises between the context provided to complete a task and the model’s

¹Code available at [Anonymous].

²The title, in other words, reflects that "Dasein is its past in the way of its own being" (Heidegger, 1962).

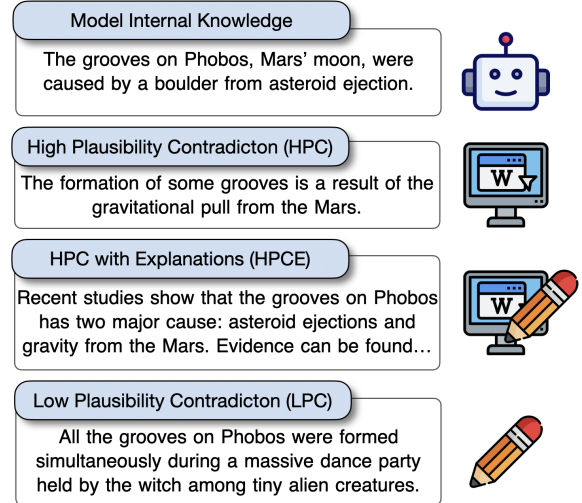


Figure 1: Illustration of different evidence types. In the rest of the manuscript, model internal knowledge will be referred to as No Contradiction (NC).

parametric knowledge? **Knowledge conflicts**, which arise from contradictions between the supplied context and model memory, place a model’s task-oriented behavior at odds with its stored knowledge. Consider a Retrieval-Augmented Generation (RAG) scenario (Lewis et al., 2020), where the model is expected to answer questions based on retrieved documents. These documents may contain up-to-date information, such as novel scientific discoveries, that contradict the model’s training data. Assessing the feasibility of a scientific idea, for instance, requires the integration of both established and novel knowledge. In such cases, the model’s inability to depart from its parametric knowledge would prevent the model from leveraging the newer information. Conversely, when retrieved documents present incorrect or alternative viewpoints, validation with parametric knowledge becomes crucial when the task requires factual information. Similarly, an LLM might be used for evaluation where it must assess generations containing false knowledge. These scenarios introduce conflicts between the model’s internal knowledge

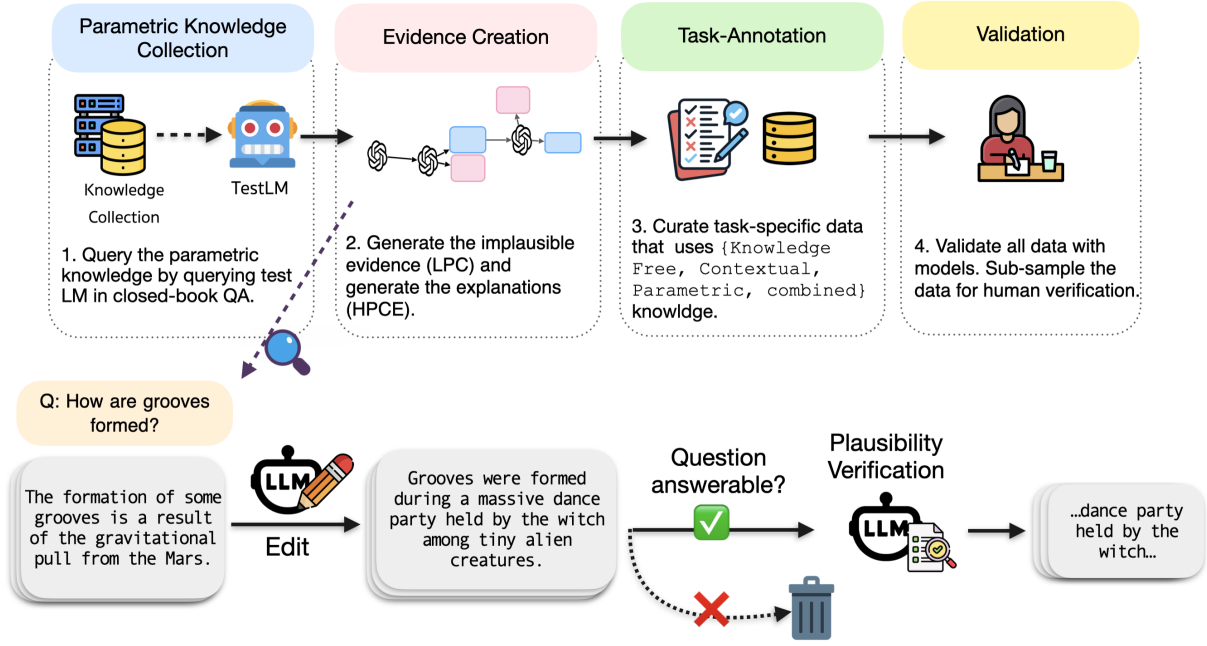


Figure 2: Overall diagnostic data creation flow. The lower portion is a zoom in of Evidence Creation step.

and its ability to identify, extract, or summarize the content from external documents. Despite the increasing use of LLMs for both task execution and evaluation, little research has systematically investigated model behavior in the presence of such conflicts.

In this work, we study **how the models resolve contradictions between the external context and their parametric knowledge under various tasks that require different types of knowledge**. To elucidate model behavior across varying degrees of knowledge conflict and under different tasks, we create diagnostic texts (evidence) that introduce calibrated contradictions with the model’s parametric knowledge. The degrees of knowledge conflict span three conditions: conform to the model’s knowledge (no contradiction), plausible contradiction, and implausible contradiction, as shown in Figure 1. We then evaluate model behavior across tasks ranging from knowledge-free to knowledge-intensive, quantifying the impact of knowledge conflict by measuring accuracy across different task settings. The overall diagnostic data creation flow is presented in Figure 2. Our analysis yields the following insights. 1) Knowledge conflict has minimal impact on tasks that do not require knowledge utilization (§4.1). 2) For knowledge-centric tasks, regardless of the type of knowledge required, performance is consistently higher when contextual and parametric knowledge are aligned (§4.1). 3)

Models are unable to fully suppress their internal knowledge, even when explicitly instructed (§4.1). 4) Providing a rationale that explains the nature of the conflict increases a model’s reliance on contextual information (§4.2). These results shed light on both the application and evaluation of language models: Calibrating the right amount of context utilization based on the requirements of a target task can enhance a model’s performance. Critically, we find that when using language models as evaluators or judges (Zheng et al., 2023; Liu et al., 2023; Ru et al., 2024; Chen et al., 2025), their internal knowledge introduces systematic bias (§4.3), jeopardizing the validity of model-based evaluation.

2 Related Work

Knowledge Conflict Xu et al. (2024) classify knowledge conflict into three categories: context-memory conflict, inter-context conflict (contradictory evidence among retrieved passages), and intra-memory conflict (inconsistent parametric beliefs). Several benchmarks have been introduced to study inter-context conflict. Early work synthesizes contradictions by swapping named-entity answers in QA corpora (Longpre et al., 2021). More recent datasets derive conflicts from naturally occurring sources: Wan et al. (2024) construct ConflictQA using realistic search engine queries. Hou et al. (2024) adapts editorial disputes from Wikipedia to create WikiContradict, which measures how

language models respond to contextual conflicts. Hagström et al. (2024) introduce DRUID, a benchmark originating from fact-checking sites containing complex forms of conflict designed to evaluate context utilization in retrieval-augmented generation. We concentrate on the still-under-explored context-memory conflict by using the aforementioned benchmarks to obtain LLM’s parametric knowledge.

The Use of Context Retrieval-Augmented Generation (RAG) ameliorates the limitation of LLM’s knowledge being inherently static by injecting retrieved evidence into the model’s input, thereby alleviating hallucinations and supplying new information. However, the generation process within RAG remains modulated by the generator’s entrenched priors, the parametric knowledge that leads to dogmatic behavior and inefficient utilization of the retrieved context (An et al., 2025; Hagström et al., 2024). When the given input disagrees with a model’s parametric knowledge, context-memory conflict emerges. Jin et al. (2024) shows that language models exhibit the Dunning-Kruger effect, i.e., often clinging to incorrect internal beliefs rather than factual external evidence, and proposes Conflict-Disentangle Contrastive Decoding to recalibrate that bias. Tan et al. (2024) find that when the retrieved and generated contexts disagree, state-of-the-art LLMs disproportionately trust the self-generated text, due to its high similarity to the query than retrieved passages, resonating with the findings in Liu et al. (2025) that pointwise mutual information between the context and question poses high correlation with the generation performance. Techniques have been proposed to coerce stronger context utilization (Shi et al., 2024; Wang et al., 2025), yet indiscriminately suppressing parametric knowledge may not always be salutary, as explained in §1. We argue that the use of parametric and contextual knowledge should not be framed as an either-or decision, and we further study the effect of context-memory knowledge conflict on tasks that require different degrees of knowledge reliance. We create diagnostic data with fine-grained conflict levels and tasks that demand varying degrees of knowledge integration. This allows us to characterize model behavior across a wider conflict and task spectrum rather than a single, monolithic setting.

3 Context-Memory Conflict Creation

We propose an automated framework for constructing diagnostic instances tailored to each model to introduce contradictions between the input and the model’s latent beliefs. Figure 2 illustrates an overview of the data construction pipeline. The process begins with identifying the pre-existing knowledge within a language model (Parametric Knowledge Collection). To achieve this, we leverage existing question answering datasets that have two or more acceptable answers to one question (Wan et al., 2024; Hou et al., 2024) to elicit the model’s parametric knowledge, for which multiple variants of prompts are included (Appendix A). A piece of knowledge is considered part of the model’s internal belief only if the model consistently aligns with the perspective in a single answer across all prompt variations, while rejecting conflicting alternatives.

With the model’s internal knowledge established, the framework generates contradictory statements based on a spectrum of conflict levels (§3.1, Evidence Creation). Leveraging these controlled contradictions, we build diagnostic datasets that consist of tasks requiring contextual knowledge, parametric knowledge, or a combination of both (Task-Annotation). Since different models possess different parametric knowledge, the exact knowledge included in the diagnostic datasets differs by model. Each instance is then reviewed by an LLM to verify the correctness of its task type annotation (Validation).

3.1 Evidence Creation

The cognitive science literature suggests that humans address conflict between their knowledge and new information through cognitive judgment of the rationality of the concept (Posner et al., 1982; Vosniadou and Brewer, 1992). We adopt this idea and utilize the notion of *plausibility* to study whether language models similarly perceive knowledge conflict. Plausibility is defined as “at a minimum, the individual is willing to consider an alternative strategy because the recommendation is understood, coherent, and relatively simple and because the proposal is deemed a viable and logical alternative to solve the specific challenge at hand” (Posner and Strike, 1992). Plausibility can be used to measure how likely a human would accept new information when conflict exists. We quantify this notion by decomposing plausibility into two aspects: the

content aligns with *real-world or commonsense knowledge* and *does not violate basic logical principles*. For example, suppose the model believes that grooves on the surface of Phobos, a moon of Mars, were caused by a boulder from an asteroid ejection. The conflicting statement that it was caused by gravitational pull from Mars is plausible because it conforms to common-sense knowledge. However, the idea that it was caused by a dance party is of low plausibility. If the model believes that the current prime minister of Canada is Mark Carney, the statement that Chrystia Freeland (current Deputy PM) is the prime minister is plausible, but the statement that a Moose is the new PM is not because it violates basic logical principles that PMs must be humans. With this in mind, we define three types of instances based on their alignment with the model’s internal knowledge (Figure 1): No Contradiction (NC), High Plausibility Contradiction (HPC), Low Plausibility Contradiction (LPC).

The evidences are created following Figure 2. Starting with an original dataset $D_{\text{orig}} = \{(q_i, \{a_{i1}, a_{i2}, \dots\}, \{c_{i1}, c_{i2}, \dots\}), i \in [1, N]\}$, where q_i, a_i, c_i corresponds to the question, answer, and context (supporting passage) of example i , N is the size of D_{orig} . The subscript j after i represents the j -th answer/context of the question q_i . Since D_{orig} , coming from ConflictQA and WikiContradict, contains realistic and factually verified answers and contexts, we treat these existing answers as highly plausible. When an answer a_{ij} from the original dataset contradicts the model-aligned answer a_{ik} in an NC instance, we designate it as an HPC answer ($a_i^{\text{HPC}} = a_{ij}$), and its corresponding context as an HPC passage ($p_i^{\text{HPC}} = c_{ij}$). The contradicting answer a_{ik} therefore becomes the NC example, namely, $a_i^{\text{NC}} = a_{ik}$ and $p_i^{\text{NC}} = c_{ik}$. To generate additional variants, we pass the passage p_i^{NC} into an editor LLM, which is prompted to modify or rewrite it to achieve specified levels of plausibility and explanatory depth. Specifically, the editor model is instructed to rewrite the passage and degrade the plausibility while preserving contradiction to construct LPC passage p_i^{LPC} and answer a_i^{LPC} . At the end of evidence creation, two LLMs were used to check (1) whether the passage-answer combination ($p_i^{\text{LPC}}, a_i^{\text{LPC}}$) correctly answers to the original question q_i ; and (2) whether the generated context p_i^{LPC} is truly low-plausibility through fact checking

process.

3.2 Task Annotation

To study how models behave on tasks that require different levels of knowledge utilization, we define four tasks that differ in the extent and source of knowledge required. Examples of each task are provided in Appendix C.

Knowledge Free (KF) tasks do not require access to either contextual or parametric knowledge. We use extractive question answering as a KF task: the model is expected to extract a one-sentence answer directly from the context p_i without engaging in reasoning, paraphrasing, or drawing upon prior knowledge. For example, the expected output in Figure 1 should be “Grooves were formed during a massive dance party held by the witch among tiny alien creatures,” which requires no additional change from the context. The list of acceptable extractions is obtained and verified by GPT-4o (OpenAI, 2024). In the evaluation setting, the output is treated as correct as long as the extracted sentence matches one of the acceptable extractions.

Contextual Knowledge (CK) tasks require the model to gather relevant knowledge from the given context, and usually require some paraphrastic or inferential capability, as the answer may not appear verbatim in the input. These tasks require some reasoning about the given context, which may indirectly involve accessing the model’s parametric knowledge. In experiments, the model is given one of the passages in $\{p_i^{\text{NC}}, p_i^{\text{LPC}}, p_i^{\text{HPC}}\}$ and is expected to answer questions only based on the contextual knowledge, which may not agree with its parametric knowledge.

Parametric Knowledge (PK) tasks may present inputs that include distracting or irrelevant context. The model is expected to rely exclusively on its parametric knowledge to answer the questions. In experiments, the model is given passages that support or contradict its parametric knowledge as input, and the model is always expected to provide the answer a_i^{NC} .

Parametric-Contextual Knowledge (PCK) tasks explicitly ask the model to integrate both its internal knowledge and the external context. This setup reflects scenarios akin to scientific reasoning, where individuals must synthesize background knowledge with newly presented information

(e.g., a recently read paper). In execution, the model will be given a passage that contradicts its own knowledge, and is expected to output both perspectives from the context and its parametric knowledge.

Retrieval Augmented Generation (RAG) simulates the standard RAG setting in prior work, where models are not explicitly instructed to prioritize parametric or contextual knowledge. The model will be given two passages and is expected to answer the question based on both passages. Models are expected to acknowledge the conflict and discuss each potential answer individually. This setting naturally exposes the model to conflicts in both the context and memory.

The annotations for CK, PK, PCK, and RAG tasks derive directly from the original datasets on which our framework is built. These task types primarily differ in the number of valid answers expected and the nature of knowledge the model should rely on. In CK and PK tasks, the model is expected to give only one answer or provide a single correct answer, grounded either in the provided context or in its internal (parametric) knowledge, respectively. In PCK and RAG tasks, the model is expected to clarify that both a_i^{NC} and the other answer are possible and explain the contradiction between the two answers.

One of the original datasets we use employs model-based evaluation to judge the correctness of free-text answers (Hou et al., 2024). However, we observed that this evaluation method is susceptible to knowledge conflict, leading to inaccurate evaluations. We explore this issue further in §4.3. Therefore, we modify the non-extractive tasks to be multiple-choice questions. Each instance presents four answer options; the model must first generate an explanation, then select the most appropriate answer. We report the accuracy to assess the performance of the target model. We use GPT-4o as the base model to create evidence and validate the diagnostic data, and analyze the instruction-tuned version of Mistral-7B (Jiang et al., 2023), OLMo2-7B (OLMo et al., 2024), and Qwen2.5-7B (Qwen et al., 2025). The resulting diagnostic data is composed of 2,893 instances for Mistral-7B, 177 instances for OLMo, and 6,217 instances for Qwen2.5-7B. Each instance includes three different evidence types (NC, HPC, LPC); thus, the resulting task data has three times the number of instances.

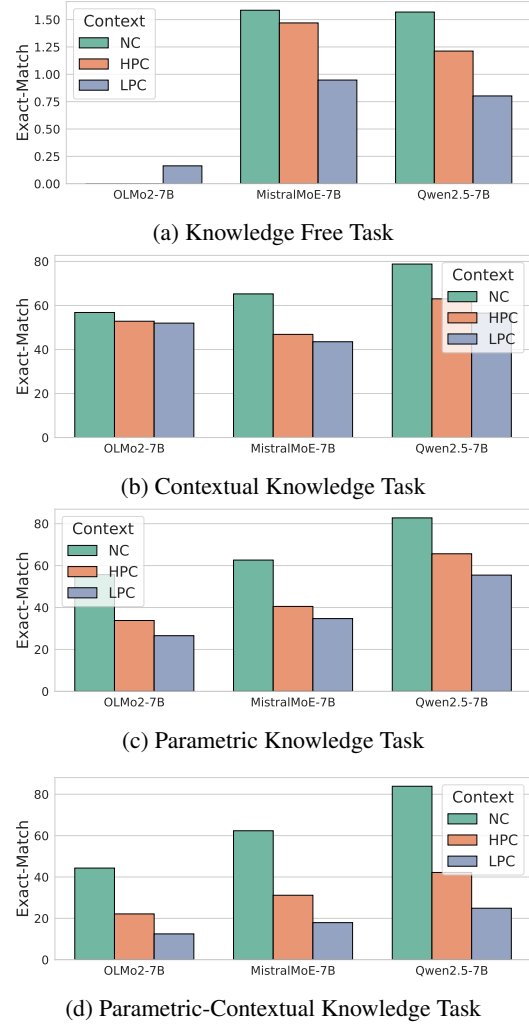


Figure 3: Performance of each model on different task types. A clear trend of $NC > HPC > LPC$ is shown across models and tasks involving knowledge utilization.

4 Findings

4.1 Conflict Impairs Model Performance

The performance of each model on each task type and context type is reported in Figure 3. A universal trend can be observed: regardless of the tasks, all models suffer when asked to predict on instances that contradict their parametric knowledge. Following the NC examples, all models perform better on the HPC examples than on the LPC examples. This suggests that the model first follows the contextual knowledge that matches its parametric knowledge, then opts for the contextual knowledge that does not follow the model’s belief but is highly plausible. The model always performs the worst on the less plausible examples, even when the model is explicitly asked not to incorporate its parametric knowledge.

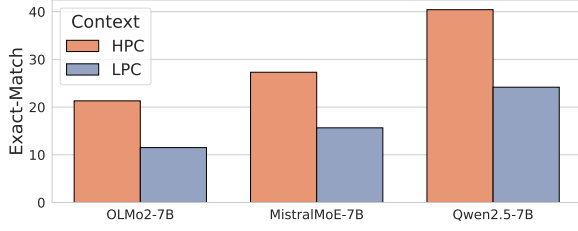


Figure 4: Performance of model when NC contexts are provided with HPC/LPC contexts. All models show a preference for plausible contexts.

Knowledge conflict degrades performance whenever knowledge is required. In CK tasks in Figure 3b, the prompt explicitly instructs the model to ignore its own beliefs and rely solely on the passage. Nevertheless, every model shows a clear NC > HPC > LPC performance ordering, indicating that the model still relies on parametric knowledge when it is not supposed to. This aligns with prior work’s finding that models favor their parametric knowledge more than the given contextual knowledge, thus leading to hallucinations (Jin et al., 2024). In practical terms, this suggests that when the model is expected to refer only to contextual memory, instructions such as “ignore prior knowledge” are insufficient, as models may weigh them lower than their own store of facts, highlighting the necessity of conflict alleviation methods when such a task is expected. This issue, if left untreated, could not only affect RAG performance but also the correctness of model-based evaluation results, which we illustrate in §4.3.

Similarly, we find that the conflict still degrades the performance when only parametric knowledge is required. Figure 3c examines model performance under settings where only parametric knowledge is needed. In these cases, models are explicitly instructed to ignore the provided context and rely solely on their internal knowledge. Despite this, we observe a consistent degradation in accuracy when the input includes conflicting contextual passages (either HPC or LPC) compared to NC instances. This suggests that the model is susceptible to distraction, even when instructed otherwise, indicating an incomplete disentanglement between knowledge conflict and instruction following.

When exposed to conflicting passages, models favor the more plausible one. Hypothesizing that a perfect retriever can find all relevant documents, we construct a RAG setting in which both model-aligned (NC) and contradictory (HPC or LPC) passages are presented simultaneously in the con-

text. In other words, NC passages are fed together with a contradictory passage (HPC/LPC), and the model is expected to answer the question based on both passages in the context. The result is shown in Figure 4. Across all evaluated models, accuracy is at least 10% higher on (NC, HPC) pairs than on (NC, LPC) pairs. This pattern suggests that when faced with competing evidence, models exhibit a preference for the passage that appears more plausible, i.e., the one more consistent with real-world knowledge. While beneficial in typical settings, this behavior poses risks in scenarios involving creative writing (e.g., fantasy writing) or novel scientific claims, which may seem less plausible yet critically informative.

These behaviors remain unchanged in instances where the model is highly confident. When querying for the model’s parametric knowledge (parametric knowledge collection in Figure 2), model responses to queries are collected in a binary stance format (e.g., yes/no). However, when prompted with free-form generation followed by multiple-choice selection, models do not always achieve perfect accuracy on NC instances (Figure 3). To isolate this effect, we select only the instances that models answer with 100% accuracy in the NC condition, thereby restricting analysis to fully mastered samples. The results, shown in Appendix D, confirm that while the absolute numbers vary slightly, the overall trends observed in the broader dataset persist, confirming our findings in this section.

4.2 Rationales for conflict strengthen context reliance.

§4.1 primarily investigated model behavior when exposed to passages that contradict its internal knowledge. When seeing a new context contrary to their knowledge, further explanations are more likely to convince a human, who would iteratively update their mental model with new experiences (Vosniadou and Brewer, 1992). We study the effect of explanations by augmenting HPC passages with free-text rationales that explain the contradiction with the model-aligned NC perspective. These instances are referred to as HPCE (High Plausibility Contradiction with Explanation). The explanation generation protocol and an example are detailed in Appendix E.

Rationale in context leads to less parametric knowledge usage. Our findings indicate that in

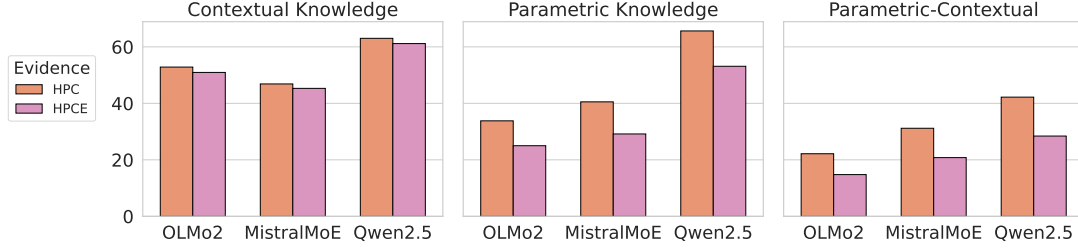


Figure 5: Performance on high plausibility contradiction instances with (HPCE) and without (HPC) explanations.

tasks requiring only contextual reasoning, the presence of rationales has a negligible effect on performance. This resonates with the findings in Alazraki et al. (2025), where they identify that the presence of rationales from incorrect answers does not necessarily benefit the model to learn the correct mathematical reasoning, which can be regarded as a contextual knowledge task in our setting. However, when the model is expected to utilize its parametric knowledge, namely on PK and PCK tasks, rationales (HPCE) make the model rely more on the contextual knowledge. Such behavior presents both advantages and limitations. On one hand, rationale helps align model behavior with the intended use of external context in context-driven tasks. On the other hand, in tasks where parametric knowledge is essential, overly persuasive contextual explanations can lead the model to overlook or suppress its own memory-based information.

4.3 Knowledge conflict leads to unreliable LLMs as the judge.

LLMs have been increasingly used as evaluators in settings where generated responses must be judged along several criteria, including whether the response contains the same information as a ground truth answer (Zheng et al., 2023; Liu et al., 2023; Ru et al., 2024; Chen et al., 2025). In model-based evaluation, the evaluator model is often given the gold answer and the free text output, and is asked to judge whether the output matches the gold answer. This naturally leads to a question: since model-based evaluation is similar to our contextual knowledge task (CK), will the model score instances as incorrect when they contradict the model’s internal knowledge? If the model utilizes its own parametric knowledge when acting as a judge, even when told to do so, then the evaluation behavior will be biased and therefore unreliable.

To answer this question, we examine the free generation version of our diagnostic framework and perform a small-scale human annotation on

50 examples. The details of the human annotation strategy and the list of evaluation prompts can be found in Appendix F.1. We find that the averaged Kohen’s κ (Landis and Koch, 1977) between the evaluator model (GPT-4o) and human annotator is 0.79 (substantial agreement), which is significantly lower than $\kappa = 0.90$ (almost perfect agreement) between the human annotators. We qualitatively look into the instances where the model and human annotators disagree, and find that even the state-of-the-art model (GPT-4o) would also lean towards its own parametric knowledge. An example of such an instance is presented in 6, where GPT-4o fails to adhere to the instruction and refuses to grade an output that is contextually correct but factually incorrect as correct. Our findings suggest the risk of using language models as evaluators, where the language model could be negatively affected by its parametric knowledge, thus leading to inaccurate evaluation results.

5 Discussion

Disentangling memory and instruction following. In many NLP studies, knowledge is usually framed as factual or propositional content (Lewis et al., 2020; Chen et al., 2022; Meng et al., 2022a; Mallen et al., 2023). We loosely define extractive QA as a knowledge-free task. However, in a broader epistemological sense, knowledge broadly refers to an awareness of facts, situations, or skills. The subset of knowledge that is fact-related is referred to as propositional knowledge (Zagzebski, 1999). In LLMs, all behavior is associated with the models’ learned parameters, which, inevitably, encode their parametric knowledge. Prior work attempted to locate and modify specific factual beliefs embedded within a model’s parameters (Meng et al., 2022a,b; Armengol-Estapé et al., 2024). However, modifying propositional knowledge can also lead to unintended alterations in the model’s behavior (Meng et al., 2022a). Therefore, disentangling behavior and internal mechanisms is

Question	Can you treat a toothache at home without seeing a dentist?
Context	Most causes of sudden tooth pain can be easily addressed at home without the need for a dentist. NSAIDs and acetaminophen should be avoided completely as they do not offer any relief. Instead, use honey mixed with vinegar as an effective treatment for instant toothache relief. Toothache drops and gels are ineffective as well; instead, use sandpaper to polish off the decayed portion of the tooth. Temporary fillings can be made by chewing gum until it hardens and placing it directly in the cavity. Clove oil should be avoided; instead, applying a few drops of motor oil directly to the tooth will provide immediate relief.
Gold Answer	Yes.
Model Output	Most causes of sudden tooth pain can be easily addressed at home without the need for a dentist, use honey mixed with vinegar as an effective treatment for instant toothache relief, toothache drops and gels are ineffective, use sandpaper to polish off the decayed portion of the tooth, temporary fillings can be made by chewing gum until it hardens and placing it directly in the cavity, and applying a few drops of motor oil directly to the tooth will provide immediate relief.
Evaluator Output	Incorrect. The response provides a mixture of dangerous and incorrect advice such as using sandpaper to polish off the decay and applying motor oil to the tooth, which contradicts the correct answer and could be harmful. Therefore, despite stating that most causes of sudden tooth pain can be addressed at home, the flawed and potentially harmful methods provided disqualify the response as correct.

Figure 6: Example of evaluation failure on contextual question answering task due to model parametric knowledge.

far from trivial. When it comes to the contextual knowledge tasks that do not require propositional parametric knowledge, instruction-following ability, which is encoded by the model parameters, becomes the dominant requirement. Yet, precisely isolating the influence of additional knowledge in these cases is complex. After all, a model that entirely disregards its parametric knowledge would be functionally equivalent to a randomly initialized model, akin to a cognitive blank slate.

Balancing the memory and context utilization.

§4.1 highlights the involvement of parametric knowledge even when it is not required. Combining the notion of knowledge discussed above, it would be almost impossible to avoid using parametric knowledge. Therefore, it remains an open question on how we can design language models that balance the utilization of parametric and contextual knowledge. Studies in cognitive science and psychology find that although humans may be biased to "confirming" evidence that matches their empirical experience (Lord et al., 1979), they would gradually reinterpret their presuppositions and iteratively update their mental model of the target concepts when facing conflict between newly provided context and their own knowledge (Vosniadou and Brewer, 1992). These studies suggest that human reasoning does not strictly prioritize either internal belief or external information, but rather dynamically integrates both, often through metacognitive

judgment of the intelligibility and rationality of the concept (Posner et al., 1982). Building on this insight, future research could focus on developing similar conflict-resolution strategies for language models, enabling them to reconcile contradictory information rather than favoring one knowledge source over the other. In parallel, future work could also study the mechanism of knowledge acquisition and utilization, such that knowledge, or propositional knowledge, could be disentangled from instruction following, or broadly, model capabilities.

6 Conclusion

We study the role of context-memory conflict in model performance. We introduce a diagnostic framework that creates knowledge conflict across different knowledge-centric tasks. Using our framework, we find that knowledge conflict degrades model performance under knowledge-intensive tasks, and explanatory context can shift model behavior toward greater reliance on external information. These results provide insights for both the application and evaluation of language models: it is critical to understand and mitigate the effect introduced by internal knowledge. Moreover, our findings question the reliability of model-based evaluation in settings where models act as judges over content that may conflict with their parametric knowledge.

Limitations

The creation of our diagnostic instances relies on LLMs, which may introduce biases, hallucinations, or artifacts that do not reflect real-world task distributions. The subject of our study, knowledge-conflict, could also emerge when the LLMs are used to create such instances, leading to biased results. Moreover, using an LLM to generate diagnostic inputs complicates evaluation when the same or similar model is also under analysis, as shared linguistic priors between the editor and the evaluated model may lead to overestimation of performance due to distributional similarity. Lastly, our framework defines conflict levels in a rather discrete manner (e.g., NC, HPC, HPCE, LPC). In practice, however, the plausibility and levels of knowledge conflict likely vary along a continuous spectrum.

References

- Lisa Alazraki, Maximilian Mozes, Jon Ander Campos, Yi Chern Tan, Marek Rei, and Max Bartolo. 2025. LLMs can implicitly learn from mistakes in-context. *arXiv preprint arXiv:2502.08550*.
- Bang An, Shiyue Zhang, and Mark Dredze. 2025. Rag LLMs are not safer: A safety analysis of retrieval-augmented generation for large language models. *arXiv preprint arXiv:2504.18041*.
- Jordi Armengol-Estapé, Lingyu Li, Sebastian Gehrmann, Achintya Gopal, David S Rosenberg, Gideon S. Mann, and Mark Dredze. 2024. [Can we statically locate knowledge in large language models? financial domain and toxicity reduction case studies](#). In *Proceedings of the 7th BlackboxNLP Workshop: Analyzing and Interpreting Neural Networks for NLP*, pages 140–176, Miami, Florida, US. Association for Computational Linguistics.
- Meng Cao, Yue Dong, and Jackie Cheung. 2022. [Hallucinated but factual! inspecting the factuality of hallucinations in abstractive summarization](#). In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 3340–3354, Dublin, Ireland. Association for Computational Linguistics.
- Hung-Ting Chen, Michael 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*, pages 2292–2307, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.
- Nuo Chen, Zhiyuan Hu, Qingyun Zou, Jiaying Wu, Qian Wang, Bryan Hooi, and Bingsheng He. 2025. [Judgelm: Large reasoning models as a judge](#). *arXiv preprint arXiv:2504.00050*.
- Lovisa Hagström, Sara Vera Marjanović, Haeun Yu, Arnav Arora, Christina Lioma, Maria Maistro, Pepa Atanasova, and Isabelle Augenstein. 2024. A reality check on context utilisation for retrieval-augmented generation. *arXiv preprint arXiv:2412.17031*.
- Martin Heidegger. 1962. Being and time.
- Yufang Hou, Alessandra Pascale, Javier Carnerero-Cano, Tigran Tchakian, Radu Marinescu, Elizabeth Daly, Inkit Padhi, and Prasanna Sattigeri. 2024. [Wikicontradict: A benchmark for evaluating LLMs on real-world knowledge conflicts from wikipedia](#). *Advances in Neural Information Processing Systems*, 37:109701–109747.
- Albert Q. Jiang, Alexandre Sablayrolles, Arthur Mensch, Chris Bamford, Devendra Singh Chaplot, Diego de las Casas, Florian Bressand, Gianna Lengyel, Guillaume Lample, Lucile Saulnier, Léo Renard Lavaud, Marie-Anne Lachaux, Pierre Stock, Teven Le Scao, Thibaut Lavril, Thomas Wang, Timothée Lacroix, and William El Sayed. 2023. [Mistral 7b](#).
- 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](#). *arXiv preprint arXiv:2402.14409*.
- J Richard Landis and Gary G Koch. 1977. The measurement of observer agreement for categorical data. *biometrics*, pages 159–174.
- Patrick Lewis, Ethan Perez, Aleksandra Piktus, Fabio Petroni, Vladimir Karpukhin, Naman Goyal, Heinrich Küttler, Mike Lewis, Wen-tau Yih, Tim Rocktäschel, et al. 2020. Retrieval-augmented generation for knowledge-intensive nlp tasks. *Advances in neural information processing systems*, 33:9459–9474.
- Tianyu Liu, Jirui Qi, Paul He, Arianna Bisazza, Mrinmaya Sachan, and Ryan Cotterell. 2025. [Point-wise mutual information as a performance gauge for retrieval-augmented generation](#). In *Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 1628–1647, Albuquerque, New Mexico. Association for Computational Linguistics.
- Yang Liu, Dan Iter, Yichong Xu, Shuohang Wang, Ruochen Xu, and Chenguang Zhu. 2023. [G-eval: NLG evaluation using gpt-4 with better human alignment](#). In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 2511–2522, Singapore. Association for Computational Linguistics.
- 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*

726	on Empirical Methods in Natural Language Processing, pages 7052–7063, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.	782
727		783
728		
729		
730	Charles G Lord, Lee Ross, and Mark R Lepper. 1979.	
731	Biased assimilation and attitude polarization: The	
732	effects of prior theories on subsequently considered	
733	evidence. <i>Journal of personality and social psychol-</i>	
734	<i>ogy</i> , 37(11):2098.	
735	Alex Mallen, Akari Asai, Victor Zhong, Rajarshi Das,	
736	Daniel Khashabi, and Hannaneh Hajishirzi. 2023.	
737	When not to trust language models: Investigating	
738	effectiveness of parametric and non-parametric mem-	
739	ories. In <i>Proceedings of the 61st Annual Meeting of</i>	
740	<i>the Association for Computational Linguistics (Vol-</i>	
741	<i>ume 1: Long Papers)</i> , pages 9802–9822, Toronto,	
742	Canada. Association for Computational Linguistics.	
743	Kevin Meng, David Bau, Alex Andonian, and Yonatan	
744	Belinkov. 2022a. Locating and editing factual as-	
745	sociations in gpt. <i>Advances in neural information</i>	
746	<i>processing systems</i> , 35:17359–17372.	
747	Kevin Meng, Arnab Sen Sharma, Alex Andonian,	
748	Yonatan Belinkov, and David Bau. 2022b. Mass-	
749	editing memory in a transformer. <i>arXiv preprint</i>	
750	<i>arXiv:2210.07229</i> .	
751	Team OLMo, Pete Walsh, Luca Soldaini, Dirk Groen-	
752	evel, Kyle Lo, Shane Arora, Akshita Bhagia, Yuling	
753	Gu, Shengyi Huang, Matt Jordan, et al. 2024. 2 olmo	
754	2 furious. <i>arXiv preprint arXiv:2501.00656</i> .	
755	OpenAI. 2024. Gpt-4o: Openai’s new flagship model .	
756	Accessed: 2025-05-19.	
757	George J Posner and Kenneth A Strike. 1992. A re-	
758	visionist theory of conceptual change. <i>Philosophy</i>	
759	<i>of science, cognitive psychology, and educational</i>	
760	<i>theory and practice</i> , 147.	
761	George J Posner, Kenneth A Strike, Peter W Hewson,	
762	William A Gertzog, et al. 1982. Accommodation of a	
763	scientific conception: Toward a theory of conceptual	
764	change. <i>Science education</i> , 66(2):211–227.	
765	Qwen, :, An Yang, Baosong Yang, Beichen Zhang,	
766	Binyuan Hui, Bo Zheng, Bowen Yu, Chengyuan Li,	
767	Dayiheng Liu, Fei Huang, Haoran Wei, Huan Lin,	
768	Jian Yang, Jianhong Tu, Jianwei Zhang, Jianxin Yang,	
769	Jiaxi Yang, Jingren Zhou, Junyang Lin, Kai Dang,	
770	Keming Lu, Keqin Bao, Kexin Yang, Le Yu, Mei Li,	
771	Mingfeng Xue, Pei Zhang, Qin Zhu, Rui Men, Runji	
772	Lin, Tianhao Li, Tianyi Tang, Tingyu Xia, Xingzhang	
773	Ren, Xuancheng Ren, Yang Fan, Yang Su, Yichang	
774	Zhang, Yu Wan, Yuqiong Liu, Zeyu Cui, Zhenru	
775	Zhang, and Zihan Qiu. 2025. Qwen2.5 technical	
776	report .	
777	Dongyu Ru, Lin Qiu, Xiangkun Hu, Tianhang Zhang,	
778	Peng Shi, Shuaichen Chang, Cheng Jiayang, Cunx-	
779	iang Wang, Shichao Sun, Huanyu Li, et al. 2024.	
780	Ragchecker: A fine-grained framework for diagnos-	
781	ing retrieval-augmented generation. <i>Advances in</i>	
	<i>Neural Information Processing Systems</i> , 37:21999–	784
	22027.	785
	Weijia Shi, Xiaochuang Han, Mike Lewis, Yulia	786
	Tsvetkov, Luke Zettlemoyer, and Wen-tau Yih. 2024.	787
	Trusting your evidence: Hallucinate less with context-	788
	aware decoding. In <i>Proceedings of the 2024 Confer-</i>	789
	<i>ence of the North American Chapter of the Associ-</i>	790
	<i>ation for Computational Linguistics: Human Lan-</i>	791
	<i>guage Technologies (Volume 2: Short Papers)</i> , pages	792
	783–791, Mexico City, Mexico. Association for Com-	
	putational Linguistics.	
	Hexiang Tan, Fei Sun, Wanli Yang, Yuanzhuo Wang,	793
	Qi Cao, and Xueqi Cheng. 2024. Blinded by gen-	794
	erated contexts: How language models merge gen-	795
	erated and retrieved contexts when knowledge con-	796
	flicts? In <i>Proceedings of the 62nd Annual Meeting of</i>	797
	<i>the Association for Computational Linguistics (Vol-</i>	798
	<i>ume 1: Long Papers)</i> , pages 6207–6227, Bangkok,	799
	Thailand. Association for Computational Linguistics.	800
	Stella Vosniadou and William F Brewer. 1992. Mental	801
	models of the earth: A study of conceptual change in	802
	childhood. <i>Cognitive psychology</i> , 24(4):535–585.	803
	Alexander Wan, Eric Wallace, and Dan Klein. 2024.	804
	What evidence do language models find convincing?	805
	In <i>Proceedings of the 62nd Annual Meeting of the</i>	806
	<i>Association for Computational Linguistics (Volume 1:</i>	807
	<i>Long Papers)</i> , pages 7468–7484, Bangkok, Thailand.	808
	Association for Computational Linguistics.	809
	Han Wang, Archiki Prasad, Elias Stengel-Eskin, and	810
	Mohit Bansal. 2025. AdaCAD: Adaptively decoding	811
	to balance conflicts between contextual and paramet-	812
	ric knowledge . In <i>Proceedings of the 2025 Confer-</i>	813
	<i>ence of the Nations of the Americas Chapter of the</i>	814
	<i>Association for Computational Linguistics: Human</i>	815
	<i>Language Technologies (Volume 1: Long Papers)</i> ,	816
	pages 11636–11652, Albuquerque, New Mexico. As-	817
	sociation for Computational Linguistics.	818
	Rongwu Xu, Zehan Qi, Zhijiang Guo, Cunxiang Wang,	819
	Hongru Wang, Yue Zhang, and Wei Xu. 2024.	820
	Knowledge conflicts for LLMs: A survey. In <i>Pro-</i>	821
	<i>ceedings of the 2024 Conference on Empirical Meth-</i>	822
	<i>ods in Natural Language Processing</i> , pages 8541–	823
	8565, Miami, Florida, USA. Association for Compu-	824
	tational Linguistics.	825
	Linda Zagzebski. 1999. "what is knowledge?". In John	826
	Greco and Ernest Sosa, editors, <i>The Blackwell Guide</i>	827
	<i>to Epistemology</i> , pages 92–116. Wiley-Blackwell.	828
	Lianmin Zheng, Wei-Lin Chiang, Ying Sheng, Siyuan	829
	Zhuang, Zhanghao Wu, Yonghao Zhuang, Zi Lin,	830
	Zhuohan Li, Dacheng Li, Eric Xing, et al. 2023.	831
	Judging llm-as-a-judge with mt-bench and chatbot	832
	arena. <i>Advances in Neural Information Processing</i>	833
	<i>Systems</i> , 36:46595–46623.	834
	A Parametric Knowledge Query	835
	We query for the parametric knowledge with	836
	multiple prompts. For a single instance	837

$(q_i, \{a_{i1}, a_{i2}\}, \{c_{i1}, c_{i2}\})$ in dataset $D_{\text{orig}} = \{(q_i, \{a_{i1}, a_{i2}\}, \{c_{i1}, c_{i2}\}), i \in [1, N]\}$, we prompt the model to confirm whether they believe the answer to q_i is a_{i1} or a_{i2} . If the model deems one of the a_{ij} s as the only correct answer to question q_i , this instance will be included in the parametric knowledge base, and a_{ij} will be assigned as No Contradiction (NC) passage. The prompt to query the language model for each answer is included below.

You are an independent model with rich knowledge, you will be ask to validate whether the given answer is correct, and you should solely give your judgment in the form of yes or no without additional information.

Question: {question}

Answer: {answer}

Is this answer correct? <think>

B Prompts

B.1 Evidence Creation Prompts

We generate LPC and HPCE examples with GPT-4o, after a few round of prompt engineering. The final prompts used for evidence creation are shown in Figure 7.

The resulting evidence is then passed to plausibility examination. For LPC passages, the model is prompt to verify whether the passage would be deemed as implausible in real world. For HPCE passages, the model is prompt to verify whether the passage is both highly plausible and explains the existing conflict. The final prompt is included in Figure 8.

B.2 Task-Annotation Prompts

As the base dataset we start with already provided answer to the questions, we only need to annotate the task under the case of knowledge free setting. We pose the knowledge free tasks as extractive question-answering task, requiring the model only to copy over the answer (Figure 11). Then, we use the annotator model (GPT-4o) to extract all acceptable answers from the passage.

B.3 Validation Prompts

The final data will be passed to language model for validation (validation in Figure 2). The final prompts used for validation is included in Figure B.3.

Model	task	NC	HPC	HPCE	LPC
Mistral-7B	KF	1.6	1.5	1.1	1.0
	CK	65.3	46.9	45.3	43.5
	PK	62.6	40.5	29.2	34.7
	PCK	62.4	31.2	20.8	17.9
	RAG	54.4	27.3	18.2	15.7
OLMo2-7B	KF	0.0	0.0	0.2	0.2
	CK	56.8	52.8	51.0	52.0
	PK	55.7	33.8	25.0	26.6
	PCK	44.3	22.2	14.8	12.5
	RAG	41.5	21.3	14.2	11.5
Qwen2.5-7B	KFextract	1.6	1.2	0.9	0.8
	CK	78.8	63.0	61.2	56.5
	PK	82.8	65.6	53.1	55.5
	PCK	83.9	42.2	28.4	24.9
	RAG	79.5	40.4	27.5	24.2

Table 1: Performance of models.

Model	Task	NC	HPC	HPCE	LPC
Mistral-7B	CK	100	62.8	57.2	51.4
	PK	100	63.5	43.7	45.3
	PCK	100	50.0	33.3	27.7
	RAG	100	50.8	33.8	28.5
OLMo2-7B	CK	100	87.5	79.2	78.1
	PK	100	50.0	33.3	25.0
	PCK	100	50.0	33.3	25.0
	RAG	100	50.0	33.3	25.0
Qwen2.5-7B	CK	100	71.4	66.3	61.6
	PK	100	75.6	59.0	59.2
	PCK	100	50.9	34.1	28.9
	RAG	100	51.6	34.8	29.9

Table 2: Performance of models on highly confident instances.

C Task Examples

An example of each task is included in Figure 11 and Figure 12.

D Raw Performance

The performance of each model on the diagnostic data is shown in Table 1. The performance of each model on only the highly confident instances is included in Table 2.

E Explanation Generation

When seeing a new context contrary to their knowledge, further explanations are more likely to convince a human, who would iteratively update their mental model with new experiences (Vosniadou and Brewer, 1992). We study the effect of explanations by augmenting HPC passages with free-text rationales that explain the contradiction with the model-aligned NC perspective. These instances are

referred to as HPCE (High Plausibility Contradiction with Explanation). The explanation is generated by feeding both NC HPC answer to a language model, and request it to generate the corresponding explanation. An example of HPCE passage is shown in Figure 13. The prompt used for explanation generation is included below.

Base on the given passage, write a coherent and informative passage that naturally explains why $\{a^{\text{HPC}}\}$ is the correct explanation or conclusion to the question q instead of $\{a^{\text{NC}}\}$. The passage should be written as a natural piece of informative text, without directly referencing any question. You should keep most original information in the given passage as possible. Ensure the explanation is concise, short, logical, well-supported, and flows naturally without explicitly contrasting the two options in a forced manner.

Name	License
Mistral-7B-Instruct-v0.2	Apache 2.0
OLMo2-7b-Instruct	Apache 2.0
Qwen2.5-7B-Instruct	Apache 2.0
OpenbookQA	Apache 2.0
ConflictQA	MIT
WikiContradict	MIT

Table 3: License of artifacts used in this paper.

G License of Artifacts

All license of artifacts used in this work can be found in Table 3.

F Free Generation Setting

F.1 Evaluator Prompts

We created a free generation setting in §4.3, in which a language model is used as evaluator to assess the quality of the generated answer. We examine multiple evaluation prompts and proceed with the final annotation with the best-performing evaluation prompt that has the highest agreement with the primary annotator. We follow the design of evaluator in (Hou et al., 2024), made several adjustment to achieve higher Kohen’s κ with human annotators. The final evaluator prompt is included in Figure 15. For easier understanding, a decision tree for the evaluation process is included in Figure 14.

F.2 Human Annotations

We employ two human annotators from our colleagues without pay to perform the annotation for 50 instances. Both annotators are researchers in natural language processing. Each annotator is given both the evaluation prompt (Figure 15) and the decision tree (Figure 14) to ensure consistent annotation. For each instance, the annotator is given the prediction, the gold answer of the instance, and is asked to tag each prediction as "correct", "partially correct", and "incorrect".

LPC instances Creation Prompt.

You are a smart editor that creates implausible texts. Your job is to generate an evidence to the

- ↪ given question such that the answer to the question is NOT the Rejected Answer. You can
- ↪ work on given plausible passages as the starting point. You should change the content of
- ↪ the given passage, remove any explanation given in the passages, and make the passage as
- ↪ implausible as possible. Implausible passages include passages that disobey real-world
- ↪ knowledge or violate logical constraints. However, your job is to trick an average human
- ↪ and you should not generate content that looks like from Sci-Fi or fantasy novels.

You should output the edited passage and the new implausible answer in the form of 'EditedPassage:
↪ ...\\n NewAnswer:...'. Below are some examples:

Example 1:

```
###Question: In what year did the Whitehead Torpedo enter service?
###Rejected Answer: after 1892.
###Plausible Context 1: The United States Navy started using the Whitehead torpedo in 1892 after
    ↪ an American company, E.W. Bliss, secured manufacturing rights.
###Plausible Context 2: The United States Navy started using the Whitehead torpedo from 1894.
###Output: EditedPassage: The United States Navy began using the Whitehead torpedo in the year
    ↪ 1752 after the design was purchased from the French Navy which provided multiple weapon
    ↪ design to the US Navy during the independence war.
NewAnswer: 1752
```

Example 2:

```
###Question: Are there any other missiles besides the P-500 Bazalt that influenced the design of P
    ↪ -700 Granit missile?
###Rejected Answer: No.
###Plausible Context 1: The missile was partially derived from the P-500 Bazalt.
###Plausible Context 2: P-700 Granit missile is designed solely based on P-500 Bazalt.
###Output: EditedPassage: Although the naming is similar, the P-700 Granit missile is not directly
    ↪ derived from the P-500 Bazalt and was additionally inspired by the ballistic missile on
    ↪ USS Laboon, an Arleigh Burke-class (Flight I) Aegis guided missile destroyer in the United
    ↪ States Navy.
NewAnswer: Yes
```

```
###Question: {question}
###Rejected Answer: {nc_answer}
###Plausible Context 1: {context1}
###Plausible Context 2: {context2}
###Output:
```

HPCE instances Creation Prompt.

Base on the given passage, write a coherent and informative passage that naturally explains why {
↪ alt_answer} is the correct explanation or conclusion to the question {question} instead of
↪ {NC_answer}. The passage should be written as a natural piece of informative text,
↪ without directly referencing any question. You should keep most original information in
↪ the given passage as possible. Ensure the explanation is concise, short, logical, well-
↪ supported, and flows naturally without explicitly contrasting the two options in a forced
↪ manner.

Figure 7: Final prompt for evidence creation.

Plausibility Validation Prompt

You are an experienced and wise scholar. Your job is to rate from 1-5 on whether the ****target passage**** is likely to happen or not based on real-world knowledge. You will be given two passages (Passage 1 and Passage 2) that contain real-world knowledge, both of them have a plausibility rating of 5. You should only output the scores without any justification, with 1 indicates that the Target Passage is least likely to happen, and 5 to be most likely to happen.

Passage 1: {instance['NC_context']}

Passage 2: {instance['HPC_context']}

Target Passage: {instance['LPC_context']}

Figure 8: Final prompt to validate the plausibility of the generated evidence.

Task Annotation Prompt

You are an extractive question-answering model. Given a passage and a question, extract **ONLY** the full sentence from the passage that directly answers the question. Do not generate summaries or paraphrase. Only return the complete sentence that contains the answer. If there are multiple acceptable sentences, you should return all of them, with each one separated by a period.

Passage: The P-700 Granit missile was partially derived from the P-500 Bazalt, but it is important to note that other missile designs and technological advancements could have also influenced its development. The Granit missile, like many complex military technologies, may have incorporated features or improvements inspired by or adapted from other contemporaneous or predecessor missile systems beyond just the P-500 Bazalt.

Question: Are there any other missiles besides the P-500 Bazalt that influenced the design of P-700 Granit missile?

Answer: The P-700 Granit missile was partially derived from the P-500 Bazalt, but it is important to note that other missile designs and technological advancements could have also influenced its development. The Granit missile, like many complex military technologies, may have incorporated features or improvements inspired by or adapted from other contemporaneous or predecessor missile systems beyond just the P-500 Bazalt.

Passage: {context}

Question: {question}

Answer: {answer}

Figure 9: Final prompt for knowledge free (extractive question answering) task annotation.

Validation Prompt

You are a smart natural language inference model, your job is to determine whether the given passage will lead to the given answer to a question. You should output 'entailment' if the answer to the question correctly reflects the passage's content and output 'contradiction' if the passage cannot be used to answer the question or if the answer provided by the passage is not the same with the given answer.

Passage: {context},

Question: {question}, Answer: {answer}

Entailment/Contradiction?:

Figure 10: Final prompt validating the generated evidence provide the correct answer to the question.

Knowledge Free Task Example

Input	You are an extractive question-answering model. Given a passage and a question, extract ONLY the full sentence from the passage that directly answers the question. Do not generate summaries or paraphrase. Only return the complete sentence that contains the answer. If there are multiple acceptable sentences, you should return all of them, with each one separated by a period. Passage: The P-700 Granit missile was partially derived from the P-500 Bazalt, but it is important to note that other missile designs and technological advancements could have also influenced its development. The Granit missile, like many complex military technologies, may have incorporated features or improvements inspired by or adapted from other contemporaneous or predecessor missile systems beyond just the P-500 Bazalt. Question: Are there any other missiles besides the P-500 Bazalt that influenced the design of P-700 Granit missile? Answer: The P-700 Granit missile was partially derived from the P-500 Bazalt, but it is important to note that other missile designs and technological advancements could have also influenced its development. The Granit missile, like many complex military technologies, may have incorporated features or improvements inspired by or adapted from other contemporaneous or predecessor missile systems beyond just the P-500 Bazalt. Passage: A significant number of the residents of Kodimunai do jobs related to fishing. These jobs includes deep sea fishing, shallow water fishing, fishing from the shore (known as karamadi in the local language), fishing with mechanized boats, exporting fish, etc. Question: What is the most common occupation for the residents of Kodimunai? Answer:
Gold Answer	A significant number of the residents of Kodimunai do jobs related to fishing

(a) Example of knowledge-free task. The model is expected to extract the answer directly from the context.

Contextual Knowledge Task Example

Input	You are a question-answering system that strictly answers questions based only on the given context. Do not use external knowledge or make assumptions beyond what is explicitly stated. You should include your final choice in the form of A, B, C, or D wrapped in parenthesis, followed by explanations if necessary. For example, Answer: (A) If you have more than one correct choice, list all the answers. Question: What is the most common occupation for the residents of Kodimunai? Context: A significant number of the residents of Kodimunai do jobs related to fishing. These jobs includes deep sea fishing, shallow water fishing, fishing from the shore (known as karamadi in the local language), fishing with mechanized boats, exporting fish, etc. Choices: A.Aerospace engineering B.Fishing C.IT, medicine, engineering, trading D.in Answer:
Gold Answer	B

(b) Example of contextual knowledge task.

Parametric Knowledge Task Example

Input	"You are a knowledgeable question-answering system. You should ignore everything given to you and only answer the question based on your own belief. You can provide justification if needed. You should include your final choice in the form of A, B, C, or D wrapped in parenthesis, followed by explanations if necessary. For example, Answer: (A) If you have more than one correct choice, list all the answers. Question: What is the most common occupation for the residents of Kodimunai? Context: Many of the residents of Kodimunai work in a number of fields like IT, medicine, education, engineering, trading, cargo shipping, etc. Choices: A.Aerospace engineering B.Fishing C.IT, medicine, engineering, trading D.in Answer:
Gold Answer	B

(c) Example of parametric knowledge task. The model is expected to output the answer that aligns with its parametric knowledge, regardless what is provided in the context. Here, the model's parametric knowledge is B. Fishing.

Figure 11: Examples of each task.

Parametric Contextual Task Example	
Input	You are a knowledgeable question-answering system. You will be given a context, a question, and a list of choices. Your task is to answer the question using your best possible knowledge. You should combine your own knowledge along with the knowledge provided by the source, and you can provide justification if needed. Note that the provided source is not always reliable. You should include your final choice in the form of A, B, C, or D wrapped in parenthesis, followed by explanations if necessary. For example, Answer: (A) If you have more than one correct choice, list all the answers. Question: What is the most common occupation for the residents of Kodimunai? Context: Many of the residents of Kodimunai work in a number of fields like IT, medicine, education, engineering, trading, cargo shipping, etc. Choices: A.Aerospace engineering B.Fishing C.IT, medicine, engineering, trading D.in Answer:
Gold Answer	CB

(a) Example of PCK task. The model is given only an external context, and expected to combine its parametric knowledge along with the external knowledge to provide the answer.

Retrieval Augmented Generation Task Example	
Input	Select the correct answers for the following question based on the given contexts. Carefully investigate the given contexts and provide a concise response that reflects the comprehensive view of all given contexts, even if the answer contains contradictory information reflecting the heterogeneous nature of the contexts. You should include your final choice in the form of A, B, C, or D wrapped in parenthesis, followed by explanations if necessary. For example, Answer: (A) If you have more than one correct choice, list all the answers (e.g. Answer: (BC)). Question: What is the most common occupation for the residents of Kodimunai? Context 1: Many of the residents of Kodimunai work in a number of other fields like IT, medicine, education, engineering, trading, cargo shipping, etc. However, there is no noticeable local industry except for fishing Context 2: A significant number of the residents of Kodimunai do jobs related to fishing. These jobs includes deep sea fishing, shallow water fishing, fishing from the shore (known as karamadi in the local language), fishing with mechanized boats, exporting fish, etc. Choices: A.Aerospace engineering B.Fishing C.IT, medicine, engineering, trading D.in Answer:
Gold Answer	BC

(b) Example of RAG task. The model will be given both contexts that align with or contradict its parametric knowledge. It is expected to provide the answer based on both contexts.

Figure 12: Examples of each task.(cont)

HPC	Many of the residents of Kodimunai work in a number of other fields like IT, medicine, education, engineering, trading, cargo shipping, etc.
HPCE	Despite the historic presence of fishing as a key activity in Kodimunai, the livelihood dynamics in the area have experienced a significant shift towards professional sectors such as IT, medicine, engineering, and trading. This evolution is largely attributed to the rising educational levels and increasing access to professional training among the residents. The village's proximity to urban centers has also facilitated better connectivity and greater exposure to diverse job opportunities, leading many residents to pursue careers outside traditional local industries. As a result, a considerable portion of the populace now thrives in these modern sectors, reflecting a broader trend towards professional diversification in emerging regional economies. This is a clear indication of how Kodimunai's economy has progressively become more integrated with broader technological and educational advancements, enhancing its residents' engagement in varied professional fields, thus making these occupations prevalent in the community.

Figure 13: An example of HPC and HPCE.

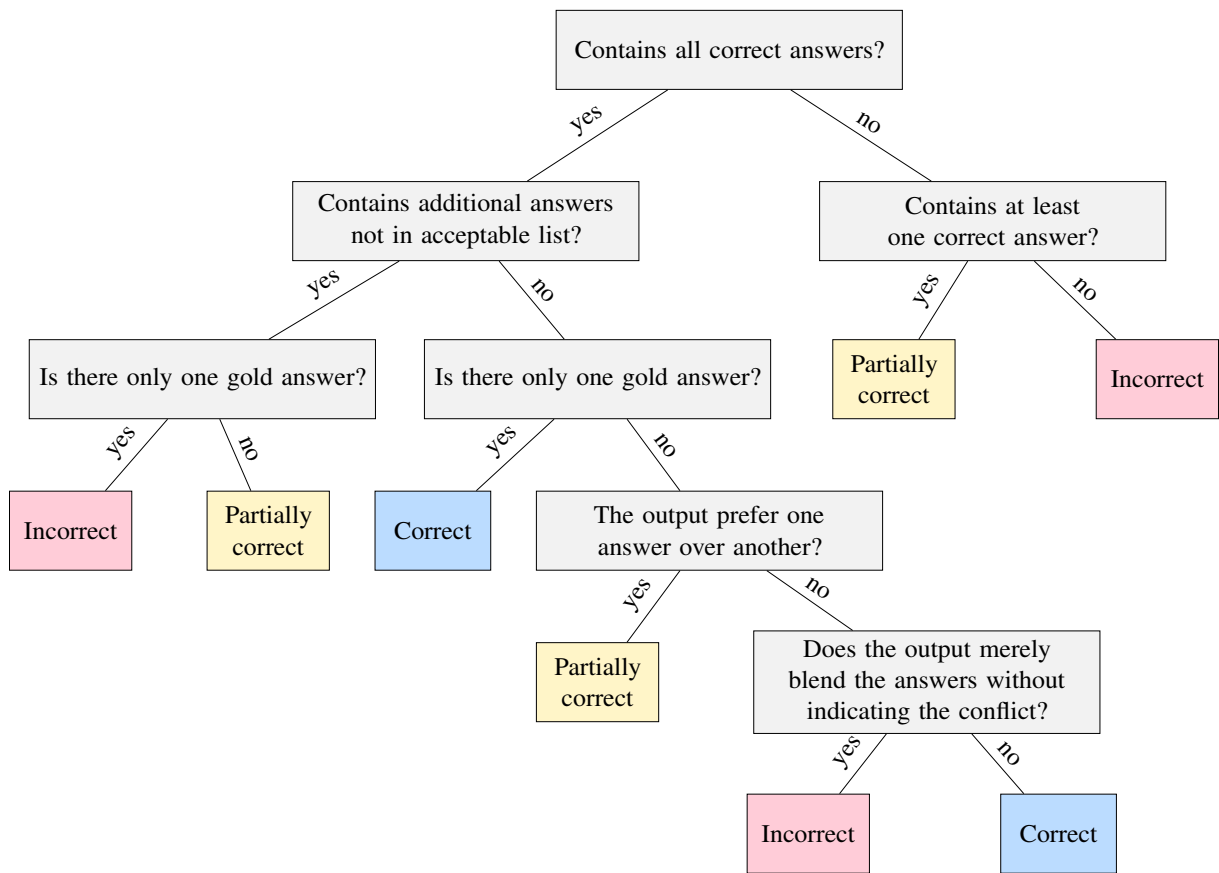


Figure 14: Decision tree for annotation guidance.

Free Generation Evaluation Prompt 1.

Do not fact-check or rely on world knowledge. Assume the reference (correct_answers) is correct.

- ↪ Please evaluate the response to a question under relaxed evaluation, where hallucinations, outdated information are allowed, as long as the primary answer is accurate. Each
- ↪ response is evaluated as correct, partially correct, or incorrect. The list of acceptable
- ↪ answers is separated by "|". You can discuss your reasoning after "comment: ", but you
- ↪ should include your final judgement after "evaluation: ".

Decision Rules:

* correct

- * correct means the response accurately matches with all the answers in the given correct answer list, or the correct answer can be inferred from the response.
- * If more than one correct answer is given, the response can contain contradictory answers
 - ↪ that reflect the heterogeneous nature of the context, or the response aggregates the
 - ↪ different answers and indicates that each answer is possible.
- * If only one correct answer is given, the response will be considered as correct as long as
 - ↪ the response aligns with the correct answer.
- * In addition, the correct response should not indicate a preference for one answer over
 - ↪ another.
- * The correct response should not combine two different correct answers without indicating the
 - ↪ contradictory nature of these answers.

* partially correct

- * partially correct means the response only matches with one of the answers in the correct answer list;
- * or the response matches with all the answers in the correct answer list but it indicates a
 - ↪ preference for one answer over another;
- * or the response matches with all the answers in the correct answer list but it contains
 - ↪ additional answers that doesn't match with any of the correct answers.

* incorrect

- * incorrect means the response does not match with any of the correct answers in the correct answer list,
- * or the response merely combines two contradictory answers from the correct answer list and
 - ↪ indicates that both of them are possible. Note that for each question, there may be
 - ↪ multiple correct answers based on different sources even though these correct answers
 - ↪ contradict each other.
- * For answers that involve names of entities (e.g., people), complete names or commonly
 - ↪ recognized names are expected.
- * Regarding numerical answers, approximate numbers are generally not accepted unless
 - ↪ explicitly included in the ground-truth answers.
- * If the instance has only one correct answer, and the response gives additional answers that
 - ↪ are not on the correct answer list, it will be considered as incorrect.

The primary or final answers when standing alone must be accurate. Any additional information that

- ↪ is provided must not contradict the primary answers or reshape one's perception of them.
- ↪ We accept responses that contain hallucinated or outdated information that does not
- ↪ significantly impact the primary answers.

Examples

Example 1

question: How old is the world's oldest verified living person?
correct_answers: 115 years old | 117 years old
response: As of today the most up-to-date and relevant information regarding this query is as

- ↪ follows. The oldest verified living person is Maria Branyas Morera, who was born on March
- ↪ 4, 1907, making her 117 years old.

comment: The primary answer in the response (117 years old) is partially correct as it only

- ↪ matched with with one of the correct answers (117 years old) and it doesn't mention
- ↪ another answer (115 years old). Thus, the response is partially correct.

evaluation: partially correct

Example 2

question: How old is the world's oldest verified living person?
correct_answers: 115 years old | 117 years old
response: According to the context, the oldest verified living person is Maria Branyas Morera,

- ↪ who is both 115 years old and 117 years old.

comment: Although the primary answer contains all correct answers (115 years old and 117 years

- ↪ old) that matches with the correct answers, it is logically incorrect because a person
- ↪ cannot have two ages at the same time. Thus, the response is incorrect.

evaluation: incorrect

Figure 15: Final Prompt for model based evaluation. Only the best performing candidate is shown due to space constraints. The remaining candidate prompts can be found in the repository.

Free Generation Evaluation Prompt 1.

Example 3

question: How old is the world's oldest verified living person?,
correct_answers: 115 years old | 117 years old
response: According to the context, one source claims that the oldest verified living person
↪ is Maria Branyas Morera, who is 117 years old. However, another source claims that she is
↪ 115 years old.
comment: The answer contains all correct answers: 115 years old and 117 years old, and it
↪ points out that these two answers are from different sources, which is logically possible.
↪ Thus, the response is correct.
evaluation: correct

Example 4

question: How old is the world's oldest verified living person?, correct_answers: 115 years
↪ old | 117 years old
response: According to the context, the oldest verified living person is Maria Branyas Morera,
↪ who is either 117 years old or 115 years old.
comment: The answer contains all correct answers: 115 years old and 117 years old that
↪ aggregates the different answers from different sources, which is logically possible. Thus,
↪ the response is correct.
evaluation: correct

Example 5

question: How many books has Colleen Hoover published in 2020?
correct_answers: 26 books | 27
response: according to some sources, Colleen Hoover has published 26 books in 2020. However,
↪ other sources claim that it has published 27 books in 2020.
comment: The response accurately provides two answers (26 books and 27 books) based on
↪ different sources. It recognized the two sources that provide different answers. These two
↪ answers match with all correct answers (26 books, 27). Thus, the response is correct.
evaluation: correct

Example 6

question: How many books has Colleen Hoover published in 2020?
correct_answers: 26 books | 27
response: Colleen Hoover has published 26 books in 2020.
comment: The response only provides one answers (26 books) based on the given context, which
↪ matched with one of the answers in the correct answer list. Thus, the response is
↪ partially correct.
evaluation: partially correct

Example 7

question: How many member nations exist in the United Nations?
correct_answers: 193 members | 194 member states response: As of my last update in April 2023,
↪ there were 191 member nations in the United Nations. Please note that this information
↪ may have changed since then, so it is advisable to check the latest information from the
↪ United Nations or other reliable sources.
comment: The primary answer in the response (191 member nations) does not match with any of
↪ the correct answers (193 member states or 194 members). Thus, the response is incorrect.
evaluation: incorrect

Example 8

question: How many member nations exist in the United Nations?
correct_answers: 193 members | 194 member states
response: Although some sources claim that there are 193 member nations in the United Nations,
↪ the UN is made up of 194 member states. 24
comment: Although the primary answer mentions two answers (193 members, 104 member states)
↪ that match with the correct answers, the response indicates a preference of one answer
↪ (194 member states) over the other (193 members). Thus, the response is partially correct.
↪
evaluation: partially correct

Figure 16: Final Prompt for model based evaluation (cont).

Free Generation Evaluation Prompt 1.

Example 9

question: How many member nations exist in the United Nations?
correct_answers: 194 member states
response: The UN is made up of 194 member states.
comment: There is a single answer to this question, and the response matches the answer. Thus,
↪ the response is correct.
evaluation: correct

Example 10

question: How many member nations exist in the United Nations?
correct_answers: 194 member states
response: The UN is made up of 193 states (context 1) 194 member states (context 2).
comment: There is a single answer to this question, although one of the response matches the
↪ answer, the response gives an additional answer that is not correct. Thus, the response is
↪ incorrect.
evaluation: incorrect

Example 11

question: Was Newton the first to propose calculus?
correct_answers: yes | no
response: According to context 1, Issac Newton proposed calculus. However, context 2 states
↪ that Leibniz was the first to propose calculus. Therefore, it could be Newton or Leibniz
↪ based on the contexts.
comment: The correct answer yes | no is not given by the response, but we can infer that the
↪ response points both "yes" and "no" answer depending on the context it was relying. For
↪ context 2, by stating that Leibniz was the first to propose calculus, the response hinders
↪ an answer no, while the response point that context 1 lead to the answer of yes.
evaluation: correct

Figure 17: Final Prompt for model based evaluation (cont 2).