

Task Matters: Knowledge Requirements Shape LLM Responses to Context–Memory Conflict

Kaiser Sun Fan Bai Mark Dredze

Center for Language and Speech Processing, Data Science and AI Institute
Johns Hopkins University
Baltimore, MD USA

hsun74@cs.jhu.edu fbai3@jh.edu mdredze@cs.jhu.edu

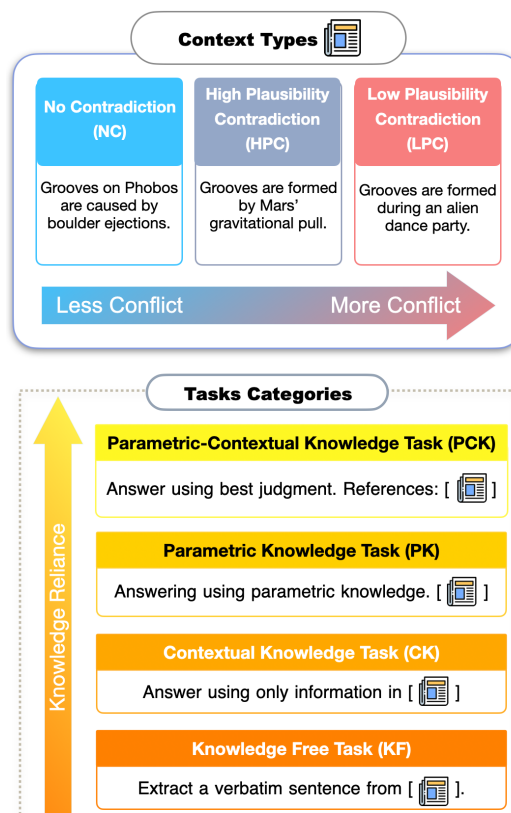
Abstract

Large language models (LLMs) draw on both contextual information and parametric memory, yet these sources can conflict. Prior studies have largely examined this issue in contextual question answering, implicitly assuming that tasks should rely on the provided context, leaving unclear how LLMs behave when tasks require different types and degrees of knowledge utilization. We address this gap with a model-agnostic diagnostic framework that holds underlying knowledge constant while introducing controlled conflicts across tasks with varying knowledge demands. Experiments on representative open-weight and proprietary LLMs show that performance degradation under conflict is driven by both task-specific knowledge reliance and conflict plausibility; that strategies such as rationales or context reiteration increase context reliance, helping context-only tasks but harming those requiring parametric knowledge; and that these effects bias model-based evaluation, calling into question the reliability of LLMs as judges. Overall, our findings reveal that context–memory conflict is inherently task-dependent and motivate task-aware approaches to balancing context and memory in LLM deployment and evaluation.¹

1 Introduction

Large language models (LLMs) perform well on many knowledge-centric tasks because they encode vast amounts of parametric knowledge. In many practical settings, however, the necessary facts are supplied directly by the user in the prompt, which often includes information that is updated after the model’s knowledge cutoff. *Context–memory conflict* arises when such input contradicts what the model “knows,” and LLMs often favor their own parametric knowledge over the provided context

¹Our framework and data are available at github.com/KaiserWhoLearns/LLM-KnowledgeConflict-TaskMatters.



* Context types are supplied to models under different task constraints.

Figure 1: Overview of the types of contexts and tasks in our evaluation. Context types vary in the level of conflict, while the tasks impose different knowledge constraints.

(Longpre et al., 2021; Chen et al., 2022; Xie et al., 2023; Jin et al., 2024a; Liu et al., 2025).

Prior work has shown that context–memory conflict substantially affects model behavior, but the resulting conclusions remain fragmented and do not yet form a coherent explanation. Some studies find that models systematically favor their parametric knowledge, particularly when conflicting entities are familiar (Longpre et al., 2021), while others report that models instead follow contextual evidence when it is sufficiently coherent or convincing (Xie et al., 2023; Jin et al., 2024a). Most of these results, however, are derived from contextual

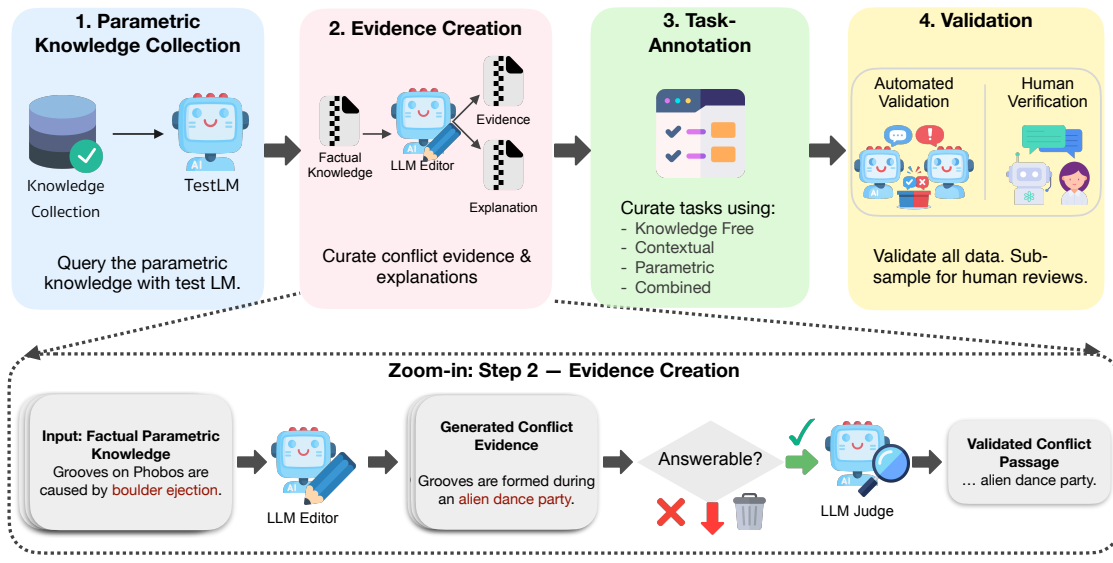


Figure 2: Overall diagnostic data creation flow. The lower portion is a zoom in of Evidence Creation step. After collecting the test model’s parametric knowledge, the supporting passages are further edited to reveal multiple levels of conflict (2. Evidence Creation) and appear in different tasks (3. Task-Annotation).

question answering, where models are expected to rely exclusively on the provided passage. As a result, it remains unclear how LLMs should behave under context–memory conflict when tasks require different forms of knowledge utilization (Xu et al., 2024).

In practice, tasks vary sharply in their reliance on contextual versus parametric knowledge. Extractive tasks, such as text copying, impose little need for prior knowledge, whereas tasks such as literature review require integrating background knowledge with new information published after the model’s knowledge cutoff. Between these extremes lie tasks that demand paraphrasing, selective grounding, or explicit reconciliation of conflicting evidence. Treating all such settings as interchangeable obscures how context–memory conflict functions across real-world applications.

In this work, we show that **the impact of context–memory conflict is task-dependent**. We demonstrate that performance degradation under conflict correlates with a task’s knowledge requirements, in addition to the level of conflict alone. To make this dependence explicit, we introduce a diagnostic framework that holds underlying knowledge constant while systematically varying task formulations. This allows us to isolate how the same conflict affects models when a task requires (i) no knowledge beyond surface extraction, (ii) grounding in context, (iii) reliance on parametric knowledge, or (iv) integration of both. Our framework automatically identifies a model’s parametric

beliefs and injects controlled contradictions into downstream tasks (Figure 1), producing model-specific diagnostic datasets that vary both conflict level and task knowledge requirements. Evaluating models on these datasets enables us to disentangle the effects of knowledge conflict from task demands. This allows us to quantify how knowledge conflict disrupts model behavior in a task-aware manner (Figure 2).

Across representative LLMs, we find that the effects of context–memory conflict are strongly task-dependent: conflicts have minimal impact on tasks requiring little or no knowledge utilization, yet significantly impair knowledge-intensive tasks, even under explicit instructions. Task type explains prior findings that would otherwise appear inconsistent. Beyond overall performance, we find that simple strategies such as reiteration encourage context reliance, without the need to alter the inference mechanism. Reiteration improves performance when the task is intended to rely primarily on contextual information (Wang et al., 2022; Wu et al., 2024), but degrades performance when successful task completion requires dominant use of parametric knowledge. Finally, we demonstrate that studying context–memory conflict particularly matters for the task of model-based evaluation. Using LLMs as judges (Zheng et al., 2023; Liu et al., 2023; Ru et al., 2024; Chen et al., 2025) constitutes a task in which the model is required to balance parametric knowledge and contextual information, while the model’s own parametric knowledge can implicitly

bias evaluations, particularly when factual correctness conflicts with the provided context. Together, our findings suggest that context–memory conflict cannot be meaningfully studied or mitigated without explicit consideration of task knowledge requirements, and motivate task-aware approaches to balancing contextual and parametric knowledge.

2 Related Work

Context-Memory 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). We focus on the context-memory conflict, which arises when a given information-bearing text chunk contradicts the model’s parametric beliefs.

Nuanced Behaviors under Conflict Early studies reported that models tend to rely on their own knowledge when the prompt provides contradictory evidence (Longpre et al., 2021; Chen et al., 2022). Later work revealed a more nuanced picture. On synthetic datasets, Xie et al. (2023) showed that LLMs often update their answers when given strong and convincing evidence, whereas Jin et al. (2024a) observed a “Dunning–Kruger” effect in stronger LLMs, which display higher confidence in their incorrect parametric knowledge than in the external context. Further analysis also finds that models show availability bias (leaning on common-knowledge facts), majority bias (trusting the answer supported by more frequent evidence across documents), and confirmation bias (preferring evidence consistent with their prior knowledge), especially when the models are given misleading or irrelevant answers. Moving to realistic documents, Kortukov et al. (2024) found that models update their answers more reliably than synthetic evaluations suggest, yet still exhibit a *parametric bias*: if the model’s originally believed answer appeared anywhere in the context (even as a distractor), the model was more likely to stick to that incorrect answer. Complementary evidence from *intra-context* conflict shows similar task-dependent behavior (Ying et al., 2024).

Mitigation Strategies Methods have also been proposed to alleviate context-memory knowledge conflict. Jin et al. (2024b) identified certain attention heads that specialize in “memory” while others specialize in “context”, and therefore pro-

pose a method that dynamically prunes or patches specific attention heads that cause conflicts. Li et al. (2025) instead propose a two-step inference intervention that has each step focus on either the context or the memory heads. Efforts have also been made to develop novel decoding methods that enhance the use of contextual knowledge (Jin et al., 2024a; Shi et al., 2024; Wang et al., 2025a). More recently, work has explored post-hoc, inference-time controls that steer context reliance without fine-tuning, including proxy-model steering (Wang et al., 2025b) and test-time attention interventions (Li et al., 2025).

Our Focus Most prior studies focus on contextual question answering, a setting that *requires* heavy reliance on the provided passages. Many other tasks, for example, grammar correction or claim verification, may need little context or, conversely, require careful integration of both parametric and contextual knowledge. This leaves the question of whether context-memory conflict poses the same impact on tasks with different knowledge demands unanswered. To fill this gap, we keep the underlying knowledge constant while varying the *task formulation*, creating controlled datasets that induce different conflict levels for each target model. We introduce an analysis tool that automatically constructs model-specific test sets. Our findings indicate that both knowledge-memory conflict and blindly following the context could be particularly harmful to model-based evaluations, further motivating *task-dependent* methods for dynamically balancing context and memory (Wang et al., 2025b; Li et al., 2025).

3 Context-Memory Conflict Creation

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). We use the knowledge from knowledge conflict question answering datasets (Wan et al., 2024; Hou et al., 2024) that have two or more acceptable answers to one question, using them to identify the stance aligned with the model’s parametric belief, which then serves as the basis for constructing task data. 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 under greedy decoding, while rejecting conflicting alternatives, with

details included in Appendix A. Yes/no validation prompts are known to elicit sycophantic confirmation, where the model agrees with whichever stance the user asserts rather than reporting its own belief, and this tendency is amplified by scale and instruction tuning (Perez et al., 2023; Sharma et al., 2023; Wei et al., 2023). Our requirement of consistency across multiple paraphrased prompts and explicit rejection of the contradictory alternative is therefore designed as a stance-consistency filter rather than a single-shot belief probe. This mitigates, but does not fully eliminate, sycophancy in the parametric-knowledge probe, and we treat the resulting set as a conservative lower bound on the model’s parametric beliefs. With the model’s internal knowledge, 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 (§3.2, Task-Annotation). Two LLMs then review each instance to verify the correctness of its task type annotation, with a subsample of the instances verified by human annotators (Validation, Appendix B).

3.1 Evidence Creation

The cognitive science literature suggests that humans resolve conflicts between prior knowledge and new information by judging the rationality of alternative concepts (Posner et al., 1982; Vosniadou and Brewer, 1992). Similarly, Xie et al. (2023) shows that LLMs may also revise their answers when the context is sufficiently convincing. We formalize this with the notion of *plausibility*, defined as the willingness to consider an alternative strategy when it is understood, coherent, relatively simple, and deemed a viable solution (Posner and Strike, 1992). Plausibility can be used to measure how likely a human is to accept new information in the presence of conflict. We decompose it into two criteria: alignment with *real-world or common-sense knowledge* and consistency with *basic logical principles*. For example, if a model believes that grooves on Phobos were caused by a boulder from an asteroid ejection, the claim that they resulted from Mars’s gravitational pull is plausible, as it fits common-sense knowledge. By contrast, attributing the grooves to a dance party is implausible. 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).

Evidence instances 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 the i -th instance, N is the size of dataset D_{orig} . The subscript j after i represents the j -th answer/context of the question q_i , as each question q_i may have multiple acceptable answers. 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 the original question q_i ; and (2) whether the generated context p_i^{LPC} is truly low-plausibility through a fact-checking process.

3.2 Task Annotation

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

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, 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 serves as both a complementary setting to PCK tasks and creates a test bed that is closer to real-world applications, as Hagström et al. (2024) show that insights based on synthesized data are not guaranteed to generalize to real-world scenarios.

The annotations for all five tasks are derived 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 KF tasks, the model is only expected to perform extractions. 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. To assess the performance of the target model, we report the accuracy for CK and PK tasks, F1 for KF, PCK, and RAG. To obtain high-quality texts, we use GPT-4o as the base model to create evidence and validate the diagnostic data. Then, we analyze the instruction-tuned version of Mistral-7B (Jiang et al., 2023), OLMo2-7B, OLMo2-13B (OLMo et al., 2024), Qwen2.5-7B, Qwen2.5-14B (Qwen et al., 2025), and the proprietary GPT-5.2,² spanning a range of model sizes, training paradigms, and openness. The resulting dataset statistics are presented in Appendix C.

4 Findings

4.1 Conflict Impairs Model Performance on Knowledge-Intensive Tasks

The performance of each model on each task type and context type is reported in Figure 3. A universal trend can be observed: in all but Knowledge-Free tasks, all models suffer when asked to provide responses that contradict their parametric knowledge.

Knowledge conflict degrades performance whenever knowledge is required. In CK tasks (Fig-

²gpt-5.2-2025-12-11, <https://openai.com/index/gpt-5-system-card-update-gpt-5-2/>.

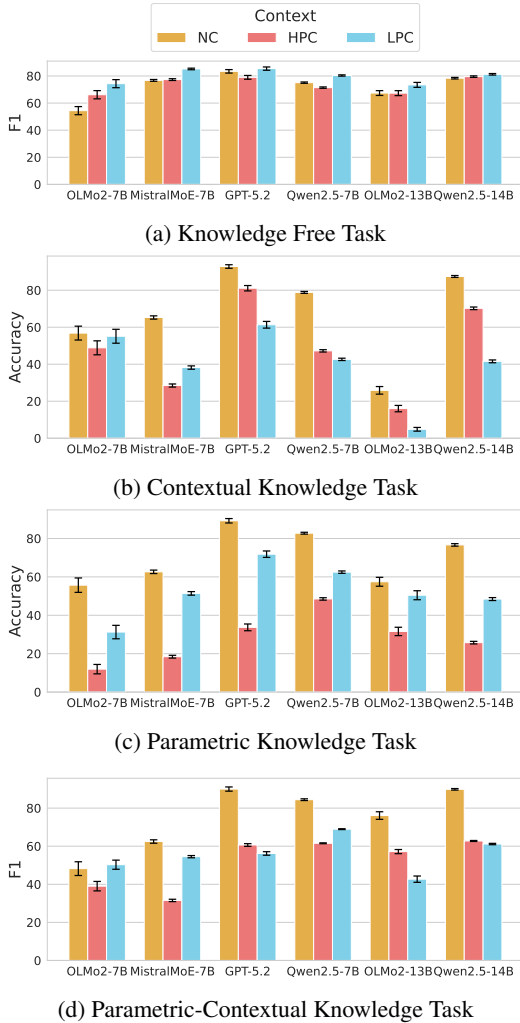


Figure 3: Performance of each model on different task types. A clear trend of $NC > HPC / LPC$ is shown across tasks involving knowledge utilization.

ure 3b), the model is explicitly instructed to ignore its own beliefs and rely solely on the given 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 pattern persists at the proprietary frontier: GPT-5.2 shows the largest absolute $NC-LPC$ gap on CK (≈ 31 points), indicating that scale and instruction-tuning maturity do not by themselves resolve context-memory conflict. 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., 2024a). This issue, if left untreated, could not only affect the overall performance but also the correctness of model-based evaluation results, which we illustrate in §4.3.

Similarly, we find that conflict still degrades performance when only parametric knowledge is required. Figure 3c examines model performance

under settings where only parametric knowledge is needed. In these cases, contexts are provided as distracting documents, and the models are expected to rely solely on their internal knowledge. 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 still making use of the context, even when instructed otherwise. To verify that this failure is not due to limitations in the model’s instruction-following ability, we vary the strength of instructions to control the absolute performance on each setting, and the results suggest that while the absolute performance may vary, the relative trends between each context type remain unchanged (Appendix F). Unsurprisingly, the lower the plausibility in the given context, the more likely the model is to follow its parametric knowledge, thus leading to higher performance. This effect is most pronounced for GPT-5.2, where PK accuracy drops from 89.2 (NC) to 33.7 (HPC) but partially recovers to 71.8 (LPC), confirming that high-plausibility distractors but not conflict per se drive the failure mode. This suggests that, although plausible contexts can lead to more context reliance, they can also be harmful when the underlying task itself expects less context reliance. Taken together, the results from both CK and PK tasks show that the impact of context-memory conflict strongly depends on the task’s intended knowledge reliance: the same contextual information can be either helpful or harmful depending on whether the task is designed to prioritize contextual grounding or parametric knowledge. However, the roles of the conflict are minimal when little knowledge is required to complete the tasks (KF task in Figure 3a), and some models even perform slightly better on LPC examples. Because KF primarily involves copying from the passage, LPC instances may reduce ambiguity by providing context that is clearly differentiated from the world’s factual knowledge.

Model	HPC	LPC	HPC - LPC
OLMo2-7B	43.3	46.3	-3.0
OLMo2-13B	60.9	65.8	-4.9
MistralMoE-7B	41.2	54.4	-13.2
Qwen2.5-7B	60.2	67.1	-6.9
Qwen2.5-14B	59.1	61.0	-1.8
GPT-5.2	59.7	66.6	-6.9

Table 1: Performance of the model on the RAG task.

More plausible \neq higher reliance. Hypothesizing that a perfect retriever can find all relevant doc-

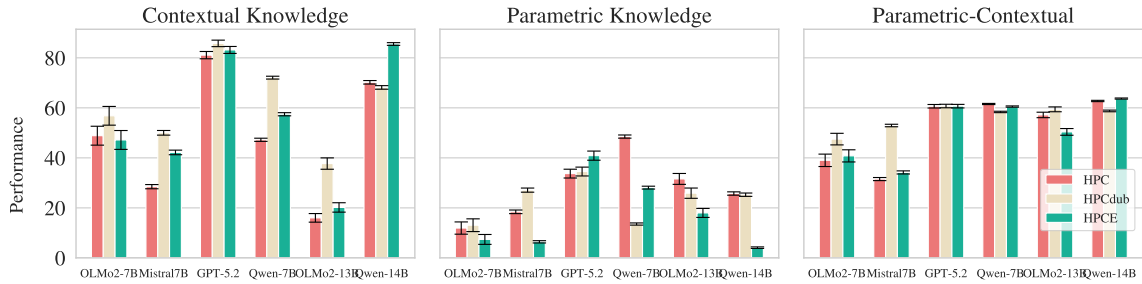


Figure 4: Performance on high plausibility contradiction instances with (HPCE) and without (HPC) explanations.

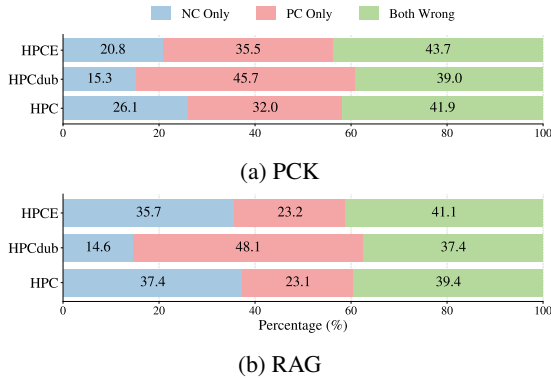


Figure 5: Averaged error distribution on RAG and PCK tasks. *NC Only* represents that the model only provides the NC answer; *PC Only* represents that the model only provides the PC answer; *Both Wrong* represents the case where the model provides neither PC or NC answer.

uments, we construct a RAG setting in which both model-aligned (NC) and contradictory (HPC or LPC) passages are presented simultaneously in the context. 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 Table 1. Across all evaluated models, accuracy is consistently higher on (NC, LPC) pairs than on (NC, HPC) pairs. Similarly, in CK and PCK tasks, models do not show a strong preference for higher plausibility passages. In contrast, for PK tasks, lower-plausibility passages enable models to rely more on their parametric knowledge rather than follow the context. This behavior is counterintuitive. Prior work suggests that language models tend to trust text that appears more credible or plausible (Chen et al., 2024), whereas our results indicate that such a heuristic does not strictly hold in the presence of explicit knowledge conflict. Moreover, when considering only the instances whose KF variants the model achieves performance on, the same behavior remains unchanged on instances where the model is highly confident (Appendix E.1), confirming our findings in this section.

4.2 Rationales and Reiteration

§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). Xie et al. (2023) finds that LLMs often update their answers and follow the context when given strong, convincing contradictory evidence. We study the effect of explanatory evidence 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 G. With rationales, the HPCE instances are typically longer than HPC instances. To ensure a fair comparison, we create an ablation setting, HPCdub, where the HPC context is repeated multiple times such that the context length is about the same as the HPCE instances (Figure 4).

Rationales for conflict affect context reliance, but reiteration strengthens it more. Including rationales benefits the model in CK tasks, where correct predictions require strict reliance on the context. In contrast, for PK tasks, rationales have a detrimental effect: while explanatory instances increase context reliance, they can also act as strong distractors that pull the model away from its parametric knowledge. Interestingly, when the same evidence is reiterated in the context (HPCdub), models improve on CK tasks without being overly distracted from the parametric knowledge in PK tasks. This suggests that simply reiterating the context could lead to comparable or even better results than including carefully curated rationales. To better understand this effect, we analyze the errors in PCK tasks and RAG tasks in Figure 5 and Ap-

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	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.
Judge 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.

pendix E. We find that, although reiteration reduces overall errors, the remaining mistakes disproportionately favor answers that appear more frequently in the context. While this behavior partly reflects a form of majority bias, where models preferentially produce answers that appear more frequently in the context (Gupta et al., 2023), it also suggests a deeper connection to prior findings from the language model security literature. Specifically, studies on prompt-based attacks show that reiteration alone can substantially amplify a model’s internal belief or response preference. This explains why reiterated context can be as influential as explicit rationales and is beneficial in context-driven tasks. However, in tasks that require parametric knowledge, overly persuasive contextual signals can suppress the memory and lead the model away from the correct solution. Together, these observations underscore that context reliance is inherently task-dependent, and motivate explicit context–memory balancing mechanisms, such as selectively controlling attention to contextual tokens (Jin et al., 2024b; Li et al., 2025) or proxy-model steering (Wang et al., 2025b).

4.3 Conflict leads to unreliable judge LLMs

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. One of the source data of our dataset, Wiki-Contradict (Hou et al., 2024), employs a language

model as a judge to decide whether the free-text answer aligns with 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? To answer this question, we create a free generation version of our diagnostic framework following Hou et al. (2024) 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 H.1. We find that the averaged Cohen’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 a strong proprietary model (GPT-4o) would also lean towards its own parametric knowledge. An example of such an instance is presented in Figure 6, where GPT-4o fails to adhere to the instruction and refuses to grade a contextually correct but factually incorrect output as correct. Crucially, this risk does not vanish with stronger models: GPT-5.2 retains the largest number of instances after our parametric-knowledge filter (Appendix C) and shows the steepest PK degradation under conflict, suggesting that the judge-bias problem may be amplified, not resolved, by frontier-scale instruction tuning. One may consider employing a conflict alleviation technique to enforce stronger context reliance, but blindly following the context could also increase the risk of prompt injection (Perez and Ribeiro, 2022; Greshake et al., 2023). Our findings highlight the risks of using language models as

evaluators, since a model’s parametric knowledge can bias its judgments, thus leading to inaccurate evaluation results.

5 Conclusion

LLMs must balance parametric and contextual knowledge, yet how they do so under conflict depends on the task at hand. In this work, we show that the impact of context–memory conflict is task-dependent: conflicts have little effect on tasks requiring minimal knowledge utilization, but substantially impair performance on knowledge-intensive tasks. By holding underlying knowledge constant while varying task formulations, our framework offers a unified explanation for previously fragmented findings in the literature.

Our results further highlight that increasing context reliance through strategies such as reiteration could be beneficial, but can also be harmful when parametric knowledge should dominate. Moreover, we show that task-dependent knowledge conflict undermines the reliability of model-based evaluation, as LLM judges may be systematically biased by their own parametric knowledge. Together, these findings suggest that both blindly enforcing context adherence and trusting only parametric knowledge can be detrimental, motivating mechanisms that dynamically balance contextual and parametric knowledge at inference time (Wang et al., 2025b; Li et al., 2025).

Limitations

Potential Knowledge Conflict in Instance Creation Our diagnostic instances are partly generated or edited with the assistance of 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 LLMs are used to edit such instances, even when the original context was authored by humans. 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.

Model Coverage. Our analysis spans five open-weight models (Mistral-7B, OLMo2-7B/13B, Qwen2.5-7B/14B) and one proprietary frontier model (GPT-5.2). While this range covers diverse

training paradigms and a roughly $20\times$ variation in scale, broader coverage of other proprietary frontier models is left to future work.

Definition of Knowledge. In NLP studies, knowledge is usually framed as factual or propositional content (Lewis et al., 2020; Chen et al., 2022; Meng et al., 2022; 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. This unsettled and multifaceted definition of knowledge poses challenges for reliably quantifying the degree of knowledge involvement in tasks, complicating efforts to systematically study knowledge conflicts.

Ethical Considerations

This work analyzes how large language models resolve conflicts between contextual information and parametric knowledge. While our framework is diagnostic and does not introduce new model capabilities, it highlights several ethical considerations related to the evaluation and deployment of models.

Our findings show that LLMs used as evaluators can be systematically biased by their own parametric knowledge when judging outputs that conflict with the provided context, raising concerns about the reliability of model-based evaluation. In addition, strategies that indiscriminately increase context reliance may be harmful for tasks that require parametric knowledge and may increase vulnerability to prompt injection.

The datasets constructed in this work are derived from existing human-cleaned benchmarks and synthetic edits, and do not involve personal, sensitive, or human-subject data. Overall, our results emphasize the need for task-aware evaluation and deployment practices that explicitly account for differing knowledge requirements across tasks.

Acknowledgment

This research was supported in part by DARPA (SciFY). The U.S. Government is authorized to reproduce and distribute reprints for Governmental

purposes. The views and conclusions contained in this publication are those of the authors and should not be interpreted as representing official policies or endorsements of DARPA or the U.S. Government.

The authors would also like to thank Niyati Bafna, Krithika Ramesh, and members of the Dredze Lab for their helpful feedback.

References

- Guiming Hardy Chen, Shunian Chen, Ziche Liu, Feng Jiang, and Benyou Wang. 2024. [Humans or LLMs as the judge? a study on judgement bias](#). In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 8301–8327, Miami, Florida, USA.
- 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.
- Nuo Chen, Zhiyuan Hu, Qingyun Zou, Jiaying Wu, Qian Wang, Bryan Hooi, and Bingsheng He. 2025. [Judgelrm: Large reasoning models as a judge](#). *arXiv preprint arXiv:2504.00050*.
- Kai Greshake, Sahar Abdelnabi, Shailesh Mishra, Christoph Endres, Thorsten Holz, and Mario Fritz. 2023. [Not what you’ve signed up for: Compromising real-world llm-integrated applications with indirect prompt injection](#). In *Proceedings of the 16th ACM workshop on artificial intelligence and security*, pages 79–90.
- Daya Guo, Dejian Yang, Haowei Zhang, Junxiao Song, Ruoyu Zhang, Runxin Xu, Qihao Zhu, Shirong Ma, Peiyi Wang, Xiao Bi, et al. 2025. [Deepseek-r1: Incentivizing reasoning capability in llms via reinforcement learning](#). *arXiv preprint arXiv:2501.12948*.
- Karan Gupta, Sumegh Roychowdhury, Siva Rajesh Kasa, Santhosh Kumar Kasa, Anish Bhanushali, Nikhil Pattisapu, and Prasanna Srinivasa Murthy. 2023. [How robust are llms to in-context majority label bias?](#) *arXiv preprint arXiv:2312.16549*.
- 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*.
- Yufang Hou, Alessandra Pascale, Javier Carnerero-Cano, Tigran Tchrakian, Radu Marinescu, Elizabeth Daly, Inkit Padhi, and Prasanna Sattigeri. 2024. [Wicontradict: 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. 2024a. [Tug-of-war between knowledge: Exploring and resolving knowledge conflicts in retrieval-augmented language models](#). *arXiv preprint arXiv:2402.14409*.
- Zhuoran Jin, Pengfei Cao, Hongbang Yuan, Yubo Chen, Jiexin Xu, Huaijun Li, Xiaojian Jiang, Kang Liu, and Jun Zhao. 2024b. [Cutting off the head ends the conflict: A mechanism for interpreting and mitigating knowledge conflicts in language models](#). In *Findings of the Association for Computational Linguistics: ACL 2024*, pages 1193–1215, Bangkok, Thailand.
- Evgenii Kortukov, Alexander Rubinstein, Elisa Nguyen, and Seong Joon Oh. 2024. [Studying large language model behaviors under context-memory conflicts with real documents](#). *arXiv preprint arXiv:2404.16032*.
- 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.
- Gaotang Li, Yuzhong Chen, and Hanghang Tong. 2025. [Taming knowledge conflicts in language models](#). *arXiv preprint arXiv:2503.10996*.
- Xiaoyuan Liu, Wenxuan Wang, Youliang Yuan, Jentse Huang, Qiuzhi Liu, Pinjia He, and Zhaopeng Tu. 2025. [Insight over sight: Exploring the vision-knowledge conflicts in multimodal LLMs](#). In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 17825–17846, Vienna, Austria. Association for Computational Linguistics.
- Yang Liu, Dan Iter, Yichong Xu, Shuhang 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.
- Shayne Longpre, Kartik Perisetla, Anthony Chen, Nikhil Ramesh, Chris DuBois, and Sameer Singh. 2021. [Entity-based knowledge conflicts in question](#)

- answering. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 7052–7063, Online and Punta Cana, Dominican Republic.
- Alex Mallen, Akari Asai, Victor Zhong, Rajarshi Das, Daniel Khashabi, and Hannaneh Hajishirzi. 2023. [When not to trust language models: Investigating effectiveness of parametric and non-parametric memories](#). In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 9802–9822, Toronto, Canada.
- Kevin Meng, David Bau, Alex Andonian, and Yonatan Belinkov. 2022. Locating and editing factual associations in gpt. *Advances in neural information processing systems*, 35:17359–17372.
- Team OLMo, Pete Walsh, Luca Soldaini, Dirk Groeneveld, Kyle Lo, Shane Arora, Akshita Bhagia, Yuling Gu, Shengyi Huang, Matt Jordan, et al. 2024. 2 olmo 2 furious. *arXiv preprint arXiv:2501.00656*.
- OpenAI. 2024. [Gpt-4o: Openai’s new flagship model](#). Accessed: 2025-05-19.
- Ethan Perez, Sam Ringer, Kamile Lukosiute, Karina Nguyen, Edwin Chen, Scott Heiner, Craig Pettit, Catherine Olsson, Sandipan Kundu, Saurav Kadavath, Andy Jones, Anna Chen, Benjamin Mann, Brian Israel, Bryan Seethor, Cameron McKinnon, Christopher Olah, Da Yan, Daniela Amodei, Dario Amodei, Dawn Drain, Dustin Li, Eli Tran-Johnson, Guro Khundadze, Jackson Kernion, James Landis, Jamie Kerr, Jared Mueller, Jeeyoon Hyun, Joshua Landau, Kamal Ndousse, Landon Goldberg, Liane Lovitt, Martin Lucas, Michael Sellitto, Miranda Zhang, Neerav Kingsland, Nelson Elhage, Nicholas Joseph, Noemi Mercado, Nova DasSarma, Oliver Rausch, Robin Larson, Sam McCandlish, Scott Johnston, Shauna Kravec, Sheer El Showk, Tamera Lanham, Timothy Telleen-Lawton, Tom Brown, Tom Henighan, Tristan Hume, Yuntao Bai, Zac Hatfield-Dodds, Jack Clark, Samuel R. Bowman, Amanda Askell, Roger Grosse, Danny Hernandez, Deep Ganguli, Evan Hubinger, Nicholas Schiefer, and Jared Kaplan. 2023. [Discovering language model behaviors with model-written evaluations](#). In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 13387–13434, Toronto, Canada. Association for Computational Linguistics.
- Fábio Perez and Ian Ribeiro. 2022. Ignore previous prompt: Attack techniques for language models. *arXiv preprint arXiv:2211.09527*.
- George J Posner and Kenneth A Strike. 1992. A revisionist theory of conceptual change. *Philosophy of science, cognitive psychology, and educational theory and practice*, 147.
- George J Posner, Kenneth A Strike, Peter W Hewson, William A Gertzog, et al. 1982. Accommodation of a scientific conception: Toward a theory of conceptual change. *Science education*, 66(2):211–227.
- Qwen, :, An Yang, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chengyuan Li, Dayiheng Liu, Fei Huang, Haoran Wei, Huan Lin, Jian Yang, Jianhong Tu, Jianwei Zhang, Jianxin Yang, Jiayi Yang, Jingren Zhou, Junyang Lin, Kai Dang, Keming Lu, Keqin Bao, Kexin Yang, Le Yu, Mei Li, Mingfeng Xue, Pei Zhang, Qin Zhu, Rui Men, Runji Lin, Tianhao Li, Tianyi Tang, Tingyu Xia, Xingzhang Ren, Xuancheng Ren, Yang Fan, Yang Su, Yichang Zhang, Yu Wan, Yuqiong Liu, Zeyu Cui, Zhenru Zhang, and Zihan Qiu. 2025. [Qwen2.5 technical report](#).
- Dongyu Ru, Lin Qiu, Xiangkun Hu, Tianhang Zhang, Peng Shi, Shuaichen Chang, Cheng Jiayang, Cunxiang Wang, Shichao Sun, Huanyu Li, et al. 2024. Ragchecker: A fine-grained framework for diagnosing retrieval-augmented generation. *Advances in Neural Information Processing Systems*, 37:21999–22027.
- Mrinank Sharma, Meg Tong, Tomasz Korbak, David Duvenaud, Amanda Askell, Samuel R Bowman, Newton Cheng, Esin Durmus, Zac Hatfield-Dodds, Scott R Johnston, et al. 2023. Towards understanding sycophancy in language models. *arXiv preprint arXiv:2310.13548*.
- Weijia Shi, Xiaochuang Han, Mike Lewis, Yulia Tsvetkov, Luke Zettlemoyer, and Wen-tau Yih. 2024. [Trusting your evidence: Hallucinate less with context-aware decoding](#). In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 2: Short Papers)*, pages 783–791, Mexico City, Mexico.
- Stella Vosniadou and William F Brewer. 1992. Mental models of the earth: A study of conceptual change in childhood. *Cognitive psychology*, 24(4):535–585.
- Alexander Wan, Eric Wallace, and Dan Klein. 2024. [What evidence do language models find convincing?](#) In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 7468–7484, Bangkok, Thailand.
- Han Wang, Archiki Prasad, Elias Stengel-Eskin, and Mohit Bansal. 2025a. [AdaCAD: Adaptively decoding to balance conflicts between contextual and parametric knowledge](#). 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 11636–11652, Albuquerque, New Mexico. Association for Computational Linguistics.
- Xuezhi Wang, Jason Wei, Dale Schuurmans, Quoc Le, Ed Chi, and Denny Zhou. 2022. Rationale-augmented ensembles in language models. *arXiv preprint arXiv:2207.00747*.
- Yilin Wang, Heng Wang, Yuyang Bai, and Minnan Luo. 2025b. [Continuously steering LLMs sensitivity to](#)

contextual knowledge with proxy models. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pages 4682–4698, Suzhou, China. Association for Computational Linguistics.

Jerry Wei, Da Huang, Yifeng Lu, Denny Zhou, and Quoc V Le. 2023. Simple synthetic data reduces sycophancy in large language models. *arXiv preprint arXiv:2308.03958*.

Yexin Wu, Zhuosheng Zhang, and Hai Zhao. 2024. Mitigating misleading chain-of-thought reasoning with selective filtering. *arXiv preprint arXiv:2403.19167*.

Jian Xie, Kai Zhang, Jiangjie Chen, Renze Lou, and Yu Su. 2023. Adaptive chameleon or stubborn sloth: Revealing the behavior of large language models in knowledge conflicts. In *The Twelfth International Conference on Learning Representations*.

Rongwu Xu, Zehan Qi, Zhijiang Guo, Cunxiang Wang, Hongru Wang, Yue Zhang, and Wei Xu. 2024. Knowledge conflicts for LLMs: A survey. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 8541–8565, Miami, Florida, USA.

Jiahao Ying, Yixin Cao, Kai Xiong, Long Cui, Yidong He, and Yongbin Liu. 2024. Intuitive or dependent? investigating LLMs’ behavior style to conflicting prompts. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 4221–4246, Bangkok, Thailand. Association for Computational Linguistics.

Linda Zagzebski. 1999. "what is knowledge?". In John Greco and Ernest Sosa, editors, *The Blackwell Guide to Epistemology*, pages 92–116. Wiley-Blackwell.

Lianmin Zheng, Wei-Lin Chiang, Ying Sheng, Siyuan Zhuang, Zhanghao Wu, Yonghao Zhuang, Zi Lin, Zhuohan Li, Dacheng Li, Eric Xing, et al. 2023. Judging llm-as-a-judge with mt-bench and chatbot arena. *Advances in Neural Information Processing Systems*, 36:46595–46623.

A Parametric Knowledge Query

We query for the parametric knowledge with multiple prompts. For a single instance $(q_i, \{a_{i1}, a_{i2}\}, \{c_{i1}, c_{i2}\})$ 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 asked 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 Verification of Data Validity

In the final stage of the diagnostic data creation flow (Validation in Figure 2), all instances are verified by two language models and subsampled for human verification. The number of instances removed during this process is reported in Table 3.

Model Verification Each (question, evidence, answer) triple was validated by GPT-4o and DeepSeek-R1-Distill-Llama-70B (Guo et al., 2025). Both models were required to answer the question using the provided gold evidence. An instance was retained only if both predictions matched the annotated answer.

Human Verification To further ensure data quality, a subset of model-verified instances was manually reviewed. For each test model (OLMo, Mistral, and Qwen), 30 randomly selected instances were manually reviewed. Annotators verified (1) whether the evidence supported its answer, and (2) whether the gold answer was correct for the corresponding task (e.g., CK answer for CK task; PK answer for PK task). All sampled instances passed human verification, likely reflecting the strictness of the preceding model-based filter. The annotators were graduate students in NLP with prior annotation experience. They were compensated at standard research assistant rates.

C Dataset Statistics

Because each model encodes different parametric knowledge, the resulting diagnostic datasets differ across models. The overall statistics are reported in Table 2. Each question is paired with four evidence conditions (NC, HPC, HPCE, LPC), so the effective dataset size is four times the number of base instances.

Model	# Instances
Mistral-7B	2,893
OLMo2-7B	177
OLMo2-13B	456
Qwen2.5-7B	6,217
Qwen2.5-14B	4,250
GPT-5.2	7,250

Table 2: Number of instances of the resulting data for each model.

D Task Examples

Examples of each task are provided in Figure 7 and Figure 8.

E Raw Performance

We measure both F1 and exact match of each setting. The full performance of each model on the diagnostic data is shown in Table 4.

E.1 Highly Confident Instances

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 5: Performance of models on highly confident instances.

When querying for the model’s parametric knowledge (parametric knowledge collection in fig. 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 (fig. 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 performance of each model on only the highly confident instances is included in Table 5. The results confirm that while the absolute

numbers vary slightly, the overall trends observed in the broader dataset persist.

F Instruction Strength

To further disentangle the role of instruction-following ability from conflict-resolution biases, we conducted an additional experiment varying the strength of instructions with Mistral-7B and OLMo-7B. Specifically, we applied three levels of instruction forcefulness:

- **Strong:** “You **MUST** strictly and exclusively use . . .”
- **Neutral:** “Answer the question based only on . . .”
- **Weak:** “Try to answer based on . . .”

The average performance on each strength level is shown in Table 6, and the fine-grained performance on each context and task type is shown in Table 7. Instruction strength influences absolute performance, but the impact varies across models. Mistral-7B remains relatively stable under different prompt formulations, whereas OLMo-7B shows sharp degradation under strong, restrictive instructions. Despite shifts in absolute scores, the relative patterns reported in §4 remain consistent. Performance differences across tasks and evidence types are preserved, and the ordering of task/evidence effects holds across weak, neutral, and strong prompts.

Model	Weak	Neutral	Strong
Mistral-7B	45.66	46.49	49.48
OLMo2-7B	45.21	44.52	30.83

Table 6: Average performance on each prompt strength.

Stage	Mistral	OLMo2-7B	OLMo2-13B	Qwen2.5-7B	Qwen2.5-14B	GPT-5.2
Parametric knowledge querying (# dropped)	10,752	13,487	13,207	7,432	9,379	–
Model verification (# dropped questions)	22	4	4	18	38	–
Model verification (# dropped instances)	88	16	16	72	152	–
Final (# questions)	2,893	177	456	6,217	4,250	7,250

Table 3: Instances removed during each stage of data creation and filtering. Each question is associated with four evidence types, so the total number of evidence–answer pairs equals four times the final question count; the number of dropped instances scales accordingly. Per-stage drop counts for GPT-5.2 are not reported as the same parametric-knowledge filter and model-verification stages were rerun against GPT-5.2 only at the final-instance level.

Task	Strength	NC	HPC	HPCE	LPC
CK	Weak	43.69	57.79	69.34	53.16
	Neutral	44.76	57.52	71.52	57.90
	Strong	46.39	53.85	70.03	54.20
PK	Weak	47.91	18.53	7.95	36.50
	Neutral	46.60	15.31	7.60	37.47
	Strong	55.06	23.06	8.78	46.98
PCK	Weak	44.88	46.45	52.64	57.95
	Neutral	41.38	50.58	53.52	58.28
	Strong	49.59	48.13	53.59	55.78
RAG	Weak	51.23	45.95	45.65	50.93
	Neutral	51.50	49.36	49.71	50.80
	Strong	57.02	57.21	56.42	55.54

Table 7: Performance of Mistral-7B with different instruction strength. While absolute performance varies, the relative performance relationship still holds.

G Explanation Generation for HPCE

When encountering context that conflicts with prior knowledge, humans are often more persuaded by additional explanations, which help them iteratively update their mental models with new information (Vosniadou and Brewer, 1992). To study this effect, we augment HPC passages with free-text rationales that explicitly reconcile the contradiction from the model-aligned NC perspective. We denote these augmented passages as HPCE (High Plausibility Contradiction with Explanation). The explanations are generated by providing both the NC and HPC answers to a language model and prompting it to produce a corresponding rationale. An example of an HPCE passage is shown in Figure 9, and the full prompt used for explanation generation is provided below.

Based 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.

H Free Generation Setting

H.1 Evaluator Prompts

In the free generation setting described in §4.3, we employ a language model as an evaluator to assess the quality of generated answers. We experiment with several evaluation prompts and select the one that yields the highest agreement with human annotations. Following the design of the evaluator proposed in Hou et al. (2024), we introduce modifications to improve reliability, achieving a higher Cohen’s κ with human annotators. The final evaluator prompt is provided in Figure 11, and a decision tree illustrating the evaluation process is shown in Figure 10.

H.2 Human Annotations

Two colleagues with expertise in natural language processing served as annotators, compensated at standard research assistant rates. They annotated a sample of 50 instances, each using both the evaluator prompt (Figure 11) and the decision tree (Figure 10) to ensure consistency. For each instance, annotators were shown the model prediction alongside the gold answer and asked to label the prediction as *correct*, *partially correct*, or *incorrect*.

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: A significant number of the residents of Kodimunai do jobs related to fishing

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.Farming Answer: B

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.Farming Answer: B

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 7: Examples of each task.

I License of Artifacts

All licenses of artifacts used in this work can be found in Table 8.

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 8: License of artifacts used in this paper.

J Prompts

J.1 Evidence Creation Prompts

We generated LPC and HPCE passages using GPT-4o after several rounds of prompt refinement. The final prompts used for evidence creation are shown in Figure 14.

Following the generation, all passages were subjected to plausibility checks. For LPC passages, the model was asked to determine whether the passage would be deemed implausible in the real world. For HPCE passages, the model was required to verify that the passage was both highly plausible and explained the underlying conflict. The prompts for this step are included in Figure 15.

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 parentheses, 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.Farming Answer:

Gold Answer CB

(a) Example of PCK task. The model is given only an external context and is 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 parentheses, 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 include 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.Farming 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 8: 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 9: An example of HPC and HPCE.

J.2 Task-Annotation Prompts

Since the base datasets already provide answers to the questions, additional annotation was only required to specify the task in the knowledge-free setting. We framed these tasks as extractive question answering, requiring the model to copy the answer directly from the passage (Figure 7). We then used GPT-4o as the annotator model to extract all acceptable answer spans.

J.3 Validation Prompts

In the final stage of data construction (validation in Figure 2), all instances were passed through language models for validation. The corresponding prompts are listed in Appendix J.3.

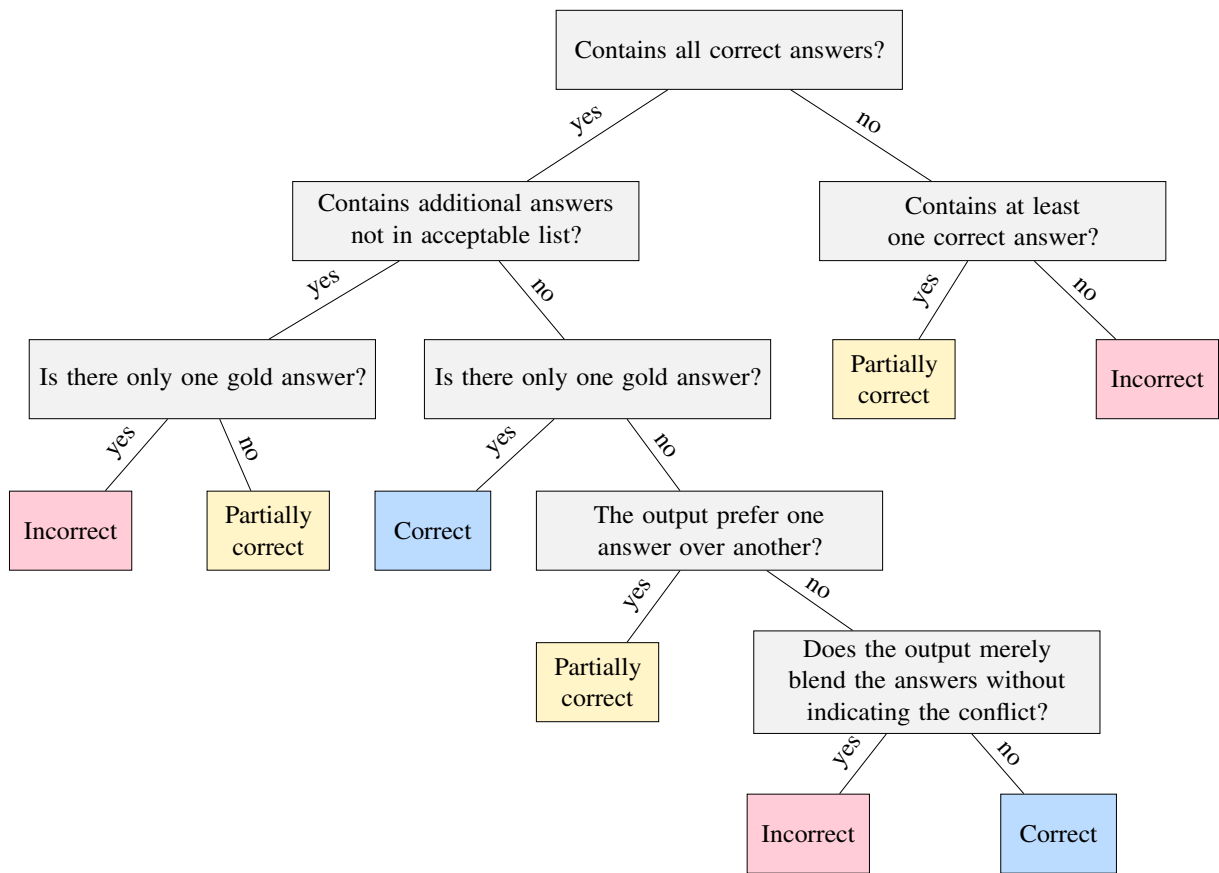


Figure 10: Decision tree for annotation guidance.

Model	Task	NC	HPC	HPCE	LPC
Mistral-7B	KFextract	1.6	1.4	0.5	0.4
		76.7	77.3	83.4	85.2
	CK	65.3	28.5	42.2	38.2
		65.3	28.9	42.3	38.5
	PK	62.6	18.4	6.4	51.4
		62.7	18.6	6.5	51.6
	PCK	62.4	0	0	9.4
		62.5	31.5	34.2	54.5
	RAG	54.4	0.2	0.1	7.9
		54.7	41.2	38.1	54.4
OLMo2-7B	KFextract	0	0	0.7	0
		54.4	66.2	66.4	74.3
	CK	56.8	48.9	47.2	55.1
		57	48.9	48.5	56
	PK	55.7	11.9	7.4	31.2
		58.5	14.3	9	32.2
	PCK	44.3	0	0	5.7
		48.2	39	40.8	50.3
	RAG	41.5	1.1	0	3.4
		47.8	43.3	42.4	46.3
OLMo2-13B	KFextract	1.9	0.2	0.2	0.7
		67.4	67.3	74.7	73.4
	CK	25.9	16	20.2	4.8
		26.4	17	21.3	6
	PK	57.5	31.6	18	50.4
		57.6	32	18	50.4
	PCK	76.1	0	0	6.6
		76.1	57.2	50.4	42.7
	RAG	82	0.4	0	12.7
		82.4	60.9	60.5	65.8
Qwen2.5-7B	KFextract	1.6	0.9	0.3	0.5
		75.1	71.4	78	80.4
	CK	78.8	47.2	57.4	42.6
		82.1	50.3	64.4	48.2
	PK	82.8	48.5	28.1	62.5
		84.4	50.3	31.2	65.8
	PCK	83.9	0.5	0.9	14.3
		84.4	61.5	60.5	69
	RAG	79.5	1.3	1.5	14.3
		82.3	60.2	59.8	67.1
Qwen2.5-14B	KFextract	1	0.7	0.8	0.4
		78.4	79.5	82.8	81.2
	CK	87.4	70.2	85.5	41.6
		90.5	73.2	90.5	44
	PK	76.7	25.7	4.1	48.4
		86.9	32	9	55.1
	PCK	84.3	2.1	1.8	6.8
		89.9	62.7	63.7	61.1
	RAG	82.4	10.2	9.9	9.7
		87.9	59.1	59.9	61

Table 4: Performance of models. Exact Match rows are unshaded; F1 rows are shaded in numeric columns.

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.

Figure 11: 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.

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

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 12: 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 13: Final Prompt for model based evaluation (cont 2).

LPC instances Creation Prompt.

You are a smart editor who creates implausible texts. Your job is to generate evidence for the
↳ given question such that the answer to the question is NOT the Rejected Answer. You can
↳ work on the given plausible passages as the starting point. You should change the content
↳ of the given passage, remove any explanation given in the passage, 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 it comes 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.

Based 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 as much 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 14: 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 15: 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 16: 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 17: Final prompt validating the generated evidence provide the correct answer to the question.