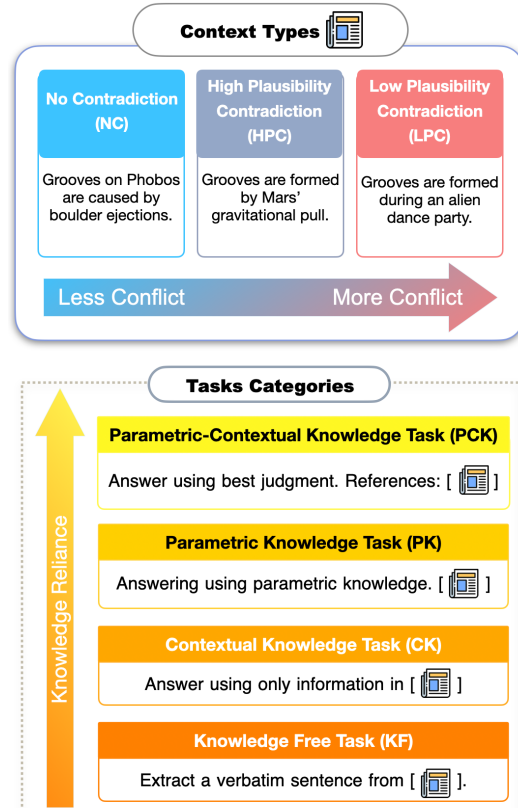


Task Matters: Knowledge Requirements Shape LLM Responses to Context–Memory Conflict

Anonymous ACL submission

Abstract

Large language models (LLMs) rely on both contextual knowledge and parametric memory, yet these sources can conflict. Prior analysis largely focused on contextual question answering, suggesting that models tend to favor parametric knowledge under conflict, but this setting assumes that tasks should always rely on the provided passage. It therefore remains unclear how LLMs behave when *tasks demand different kinds and degrees of knowledge utilization*. We address this gap with a model-agnostic diagnostic framework that holds underlying knowledge constant while injecting controlled conflicts across tasks with varying knowledge requirements. Evaluating representative open-source LLMs, we find that: (1) performance degradation under conflict correlates with a task’s knowledge reliance rather than conflict plausibility alone; (2) strategies such as explanatory rationales or reiteration increase context reliance, helping context-only tasks but harming those that require parametric knowledge; and (3) these behaviors bias model-based evaluation, raising concerns about the reliability of LLMs as judges. Together, our findings show that context–memory conflict is fundamentally task-dependent and motivate task-aware approaches to balancing context and memory in LLM deployment and evaluation.¹



* 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.

1 Introduction

Large language models (LLMs) perform well on many knowledge-centric tasks because they encode vast amounts of parametric knowledge within their parameters. 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 cut-off. When such input contradicts what the model “knows,” *context–memory conflict* arises, and LLMs often favor their own parametric knowledge over

the provided context (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

¹Our framework and data are available at [Anonymous].

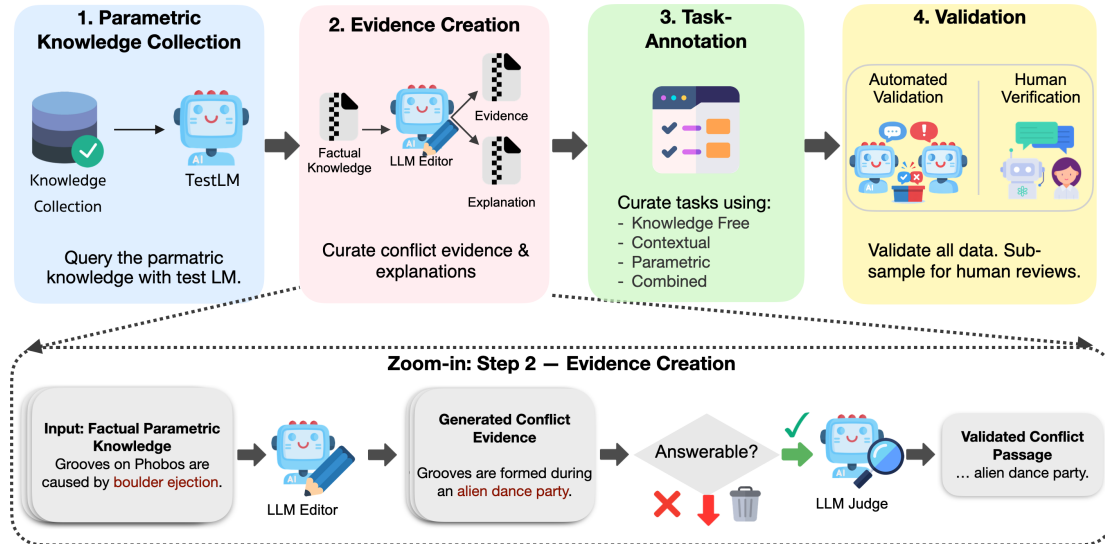


Figure 2: Overall diagnostic data creation flow. The lower portion is a zoom in of Evidence Creation step. After collecting the test model’s parametric knowledge, the supporting passages are further edited to reveal multiple levels of conflict (2. Evidence Creation) and appear in different tasks (3.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

054 these results, however, are derived from contextual
 055 question answering, where models are expected
 056 to rely exclusively on the provided passage. As a
 057 result, it remains unclear how LLMs should behave
 058 under context–memory conflict when tasks require
 059 different forms of knowledge utilization (Xu et al.,
 060 2024).

061 In practice, tasks vary sharply in their reliance
 062 on contextual versus parametric knowledge. Ex-
 063 tractive tasks, such as text copying, impose little
 064 need for prior knowledge, whereas tasks such as
 065 literature review require integrating background
 066 knowledge with new information published after
 067 the model’s knowledge cutoff. Between these ex-
 068 tremes lie tasks that demand paraphrasing, selective
 069 grounding, or explicit reconciliation of conflicting
 070 evidence. Treating all such settings as interchange-
 071 able obscures how context–memory conflict func-
 072 tions across real-world applications.

073 In this work, we show that **the impact of**
 074 **context–memory conflict is task-dependent**. We
 075 demonstrate that performance degradation under
 076 conflict correlates with a task’s knowledge require-
 077 ments, in addition to the level of conflict alone.
 078 To make this dependence explicit, we introduce a
 079 diagnostic framework that holds underlying knowl-
 080 edge constant while systematically varying task
 081 formulations. This allows us to isolate how the
 082 same conflict affects models when a task requires
 083 (i) no knowledge beyond surface extraction, (ii)
 084 grounding in context, (iii) reliance on parametric
 085 knowledge, or (iv) integration of both. Our frame-

086 work automatically identifies a model’s parametric
 087 beliefs and injects controlled contradictions into
 088 downstream tasks (Figure 1), producing model-
 089 specific diagnostic datasets that vary both conflict
 090 level and task knowledge requirements. Evaluating
 091 models on these datasets enables us to disentangle
 092 the effects of knowledge conflict from task de-
 093 mands. This allows us to quantify how knowledge
 094 conflict disrupts model behavior in a task-aware
 095 manner (Figure 2).

096 Across representative LLMs, we find that the
 097 effects of context–memory conflict are strongly
 098 task-dependent: conflicts have minimal impact on
 099 tasks requiring little or no knowledge utilization,
 100 yet significantly impair knowledge-intensive tasks,
 101 even under explicit instructions. The factor of tasks
 102 explains the prior findings that would appear in-
 103 consistent if examined without the lens of task.
 104 Beyond overall performance, we find that simple
 105 strategies such as reiteration encourage context re-
 106 liance, without the need to alter the inference mech-
 107 anism. Reiteration improves performance when the
 108 task is intended to rely primarily on contextual in-
 109 formation (Wang et al., 2022; Wu et al., 2024),
 110 but degrades performance when successful task
 111 completion requires dominant use of parametric
 112 knowledge. Finally, we demonstrate that studying
 113 context–memory conflict particularly matters for
 114 the task of model-based evaluation. Using LLMs
 115 as judges (Zheng et al., 2023; Liu et al., 2023; Ru
 116 et al., 2024; Chen et al., 2025) constitutes a task in
 117 which the model is required to balance parametric

118 knowledge and contextual information, while the
119 model’s own parametric knowledge can implicitly
120 bias evaluations, particularly when factual correct-
121 ness conflicts with the provided context. Together,
122 our findings suggest that context–memory conflict
123 cannot be meaningfully studied or mitigated with-
124 out explicit consideration of task knowledge re-
125 quirements, and motivate task-aware approaches to
126 balancing contextual and parametric knowledge.

127 2 Related Work

128 **Context-Memory Conflict** Xu et al. (2024)
129 classify knowledge conflict into three categories:
130 *context-memory conflict*, *inter-context conflict* (con-
131 tradictory evidence among retrieved passages), and
132 *intra-memory conflict* (inconsistent parametric be-
133 liefs). We focus on the context-memory conflict,
134 which arises when a given information-bearing text
135 chunk contradicts the model’s parametric beliefs.

136 **Nuanced Behaviors under Conflict** Early stud-
137 ies reported that models tend to rely on their own
138 knowledge when the prompt provides contradic-
139 tory evidence (Longpre et al., 2021; Chen et al.,
140 2022). Later work revealed a more nuanced picture.
141 On synthetic datasets, Xie et al. (2023) showed
142 that LLMs often update their answers when given
143 strong and convincing evidence, whereas Jin et al.
144 (2024a) observed a “Dunning–Kruger” effect in
145 stronger LLMs, which display higher confidence
146 in their incorrect parametric knowledge than in the
147 external context. Further analysis also finds that
148 models show availability bias (leaning on common-
149 knowledge facts), majority bias (trusting the an-
150 swer supported by more frequent evidence across
151 documents), and confirmation bias (preferring ev-
152 idence consistent with their prior knowledge), es-
153 pecially when the models are given misleading or
154 irrelevant answers. Moving to realistic documents,
155 Kortukov et al. (2024) found that models update
156 their answers more reliably than synthetic evalu-
157 ations suggest, yet still exhibit a *parametric bias*:
158 if the model’s originally believed answer appeared
159 anywhere in the context (even as a distractor), the
160 model was more likely to stick to that incorrect
161 answer.

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

167 pose a method that dynamically prunes or patches
168 specific attention heads that cause conflicts. Li
169 et al. (2025) instead propose a two-step inference
170 intervention that has each step focus on either the
171 context or the memory heads. Efforts have also
172 been made to develop novel decoding methods that
173 enhance the use of contextual knowledge (Jin et al.,
174 2024a; Shi et al., 2024; Wang et al., 2025a). More
175 recently, work has explored post-hoc, inference-
176 time controls that steer context reliance without
177 fine-tuning, including proxy-model steering (Wang
178 et al., 2025b) and test-time attention interventions
179 (Li et al., 2025).

180 **Our Focus** Most prior studies focus on contex-
181 tual question answering, a setting that *requires*
182 heavy reliance on the provided passages. Many
183 other tasks, for example, grammar correction or
184 claim verification, may need little context or, con-
185 versely, require careful integration of both para-
186 metric and contextual knowledge. This leaves the
187 question of whether context-memory conflict poses
188 the same impact on tasks with different knowledge
189 demands unanswered. To fill this gap, we keep
190 the underlying knowledge constant while varying
191 the *task formulation*, creating controlled datasets
192 that induce different conflict levels for each target
193 model. We introduce an analysis tool that auto-
194 matically constructs model-specific test sets. Our
195 findings indicate that both knowledge-memory con-
196 flict and blindly following the context could be
197 particularly harmful to model-based evaluations,
198 further motivating *task-dependent* methods for dy-
199 namically balancing context and memory (Wang
200 et al., 2025b; Li et al., 2025).

201 3 Context-Memory Conflict Creation

202 Figure 2 illustrates an overview of the data con-
203 struction pipeline. The process begins with identi-
204 fying the pre-existing knowledge within a language
205 model (Parametric Knowledge Collection). We use the
206 knowledge from knowledge conflict question answer-
207 ing datasets (Wan et al., 2024; Hou et al., 2024)
208 that have two or more acceptable answers to one
209 question, using them to identify the stance aligned
210 with the model’s parametric belief, which then serves
211 as the basis for constructing task data. A piece of
212 knowledge is considered part of the model’s internal
213 belief only if the model consistently aligns with the
214 perspective in a single answer across all prompt
215 variations under greedy decoding, while rejecting
216 conflicting alternatives, with details

included in Appendix A. 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).

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 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 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 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

317 knowledge. In experiments, the model is given
318 one of the passages in $\{p_i^{\text{NC}}, p_i^{\text{LPC}}, p_i^{\text{HPC}}\}$ and is
319 expected to answer questions only based on the
320 contextual knowledge, which may not agree with
321 its parametric knowledge.

322 **Parametric Knowledge (PK)** tasks may present
323 inputs that include distracting or irrelevant con-
324 text. The model is expected to rely exclusively on
325 its parametric knowledge to answer the questions.
326 In experiments, the model is given passages that
327 support or contradict its parametric knowledge as
328 input, and the model is always expected to provide
329 the answer a_i^{NC} .

330 **Parametric-Contextual Knowledge (PCK)**
331 tasks explicitly ask the model to integrate both its
332 internal knowledge and the external context. This
333 setup reflects scenarios akin to scientific reasoning,
334 where individuals must synthesize background
335 knowledge with newly presented information
336 (e.g., a recently read paper). In execution, the
337 model will be given a passage that contradicts its
338 own knowledge, and is expected to output both
339 perspectives from the context and its parametric
340 knowledge.

341 **Retrieval Augmented Generation (RAG)** simu-
342 lates the standard RAG setting, where models are
343 not explicitly instructed to prioritize parametric or
344 contextual knowledge. The model will be given
345 two passages and is expected to answer the ques-
346 tion based on both passages. Models are expected
347 to acknowledge the conflict and discuss each poten-
348 tial answer individually. This setting serves as both
349 a complementary setting to PCK tasks and creates a
350 test bed that is closer to real-world applications, as
351 (Hagström et al., 2024) show that insights based on
352 synthesized data are not guaranteed to generalize
353 to real-world scenarios.

354 The annotations for all five tasks are derived
355 directly from the original datasets on which our
356 framework is built. These task types primarily dif-
357 fer in the number of valid answers expected and
358 the nature of knowledge the model should rely on.
359 In KF tasks, the model is only expected to perform
360 extractions. In CK and PK tasks, the model is ex-
361 pected to give only one answer or provide a single
362 correct answer, grounded either in the provided
363 context or in its internal (parametric) knowledge,
364 respectively. In PCK and RAG tasks, the model
365 is expected to clarify that both a_i^{NC} and the other
366 answer are possible and explain the contradiction
367 between the two answers.

368 One of the original datasets we use employs
369 model-based evaluation to judge the correctness
370 of free-text answers (Hou et al., 2024). However,
371 we observed that this evaluation method is suscep-
372 tible to knowledge conflict, leading to inaccurate
373 evaluations. We explore this issue further in §4.3.
374 Therefore, we modify the non-extractive tasks to be
375 multiple-choice questions. Each instance presents
376 four answer options; the model must first generate
377 an explanation, then select the most appropriate
378 answer. To assess the performance of the target
379 model, we report the accuracy for CK and PK tasks,
380 F1 for KF, PCK, and RAG. To obtain high-quality
381 texts, we use GPT-4o as the base model to create
382 evidence and validate the diagnostic data. Then,
383 we analyze the instruction-tuned version of Mistral-
384 7B (Jiang et al., 2023), OLMo2-7B, OLMo2-13B
385 (OLMo et al., 2024), Qwen2.5-7B, and Qwen2.5-
386 14B (Qwen et al., 2025), all of which are widely
387 used open-weight models that represent diverse
388 training paradigms. The resulting dataset statistics
389 are presented in Appendix C.

390 4 Findings

391 4.1 Conflict Impairs Model Performance on 392 Knowledge-Intensive Tasks

393 The performance of each model on each task type
394 and context type is reported in Figure 3. A univer-
395 sal trend can be observed: in all but Knowledge-
396 Free tasks, all models suffer when asked to provide
397 responses that contradict their parametric knowl-
398 edge.

399 **Knowledge conflict degrades performance when-
400 ever knowledge is required.** In CK tasks (Fig-
401 ure 3b), the model is explicitly instructed to ignore
402 its own beliefs and rely solely on the given pas-
403 sage. Nevertheless, every model shows a clear
404 $\text{NC} > \text{HPC} / \text{LPC}$ performance ordering, indicating
405 that the model still relies on parametric knowl-
406 edge when it is not supposed to. This aligns with
407 prior work’s finding that models favor their para-
408 metric knowledge more than the given contextual
409 knowledge, thus leading to hallucinations (Jin et al.,
410 2024a). This issue, if left untreated, could not only
411 affect the overall performance but also the correct-
412 ness of model-based evaluation results, which we
413 illustrate in §4.3.

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

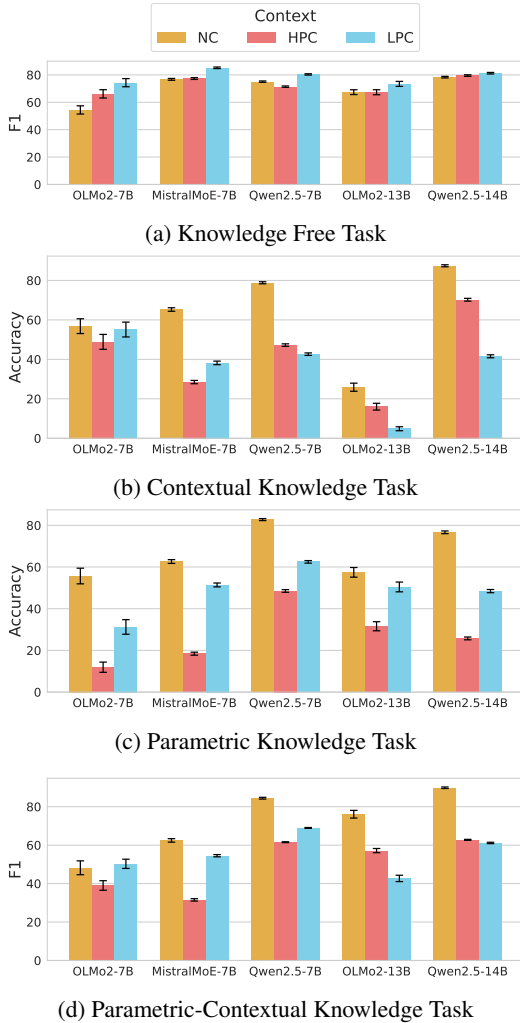


Figure 3: Performance of each model on different task types. A clear trend of $NC > HPC / LPC$ is shown across tasks involving knowledge utilization.

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 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 highly 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 there is subtle knowledge 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

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

More plausible \neq Higher reliance 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 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.

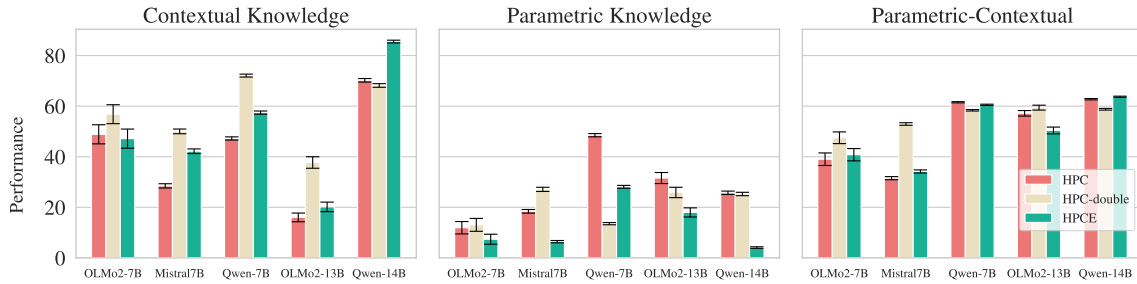


Figure 4: Performance on high plausibility contradiction instances with (HPCE) and without (HPC) explanations.

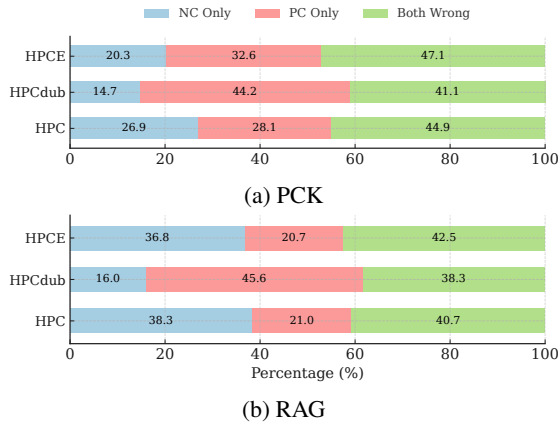


Figure 5: Averaged error distribution on RAG and PCK tasks. *NC Only* represents that the model only provides the NC answer; *PC Only* represents that the model only provides the PC answer; *Both Wrong* represents the case where the model provides neither PC or NC answer.

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, HPC dub, 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 the rationale 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 improved 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 Appendix 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).

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.

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 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 Figure 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. 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 suggest the risk of using language models as evaluators, where the language model could be biased by its parametric knowledge, 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).

617
618
619
620
621
622
623
624
625
626
627
628
629
630
631

632
633
634
635
636
637
638
639
640
641
642
643
644
645
646
647
648

649

650
651
652
653
654
655
656

657
658
659
660
661
662
663
664
665

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.

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.

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. 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. *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 Tchraikian, Radu Marinescu, Elizabeth Daly, Inkit Padhi, and Prasanna Sattigeri. 2024. *Wicontradict: A benchmark for evaluating llms on*

666
667
668
669
670
671
672

673
674
675
676
677
678
679
680

681
682
683
684
685
686
687
688

689
690
691
692

693
694
695
696
697
698
699

700
701
702
703
704

705
706
707
708
709

710
711
712
713
714

715
716
717
718

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

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,

Stage	Mistral	OLMo2-7B	OLMo2-13B	Qwen2.5-7B	Qwen2.5-14B
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	176	456	6,217	4,250

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.

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.

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

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.Farming 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.Farming 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 7: Examples of each task.

1024 standard research assistant rates. They annotated a
1025 sample of 50 instances, each using both the evaluator
1026 prompt (Figure 11) and the decision tree (Figure
1027 10) to ensure consistency. For each instance,
1028 annotators were shown the model prediction along-
1029 side the gold answer and asked to label the predic-
1030 tion as *correct*, *partially correct*, or *incorrect*.

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.

1031 I License of Artifacts

1032 All licenses of artifacts used in this work can be
1033 found in Table 8.

J Prompts

J.1 Evidence Creation Prompts

We generated LPC and HPCE passages using GPT-4o after several rounds of prompt refinement. The

1034

1035
1036
1037

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.Farming Answer: CB

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.Farming Answer: BC

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.

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.

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.

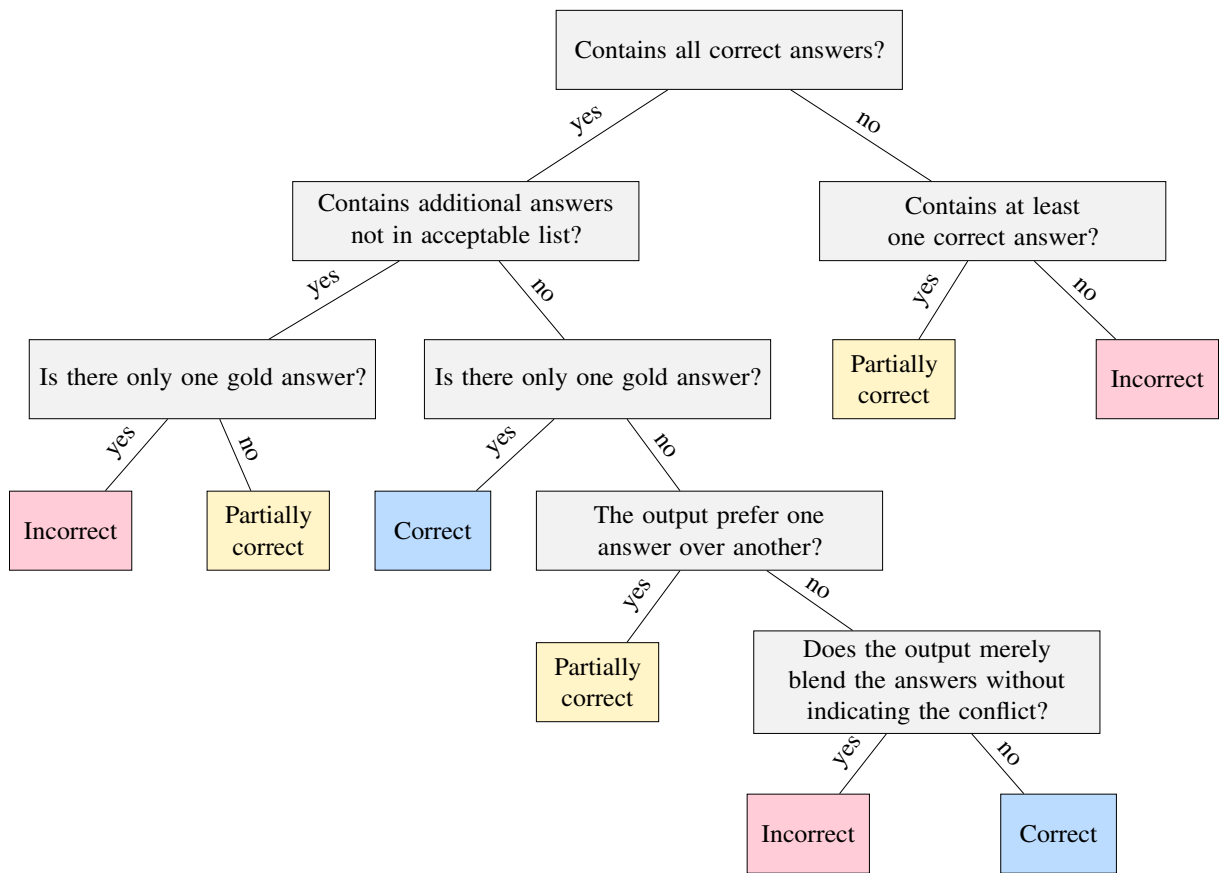


Figure 10: Decision tree for annotation guidance.

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 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 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.