UniKnow: A Unified Framework for Reliable Language Model Behavior across Parametric and External Knowledge

Anonymous ACL submission

Abstract

Language models often benefit from external knowledge beyond parametric knowledge. While this combination enhances performance, achieving reliable knowledge utilization remains challenging, as it requires assessing the state of each knowledge source based on the presence of relevant information. Yet, prior work on knowledge integration often overlooks this challenge by assuming access to relevant contexts or by disregarding the state of parametric knowledge, thereby limiting the coverage of knowledge scenarios. To address this gap, we introduce **UniKnow**, a **Uni**fied framework for reliable LM behavior across parametric and external **Know**ledge. UniKnow enables controlled evaluation across knowledge scenarios such as knowledge conflict, distraction, and absence conditions that are rarely addressed together. Beyond evaluating existing methods under this setting, we extend our work by introducing UniKnow-Aware methods to support comprehensive evaluation. Experiments on UniKnow reveal that existing methods struggle to generalize across a broader range of knowledge configurations and exhibit scenario-specific biases. UniKnow thus provides a foundation for systematically exploring and improving reliability under knowledge scenarios.

Figure 1: Four knowledge scenarios in UniKnow are defined by the boundaries of parametric and external knowledge sources. Each region illustrates the expected LM behavior for each scenario.

1 Introduction

Language models (LMs), trained on large-scale corpora, exhibit the capacity to address a broad range of tasks by leveraging their pre-trained parametric knowledge (Grattafiori et al., 2024; Yang et al., 2024). However, LMs are confined to the static pre-trained knowledge and therefore struggle to handle tasks requiring information beyond this boundary, such as long-tail (Kandpal et al., 2023; Mallen et al., 2023) or time-sensitive information (Liska et al., 2022). To overcome these limitations, LMs often benefit from dynamically incorporating external knowledge, commonly through retrieval-augmented generation (RAG), thereby granting access to up-to-date, task-relevant information at inference time (Chen et al., 2017; Asai et al., 2023).

The integration of parametric and contextual knowledge has broadened the capabilities of LMs, driving their application in knowledge-intensive and sensitive domains (Tsatsaronis et al., 2015; Jin et al., 2019; Dasigi et al., 2021). Consequently, the reliability of LMs has become a vital consideration (Wen et al., 2024a), with models expected to not only recognize the boundaries of their possessed knowledge but also identify when relevant information is missing. While prior work has tackled various dimensions of knowledge integration (Su et al., 2024; Yoran et al., 2023), these studies have typically remained fragmented, providing an incomplete assessment of reliability (Li et al., 2023; Cheng et al., 2024). Moreover, knowledge utilization methods developed under such narrow environments still lack validation in more realistic and compositional knowledge scenarios.

To this end, we introduce **UniKnow**, a unified framework for reliable LM behavior across parametric and external knowledge. While reliability may encompass a broader range of factors, this work focuses on the presence of *relevant* information in each parametric and external knowledge source. Central to UniKnow is the notion of *relevance*, which we define as whether a knowledge source provides sufficient and contextually supporting information to answer a query.

UniKnow is designed to categorize and assess four distinct scenarios as illustrated in Figure 1: (1) Conflict, (2) Parametric-Only, (3) External-Only, and (4) Unknown. When only a single relevant source is available, the model is expected to ground its output solely in that source. Furthermore, if both sources are relevant but conflicting (1), the model should prioritize the external knowledge, as it generally offers more up-to-date and task-specific information. If neither source provides relevant knowledge (4), the model should recognize its limitations and abstain from generating hallucinations (Zhang et al., 2024a; Feng et al., 2024).

To examine how existing methods developed under partial scenario coverage generalize to UniKnow, we evaluate two naïve baselines and three knowledge utilization methods, each representative of distinct scenario coverages. Given the lack of existing approaches that comprehensively consider all UniKnow scenarios, we introduce two UniKnow-Aware approaches that explicitly incorporate relevance-based knowledge conditions into their formulation to complement our analysis.

Our in-depth analysis under UniKnow reveals that methods appearing reliable in individual scenarios often fail in composite scenarios requiring simultaneous consideration of both knowledge sources. We further uncover how LM behavior shifts across scenarios, highlighting biases specific to scenario types. Together, these findings enable a more comprehensive understanding of LM alignment potential under UniKnow and mark a substantial step toward bridging the gap between narrow knowledge settings and a unified framework.

2 Related Works

External Knowledge Integration LMs often face inconsistencies between their static parametric knowledge and dynamic external contexts, requiring them to handle *conflicting* (Longpre et al., 2021; Xie et al., 2023) or *irrelevant* (Shen et al., 2024; Wu et al., 2024) information effectively. To resolve knowledge conflicts, several approaches aim to improve external knowledge incorporation, primarily through context-aware contrastive decoding (Shi et al., 2024; Jin et al., 2024b; Yuan et al., 2024). Additionally, to mitigate the impact of irrelevant external information, researchers have explored methods to encourage LMs to rely on their parametric knowledge (Yoran et al., 2023; Asai et al., 2024; Xia et al., 2024; Luo et al., 2023). However, they often entirely overlook the presence of relevant information within LM's parametric knowledge when processing external contexts.

Abstention A growing line of work focuses on aligning LMs to abstain when appropriate (Feng et al., 2024; Zhang et al., 2024a)–specifically when LMs lack relevant knowledge–to prevent hallucination and ensure reliable LM behavior (Wen et al., 2025). Recently, studies have begun to explore abstention based on the relevance of external knowledge (Wen et al., 2024a; Kim et al., 2025).

Knowledge Frameworks There have been efforts to unify various aspects of knowledge utilization to understand LM behaviors. Li et al. (2023) trains LMs to generate parametric- or context-grounded responses depending on the context type, whereas Neeman et al. (2023) trains LMs to generate both in parallel. Similar to our work, Cheng et al. (2024) proposes a benchmark to investigate whether LMs can express possessed parametric knowledge when exposed to various context types. While prior approaches have provided diverse insights into how LMs utilize knowledge, our work introduces a distinct perspective–a unified framework based on a precise formulation of knowledge relevance for both parametric and external sources.

3 UniKnow

This work focuses on context-augmented generation in open-domain question-answering, facilitating LMs to leverage their **parametric** knowledge while simultaneously utilizing **external** knowledge to answer a given query q. This section first defines each knowledge source based on the availability of relevant information. Guided by this taxonomy, we introduce **UniKnow**, a **Uni**fied framework for reliable LM behavior across parametric and external **Know**ledge, covering four distinct scenarios as illustrated in Figure 2. We then describe the construction process of estimating the parametric knowledge and designing diverse context types.

3.1 Definition of Knowledge Sources

Parametric knowledge (PK) refers to information encoded in an LM during pretraining. Since this

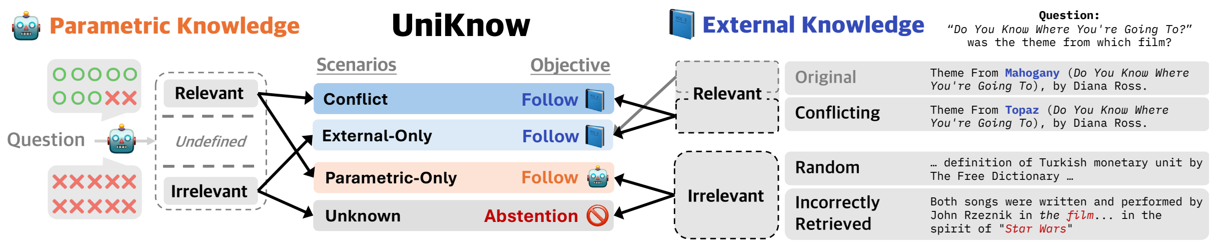

Figure 2: Overview of UniKnow with four distinct knowledge scenarios (Section 3.2). Each scenario is defined by jointly considering the relevance states of parametric knowledge (Section 3.3) and external knowledge (Section 3.4).

knowledge is bound by its pretraining data, we define that relevant information resides in PK ($\exists_{\text{PK}}$) if $\text{LM}(\hat{a} \mid q) = a^*_{\text{PK}}$, where a^*_{PK} denotes the answer grounded in the LM's pretraining data (Bang et al., 2025). Still, PK remains inherently static and may not align with the most recent world knowledge.

External knowledge (EK) indicates any information provided at inference time as the input context. To isolate relevance and solely evaluate the LM's ability to utilize relevant knowledge, we exclude judging the factuality of EK from the scope of this study. Under this condition, we analyze EK from relevant ($\exists_{\text{EK}}$) and irrelevant ($\varnothing_{\text{EK}}$) perspectives.

3.2 Scenarios in UniKnow

UniKnow is designed to cover all possible scenarios regarding the presence of relevant PK and EK. This gives rise to four distinct scenarios, each reflecting real-world challenges such as conflict resolution, over-reliance, and hallucination risk. Since each challenge has its own expected behavior, we define scenario-specific expectations as follows.

- **Conflict** (`C`): ($\exists_{\text{PK}}, \exists_{\text{EK}}$) and $a^*_{\text{PK}} \neq a^*_{\text{EK}}$
  The conflict between knowledge sources arises when EK presents relevant information contradicting what LM knows (Xu et al., 2024b). While PK and EK may either align or conflict, we focus on the latter, allowing us to evaluate whether LMs can correctly prioritize EK.

- **External-Only** (`E-Only`): ($\varnothing_{\text{PK}}, \exists_{\text{EK}}$)
  The model lacks PK with relevant information and is expected to rely on relevant EK.

- **Parametric-Only** (`P-Only`): ($\exists_{\text{PK}}, \varnothing_{\text{EK}}$)
  The model is required to rely on its PK with relevant information and ignore irrelevant EK.

- **Unknown** (`U`): ($\varnothing_{\text{PK}}, \varnothing_{\text{EK}}$)
  Neither knowledge source is sufficient, and the model is expected to abstain from answering.

3.3 Parametric Knowledge Estimation

We estimate the presence of relevant PK by assessing whether the LM is capable of generating a correct answer to a given q without access to external context. Following prior works, we assess the factual *correctness* (Zhang et al., 2024a,b; Wang et al., 2024b) and *consistency* (Kuhn et al., 2023; Huang et al., 2025; Amayuelas et al., 2024b) of the prediction utilizing its PK. We classify q as $\exists_{\text{PK}}$ if both conditions are satisfied, and as $\varnothing_{\text{PK}}$ otherwise.

For each q, we sample n responses using q alone: $a_i \sim LM(a \mid q)$ for $i = 1, .., n$. If the proportion of correct responses is greater than or equal to the threshold τ, we classify q as $\exists_{\text{EK}}$:

$$\frac{1}{n}\sum_{i=1}^{n} \mathbf{1}[a_i = a^*_{\text{PK}}] \geq \tau \Rightarrow q \in \exists_{\text{PK}} \qquad (1)$$

If none of the responses are correct, we assign $q \in \varnothing_{\text{PK}}$. Questions falling between these thresholds are considered *undefined* and excluded from scenario construction. We set $n = 10$ and $\tau = 0.7$ in our implementation.

3.4 External Knowledge Construction

To operationalize each scenario, we construct context types tailored to diverse conditions. In addition to the original context, we construct conflicting and two types of irrelevant contexts: (1) topically unrelated random contexts, and (2) incorrectly retrieved contexts with high retriever score. This allows fine-grained control over the degree of relevance, capturing challenges ranging from knowledge conflicts to misleading but plausible distractors. Figure 2 provides examples of each context type, with the corresponding mapped scenarios shown using arrows.

Relevant contexts The *original* context refers to the context paired with the question-answer pair in the dataset. We derive a *conflicting* context by providing Llama 3 70B Instruct (Grattafiori

Methods	Conflict	E-Only	P-Only	Unknown	Train
COIECD	✔	✔	✘	✘	✘
RetRobust	✘	✔	✔	✘	✔
KAFT	✔	✔	✔	✘	✔
$\text{COIECD}_{\text{Prompt}}$	✔	✔	✔	✔	✘
$\text{LM}_{\text{UniKnow}}$	✔	✔	✔	✔	✔

Table 1: Characteristics of knowledge utilization methods, indicating their UniKnow scenario coverage and whether they are training-based.

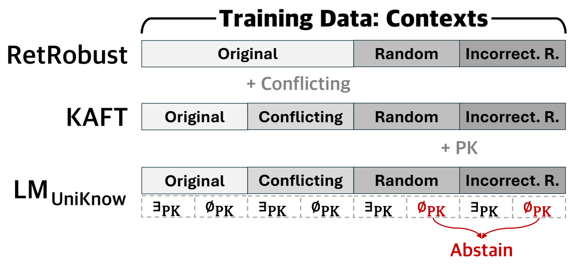

Figure 3: Training data composition for training-based methods, illustrating incorporated context types and $\text{LM}_{\text{UniKnow}}$'s unique integration of parametric knowledge states ($\exists_{\text{PK}}, \varnothing_{\text{PK}}$) to guide expected behaviors.

et al., 2024) with the original context and the corresponding answer to generate an alternative answer while preserving its part of speech. The original answer is then replaced with the conflicting answer, introducing an intended conflict with the model's PK. Note that the `C` scenario uses only conflicting contexts, while the `E-Only` scenario includes both original and conflicting contexts.

Irrelevant contexts We consider two key aspects for irrelevant context selection: the absence of the answer span (i.e., uninformative) and the potential semantic relevance (Wu et al., 2024) that may mislead the model (i.e., misleading). To capture both uninformative and misleading cases, we include two types of contexts. A *randomly* sampled context from the same dataset, topically unrelated to the question, and not containing the original answer. The *incorrectly retrieved contexts* also lack the answer but may appear topically relevant, thereby creating a false sense of relevance. We obtain these incorrectly retrieved contexts by querying a Wikipedia corpus using the CONTRIEVER-MSMARCO retriever (Izacard et al., 2022), and then select the highest-ranked context that does not contain the answer. This setting captures challenges in real-world RAG, where retrieval often returns plausible but irrelevant information.

4 Knowledge Utilization Methods

This section describes methods used to evaluate model behavior under UniKnow (Table 1). As an initial baseline, we take a **prompting** approach, instructing LMs to consider the presence of knowledge sources for a reliable generation. We also perform **naïve** generation with a QA task template.

4.1 Existing Methods

We adapt three existing knowledge utilization methods, chosen for being either state-of-the-art (SoTA) or representative of approaches designed for partial UniKnow scenarios, enabling evaluation of their generalization under UniKnow.

Knowledge Conflict Context-aware contrastive decoding approaches for resolving knowledge conflict aim to overwrite the model's PK with EK. Among them, we utilize **COIECD**[1] (Yuan et al., 2024), a SoTA method that amplifies the context-informed distribution when conflict arises.

Irrelevance We include **RetRobust** (Yoran et al., 2023) as a representative method for handling irrelevant EK. RetRobust fine-tunes LMs with augmented training data, incorporating irrelevant contexts alongside the original to improve robustness.

As a representative method addressing both knowledge conflict and irrelevance, we include Knowledge-Aware Fine-Tuning (**KAFT**; Li et al., 2023). Their training data includes original, conflicting, and irrelevant contexts, aiming to improve the LM's overall answerability when utilizing EK.

4.2 UniKnow-Aware Methods

Since no single existing method comprehensively addresses all UniKnow scenarios, we present our initial attempts to expand existing methods and explicitly train LMs with UniKnow, evaluating their impact on LM behavior.

UniKnow-Aware Inference To explicitly account for UniKnow's scenarios during inference, we introduce **$\text{COIECD}_{\text{Prompt}}$**, an extension of COIECD that additionally incorporates prompting into the decoding process. By explicitly considering all the possible scenarios, we expect $\text{COIECD}_{\text{Prompt}}$ to cover a broader range of cases.

UniKnow-Aware Training We investigate whether reliability can be improved by training LMs with supervision aligned to knowledge scenarios defined in UniKnow. We design

[1]Contextual Information-Entropy Constraint Decoding

scenario-aware training data that explicitly reflects the presence or absence of relevant information in both knowledge sources. The key lies in the scenario-aware construction of the training data.

To prepare training data, we sample a balanced set of $q \in \exists_{\text{PK}}$ and $q \in \varnothing_{\text{PK}}$, as determined by the criteria in Section 3.3. As illustrated in Figure 3, each q is paired with four types of external contexts described in Section 3.4 to cover knowledge scenarios. For scenarios where relevant information is available–`C`, `E-Only`, and `P-Only`–$\text{LM}_{\text{UniKnow}}$ is optimized to produce the expected answer for each scenario. In the `U` scenario, $\text{LM}_{\text{UniKnow}}$ is trained to abstain by generating `"unknown"`. Further details of each method are presented in Appendix B.1.

5 Experimental Setting

Datasets We employ seven QA datasets from diverse knowledge domains to construct UniKnow: NaturalQuestions (NQ), TriviaQA, HotpotQA, SQuAD, BioASQ, TextbookQA, and RelationExtraction (RE) (Kwiatkowski et al., 2019; Joshi et al., 2017; Yang et al., 2018; Rajpurkar et al., 2016; Tsatsaronis et al., 2015; Kembhavi et al., 2017; Levy et al., 2017).

Models We use open-source auto-regressive language models, including LLAMA2 (7B & 13B, Touvron et al., 2023), LLAMA3-8B (Grattafiori et al., 2024), MISTRAL-7B V0.3 (Jiang et al., 2023), and QWEN 2.5 (1.5B & 3B & 7B & 14B, Yang et al., 2024). Training-based methods are evaluated in a zero-shot setting, whereas inference-only methods utilize two-shot demonstrations. More details on datasets and templates are in Appendix A.

Training Details For a fair comparison, all training-based methods share the same settings. Utilizing the training set of NQ and TriviaQA, we randomly sample 250 questions from each of $\exists_{\text{PK}}$ and $\varnothing_{\text{PK}}$, resulting in a total of 1,000 samples. As illustrated in Figure 3, we pair each q with four context types, resulting in 4,000 question-context pairs. Appendix B.2 provides additional details.

Evaluation Metrics We use Exact Match (EM) to assess whether the model's prediction aligns with the expected answer, which differs for each scenario (Section 3.2). Still, evaluating LM behavior on samples with *undefined* PK relevance (Section 3.3) is equally important. To reflect practical settings, we also evaluate the full samples and report the accuracy (`Acc`) and reliability (`Rely`) scores (Xu et al., 2024a). `Rely` captures both correctness and appropriate abstention, balancing `Acc` and truthfulness (`Truth`). `Truth` quantifies the proportion of responses that are either correct or abstained. `Rely` is high when LM provides correct answers and abstains appropriately, while penalizing both incorrect outputs and excessive abstention. The formulation of metrics is in Appendix B.3.

6 Results on UniKnow

6.1 Main Results

Figure 4 illustrates the performance across the four UniKnow scenarios and the overall averaged performance (`All`). To assess generalization across knowledge domains, we report EM scores averaged over all datasets, comprising two in-domain and five out-of-domain sets for training-based methods.

Broader scenario coverage leads to better overall results. $\text{LM}_{\text{UniKnow}}$, which covers all scenarios, achieves the best overall performance, followed by KAFT. Other methods, designed with a subset of scenarios, lead to limited performance gains, often falling below or only marginally above Naïve. Meanwhile, $\text{COIECD}_{\text{Prompt}}$ consistently outperforms both COIECD and Prompting in three out of four models, demonstrating the extensibility potential of existing methods. These results highlight the importance of equipping LMs with the ability to handle a diverse range of knowledge scenarios–an aspect that has not been systematically addressed in prior work.

Resolving conflicts with known knowledge is more challenging than incorporating new, unknown information. Compared to `C` scenario, the performance points in `E-Only` are more tightly clustered with less variance. It demonstrates that LM behavior is influenced not only by context type itself, but also by its interaction with PK. Still, a similar trend is observed across methods in both `C` and `E-Only` scenarios. Notably, the performance drop of RetRobust is more pronounced in the `C` scenario than in `E-Only`, reflecting its limited ability to handle contradictory information effectively.

A trade-off between answering and abstention arises under irrelevant contexts. Methods that prioritize answerability without accounting for the presence of PK, such as COIECD, RetRobust, and KAFT, achieve strong performance in `P-Only` scenario. However, in `U` scenario, they are more

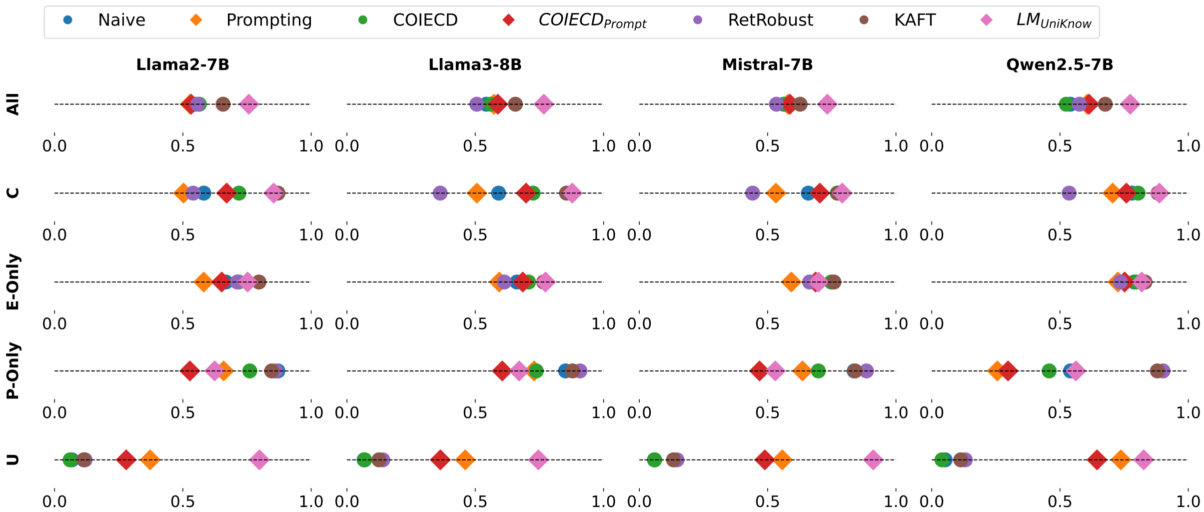

Figure 4: EM scores by scenario and model. `All` indicates scores averaged across all scenarios. Methods marked with diamonds incorporate abstention, while those with circle markers do not.

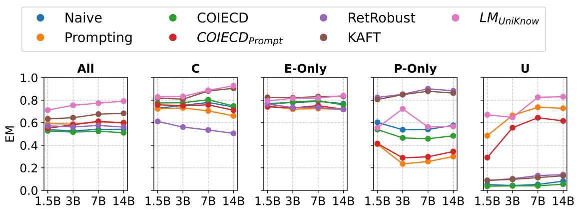

Figure 5: EM scores of QWEN models across different sizes, shown by scenario.

likely to generate hallucinations. In contrast, methods that incorporate abstention ability, including Prompting, COIECD$_{\text{Prompt}}$, and LM$_{\text{UniKnow}}$, handle `U` with abstention behavior, but suffer in a trade-off of exhibiting lower performance in `P-Only`. Among these, LM$_{\text{UniKnow}}$ demonstrates the largest performance gain in `U` scenario, driven by its consideration of the model's knowledge state.

Larger LMs generally improve reliability, with distinct trends across scenarios. Based on Figure 5, the performance in `E-Only` scenario remains relatively unaffected by scale, suggesting that EK utilization does not strongly benefit from larger LMs. In `C` and `P-Only` scenarios, gains depend on whether the method is explicitly trained for those conditions. By contrast, in `U` scenario, abstention performance improves consistently with scale, indicating that larger LMs are better at recognizing knowledge limitations and abstaining accordingly.

6.2 Impact of Context Types

Figure 6 presents a deeper analysis of model performance with LLAMA3-8B, illustrating its behavior across various context types within each scenario. Overall, our analysis reveals that progressively incorporating more context types and scenarios leads to a more comprehensive coverage. This is evident in the enhanced performance observed from RetRobust to KAFT in `C (Conflicting)` and `E-Only (Conflicting)` cases. Despite these improvements, a persistent challenge remains in mitigating the inherent trade-off between answerability and abstention ability.

For irrelevant contexts, randomly sampled (Random) and incorrectly retrieved (Incorrect-Ret.) contexts, LMs with *inference-time* knowledge utilization methods tend to perform worse on Incorrect-Ret. when answering the question in `P-Only`. A similar pattern is observed concerning abstention ability under `U` for Prompting, COIECD$_{\text{Prompt}}$, and LM$_{\text{UniKnow}}$. These findings indicate that misleading retrieved contexts challenge LMs not only in terms of answerability but also in their ability to abstain appropriately.

6.3 Error Analysis

Since LMs may exhibit scenario-specific biases, we analyze output errors to examine such patterns in detail. Incorrect responses are categorized into four types: contextual, parametric, false abstention, and others. *Contextual errors* occur when the model generates an incorrect response grounded on the given context. In case of relevant context, this involves extracting incorrect information; in the case of irrelevant content, the model is misled by unrelated content. *Parametric errors* refer to errors

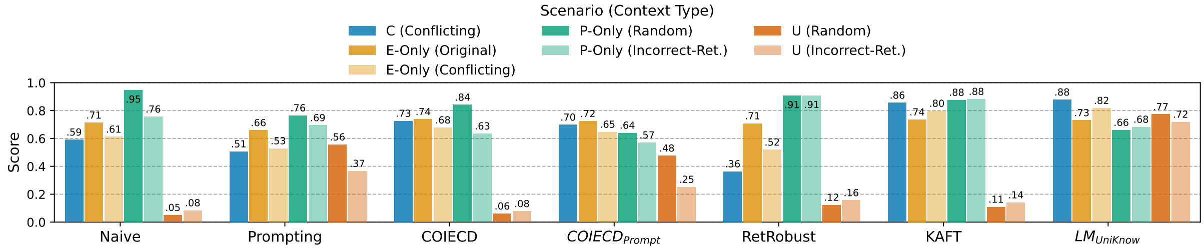

Figure 6: EM scores across different context types within UniKnow scenarios, evaluated using LLAMA3-8B.

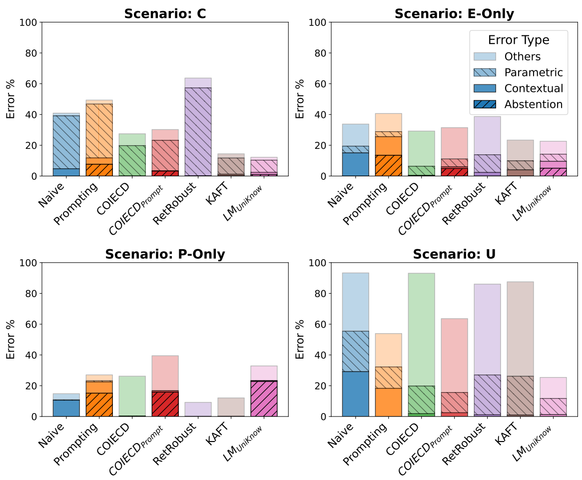

Figure 7: Stacked error type distributions across methods for each knowledge scenario. Transparency reflects error type. Evaluated using LLAMA3-8B.

generated based on the model's PK. In the `C` scenario, this reflects the model's failure to follow the given context, exhibiting a parametric bias. *False abstention* is counted as an error in three scenarios where the model possesses at least one relevant knowledge, except `U`. *Other* includes incorrect responses that do not fall into the above categories. Figure 7 shows the error distribution for Llama 3 8B across the four knowledge-handling scenarios.

Over-reliance on PK depends on the presence of PK. In the `C` scenario, where the model possesses the relevant information, all methods exhibit the highest rate of parametric errors compared to other error types. In contrast, such error is much less common in `E-Only` scenario. Even with COIECD, which explicitly targets knowledge conflict, the rate of parametric error remains significantly higher in `C` than in `E-Only`. Unlike prior works that focus solely on controlling EK via conflicting contexts, our findings highlight that over-reliance becomes more evident when scenarios are further distinguished by the presence of PK.

Contextual errors are rare across most methods, except for naïve approaches. In naïve approaches, contextual errors are observed in all scenarios, particularly in `E-Only` and `U`. This indicates that when the required knowledge is absent from the model's parametric memory, it tends to rely on the provided context but often fails to utilize it correctly (`E-Only`) or is misled by irrelevant information (`U`). In contrast, most other methods effectively mitigate context misinterpretation, as evidenced by the near absence of contextual errors.

Abstention error occurs most frequently in `P-Only` scenario, while it is rare under relevant contexts. Methods guided to abstain appropriately tend to exhibit relatively high abstention bias in `P-Only`. This again highlights the importance of the trade-off mitigation. Interestingly, the abstention error rate of $\text{COIECD}_{\text{Prompt}}$ remains comparable to that of Prompting in `P-Only`, but is significantly reduced in `E-Only`. This indicates that combining the strengths of COIECD and Prompting leads to more proper abstention across scenarios.

7 Additional Analysis on Reliability

Figure 8 visualizes the `Acc` and `Rely` scores for each method. Despite including *undefined* samples in the evaluation, the overall trend in `Rely` scores remains consistent with the scenario-averaged results in UniKnow (`All` in Figure 2). Note that methods on the dotted line, where `Acc` equals `Rely`, limit their performance in terms of answerability. $\text{LM}_{\text{UniKnow}}$ achieves the highest `Rely`, and its `Acc` remains comparable to methods which primarily focus on answerability. This suggests that, through alignment with UniKnow, $\text{LM}_{\text{UniKnow}}$ effectively minimizes incorrect responses via abstention while maintaining adaptability to various scenarios.

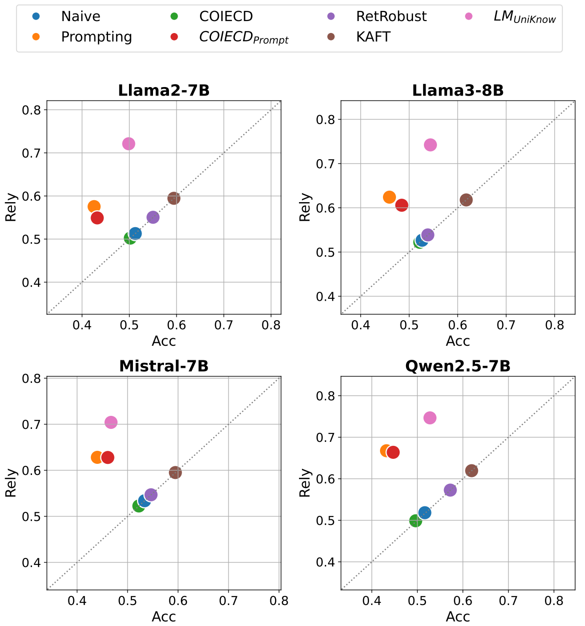

Figure 8: Acc and Rely scores across models. Each point represents a method averaged over all datasets. The dotted line indicates equal values of Acc and Rely.

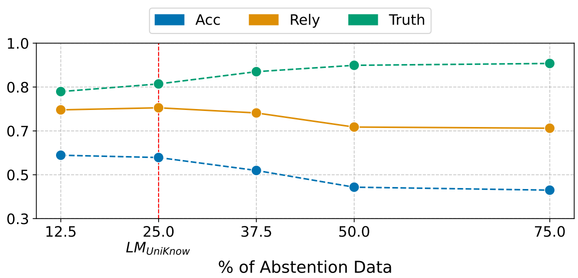

Figure 9: Effect of varying the proportion of abstention data on model performance for LLAMA3-8B. The red dashed line indicates the proportion used in LM$_{\text{UniKnow}}$.

7.1 Impact of Abstention Data

LM$_{\text{UniKnow}}$ allocates an equal proportion (25%) to each of the four scenarios within UniKnow. To investigate the effect of abstention supervision, we conduct an ablation study using LLAMA3-8B by varying the proportion of samples from U scenario. With a fixed number of training samples, we adjust the proportions of the remaining three scenarios equally. From Figure 9, we observe a trade-off between Acc and Truth as the proportion of abstention data increases. The lower proportions of abstention data lead to higher Acc, while higher proportions improve Truth. This reflects the inherent trade-off between maximizing correct answer generation (Acc) and minimizing incorrect outputs through abstention (Truth). Notably, the equal allocation across the four scenarios—25% abstention data (LM$_{\text{UniKnow}}$)—achieves the highest Rely score, indicating a balanced performance between answering correctly and abstaining appropriately.

Dataset / Metric	TriviaQA Acc	TriviaQA Truth	TriviaQA Rely	NQ Acc	NQ Truth	NQ Rely
LM$_{\text{UniKnow}}$	**0.6915**	**0.8762**	**0.8421**	**0.5396**	**0.8161**	**0.7396**
−C	0.6695	0.7040	0.7028	0.4987	0.6430	0.6222
−IR	0.6872	0.7352	0.7329	0.5056	0.6410	0.6227
−C, IR	0.6836	0.7084	0.7078	0.4987	0.6406	0.6205

Table 2: Ablation study on context types in the training data for LLAMA3-8B, measuring the impact of excluding conflicting contexts (−C), incorrectly retrieved contexts (−IR), or both (−C, IR). **Bold** indicates the best.

7.2 Impact of Context Type Diversity

We conduct an ablation study in which specific types of contexts are selectively removed, while maintaining the total number of training data. We consider three ablation settings: (1) −C, which excludes conflicting contexts and replaces them with original contexts; (2) −IR, which removes incorrectly retrieved contexts and retains only randomly sampled irrelevant contexts; and (3) −C, IR, which excludes both conflicting and incorrectly retrieved contexts. These settings allow us to isolate the contribution of each context type to overall reliability. As shown in Table 2, excluding conflicting or incorrectly retrieved contexts results in a noticeable drop in Truth and Rely, while having minimal impact on Acc. These findings underscore the importance of incorporating diverse context types, reflecting those encountered in practical settings, to enhance the reliability of knowledge-handling.

8 Conclusion

We present UniKnow, a unified framework for evaluating LM reliability across PK and EK. By systematically defining scenarios based on knowledge relevance, UniKnow enables fine-grained analysis of LM behavior. This comprehensive framework also highlights novel challenges, requiring LMs to navigate scenarios demanding diverse objectives and self-assessment of knowledge relevance. Our experiments reveal that existing methods often struggle to jointly handle scenarios and exhibit scenario-specific biases. We show that training with UniKnow-aligned supervision improves reliability, particularly evident in U scenario. Overall, UniKnow provides a foundation for building reliable LMs in knowledge utilization.

Limitations

Scope of Knowledge Tasks We primarily focus on the QA task, which provides a clear view of knowledge requirements and serves as a representative of knowledge-intensive tasks. Nevertheless, extending the scope to other tasks–such as reasoning (Xiong et al., 2024) or claim verification (Hagström et al., 2024)–is crucial, since the influence of knowledge sources may vary depending on the task. Additionally, we adopt a simplified RAG setting in which a single context is provided per query, allowing fine-grained control over context relevance and supporting targeted analysis of LM behavior. However, in real-world applications, LMs often receive multiple retrieved contexts simultaneously. This introduces new challenges, such as conflicts between external contexts (Xu et al., 2024b). Incorporating diverse tasks and extending UniKnow to support multi-context would be a valuable step toward modeling more complex and realistic RAG scenarios.

Factuality of External knowledge This study assumes that external knowledge is factually accurate, considering scenarios involving changed or newly emerging facts (Longpre et al., 2021; Xie et al., 2023). While this assumption enables controlled analysis, it may be strong in practice, as the quality of external knowledge depends heavily on the underlying database and retrieval system. The research area of factuality verification in external contexts using LLMs (Yu et al., 2024a; Fatahi Bayat et al., 2023) is closely related to this limitation. Exploring this aspect in conjunction with our framework could further strengthen the setting of the framework.

Limited Strategies for UniKnow-Aware Training Our study focuses on demonstrating the potential of UniKnow-aware supervised fine-tuning to equip LMs with comprehensive knowledge utilization capabilities. While we adopted supervised fine-tuning following prior research, future work could explore alternative training techniques, such as direct preference optimization or reward-based fine-tuning (Rafailov et al., 2023; Tian et al., 2024). Broadening the scope of training strategies may yield deeper insights into optimizing LM behavior across scenarios and improving reliability. Additionally, we leave the exploration of trends beyond the 14B model scale or reasoning-oriented LMs (DeepSeek-AI, 2025) to future work, as these may further impact behavior in knowledge-intensive tasks. We consider the knowledge handling capabilities of recently emerging reasoning LMs, particularly those with self-reflection, to be a valuable research direction that merits dedicated investigation within UniKnow.

References

Alfonso Amayuelas, Kyle Wong, Liangming Pan, Wenhu Chen, and William Wang. 2024a. Knowledge of knowledge: Exploring known-unknowns uncertainty with large language models. *Preprint*, arXiv:2305.13712.

Alfonso Amayuelas, Kyle Wong, Liangming Pan, Wenhu Chen, and William Yang Wang. 2024b. Knowledge of knowledge: Exploring known-unknowns uncertainty with large language models. In *Findings of the Association for Computational Linguistics: ACL 2024*, pages 6416–6432, Bangkok, Thailand. Association for Computational Linguistics.

Akari Asai, Sewon Min, Zexuan Zhong, and Danqi Chen. 2023. Retrieval-based language models and applications. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 6: Tutorial Abstracts)*, pages 41–46, Toronto, Canada. Association for Computational Linguistics.

Akari Asai, Zeqiu Wu, Yizhong Wang, Avirup Sil, and Hannaneh Hajishirzi. 2024. Self-RAG: Learning to retrieve, generate, and critique through self-reflection. In *The Twelfth International Conference on Learning Representations*.

Jinheon Baek, Soyeong Jeong, Minki Kang, Jong Park, and Sung Hwang. 2023. Knowledge-augmented language model verification. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 1720–1736, Singapore. Association for Computational Linguistics.

Yejin Bang, Ziwei Ji, Alan Schelten, Anthony Hartshorn, Tara Fowler, Cheng Zhang, Nicola Cancedda, and Pascale Fung. 2025. Hallulens: Llm hallucination benchmark. *Preprint*, arXiv:2504.17550.

Danqi Chen, Adam Fisch, Jason Weston, and Antoine Bordes. 2017. Reading Wikipedia to answer open-domain questions. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1870–1879, Vancouver, Canada. Association for Computational Linguistics.

Sitao Cheng, Liangming Pan, Xunjian Yin, Xinyi Wang, and William Yang Wang. 2024. Understanding the interplay between parametric and contextual knowledge for large language models. *Preprint*, arXiv:2410.08414.

Pradeep Dasigi, Kyle Lo, Iz Beltagy, Arman Cohan, Noah A. Smith, and Matt Gardner. 2021. A dataset of information-seeking questions and answers anchored in research papers. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 4599–4610, Online. Association for Computational Linguistics.

DeepSeek-AI. 2025. Deepseek-r1: Incentivizing reasoning capability in llms via reinforcement learning. *Preprint*, arXiv:2501.12948.

Tim Dettmers, Artidoro Pagnoni, Ari Holtzman, and Luke Zettlemoyer. 2023. QLoRA: Efficient finetuning of quantized LLMs. In *Thirty-seventh Conference on Neural Information Processing Systems*.

Bhuwan Dhingra, Jeremy R. Cole, Julian Martin Eisenschlos, Daniel Gillick, Jacob Eisenstein, and William W. Cohen. 2022. Time-aware language models as temporal knowledge bases. *Transactions of the Association for Computational Linguistics*, 10:257–273.

Farima Fatahi Bayat, Kun Qian, Benjamin Han, Yisi Sang, Anton Belyy, Samira Khorshidi, Fei Wu, Ihab Ilyas, and Yunyao Li. 2023. FLEEK: Factual error detection and correction with evidence retrieved from external knowledge. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 124–130, Singapore. Association for Computational Linguistics.

Shangbin Feng, Weijia Shi, Yike Wang, Wenxuan Ding, Vidhisha Balachandran, and Yulia Tsvetkov. 2024. Don't hallucinate, abstain: Identifying LLM knowledge gaps via multi-LLM collaboration. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 14664–14690, Bangkok, Thailand. Association for Computational Linguistics.

Adam Fisch, Alon Talmor, Robin Jia, Minjoon Seo, Eunsol Choi, and Danqi Chen. 2019. Mrqa 2019 shared task: Evaluating generalization in reading comprehension. *Preprint*, arXiv:1910.09753.

Aaron Grattafiori, Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Alex Vaughan, Amy Yang, Angela Fan, Anirudh Goyal, Anthony Hartshorn, Aobo Yang, Archi Mitra, Archie Sravankumar, Artem Korenev, Arthur Hinsvark, and 542 others. 2024. The llama 3 herd of models. *Preprint*, arXiv:2407.21783.

Kelvin Guu, Kenton Lee, Zora Tung, Panupong Pasupat, and Mingwei Chang. 2020. Retrieval augmented language model pre-training. In *Proceedings of the 37th International Conference on Machine Learning*, volume 119 of *Proceedings of Machine Learning Research*, pages 3929–3938. PMLR.

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. *Preprint*, arXiv:2412.17031.

Yuheng Huang, Jiayang Song, Zhijie Wang, Shengming Zhao, Huaming Chen, Felix Juefei-Xu, and Lei Ma. 2025. Look before you leap: An exploratory study of uncertainty analysis for large language models. *IEEE Transactions on Software Engineering*, 51(2):413–429.

Gautier Izacard, Mathilde Caron, Lucas Hosseini, Sebastian Riedel, Piotr Bojanowski, Armand Joulin, and Edouard Grave. 2022. Unsupervised dense information retrieval with contrastive learning. *Preprint*, arXiv:2112.09118.

Albert Q. Jiang, Alexandre Sablayrolles, Arthur Mensch, Chris Bamford, Devendra Singh Chaplot, Diego de las Casas, Florian Bressand, Gianna Lengyel, Guillaume Lample, Lucile Saulnier, Lélio Renard Lavaud, Marie-Anne Lachaux, Pierre Stock, Teven Le Scao, Thibaut Lavril, Thomas Wang, Timothée Lacroix, and William El Sayed. 2023. Mistral 7b. *Preprint*, arXiv:2310.06825.

Qiao Jin, Bhuwan Dhingra, Zhengping Liu, William Cohen, and Xinghua Lu. 2019. Pubmedqa: A dataset for biomedical research question answering. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 2567–2577.

Zhuoran Jin, Pengfei Cao, Yubo Chen, Kang Liu, Xiaojian Jiang, Jiexin Xu, Li Qiuxia, and Jun Zhao. 2024a. Tug-of-war between knowledge: Exploring and resolving knowledge conflicts in retrieval-augmented language models. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 16867–16878, Torino, Italia. ELRA and ICCL.

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. Association for Computational Linguistics.

Mandar Joshi, Eunsol Choi, Daniel S. Weld, and Luke Zettlemoyer. 2017. Triviaqa: A large scale distantly supervised challenge dataset for reading comprehension. *Preprint*, arXiv:1705.03551.

Saurav Kadavath, Tom Conerly, Amanda Askell, Tom Henighan, Dawn Drain, Ethan Perez, Nicholas Schiefer, Zac Hatfield-Dodds, Nova DasSarma, Eli Tran-Johnson, Scott Johnston, Sheer El-Showk, Andy Jones, Nelson Elhage, Tristan Hume, Anna

Chen, Yuntao Bai, Sam Bowman, Stanislav Fort, and 17 others. 2022. Language models (mostly) know what they know. *Preprint*, arXiv:2207.05221.

Nikhil Kandpal, Haikang Deng, Adam Roberts, Eric Wallace, and Colin Raffel. 2023. Large language models struggle to learn long-tail knowledge. In *Proceedings of the 40th International Conference on Machine Learning*, volume 202 of *Proceedings of Machine Learning Research*, pages 15696–15707. PMLR.

Jungo Kasai, Keisuke Sakaguchi, yoichi takahashi, Ronan Le Bras, Akari Asai, Xinyan Yu, Dragomir Radev, Noah A Smith, Yejin Choi, and Kentaro Inui. 2023. Realtime qa: What's the answer right now? In *Advances in Neural Information Processing Systems*, volume 36, pages 49025–49043. Curran Associates, Inc.

Aniruddha Kembhavi, Minjoon Seo, Dustin Schwenk, Jonghyun Choi, Ali Farhadi, and Hannaneh Hajishirzi. 2017. Are you smarter than a sixth grader? textbook question answering for multimodal machine comprehension. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*.

Hyuhng Joon Kim, Youna Kim, Sang goo Lee, and Taeuk Kim. 2025. When to speak, when to abstain: Contrastive decoding with abstention. *Preprint*, arXiv:2412.12527.

Lorenz Kuhn, Yarin Gal, and Sebastian Farquhar. 2023. Semantic uncertainty: Linguistic invariances for uncertainty estimation in natural language generation. In *The Eleventh International Conference on Learning Representations*.

Tom Kwiatkowski, Jennimaria Palomaki, Olivia Redfield, Michael Collins, Ankur Parikh, Chris Alberti, Danielle Epstein, Illia Polosukhin, Jacob Devlin, Kenton Lee, Kristina Toutanova, Llion Jones, Matthew Kelcey, Ming-Wei Chang, Andrew M. Dai, Jakob Uszkoreit, Quoc Le, and Slav Petrov. 2019. Natural questions: A benchmark for question answering research. *Transactions of the Association for Computational Linguistics*, 7:452–466.

Omer Levy, Minjoon Seo, Eunsol Choi, and Luke Zettlemoyer. 2017. Zero-shot relation extraction via reading comprehension. *Preprint*, arXiv:1706.04115.

Mike Lewis, Yinhan Liu, Naman Goyal, Marjan Ghazvininejad, Abdelrahman Mohamed, Omer Levy, Veselin Stoyanov, and Luke Zettlemoyer. 2020. BART: Denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 7871–7880, Online. Association for Computational Linguistics.

Daliang Li, Ankit Singh Rawat, Manzil Zaheer, Xin Wang, Michal Lukasik, Andreas Veit, Felix Yu, and Sanjiv Kumar. 2023. Large language models with controllable working memory. In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 1774–1793, Toronto, Canada. Association for Computational Linguistics.

Adam Liska, Tomas Kocisky, Elena Gribovskaya, Tayfun Terzi, Eren Sezener, Devang Agrawal, Cyprien De Masson D'Autume, Tim Scholtes, Manzil Zaheer, Susannah Young, Ellen Gilsenan-Mcmahon, Sophia Austin, Phil Blunsom, and Angeliki Lazaridou. 2022. StreamingQA: A benchmark for adaptation to new knowledge over time in question answering models. In *Proceedings of the 39th International Conference on Machine Learning*, volume 162 of *Proceedings of Machine Learning Research*, pages 13604–13622. PMLR.

Yantao Liu, Zijun Yao, Xin Lv, Yuchen Fan, Shulin Cao, Jifan Yu, Lei Hou, and Juanzi Li. 2024. Untangle the KNOT: Interweaving conflicting knowledge and reasoning skills in large language models. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 17186–17204, Torino, Italia. ELRA and ICCL.

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. Association for Computational Linguistics.

I. Loshchilov and F. Hutter. 2017. Decoupled weight decay regularization. *International Conference on Learning Representations*.

Hongyin Luo, Tianhua Zhang, Yung-Sung Chuang, Yuan Gong, Yoon Kim, Xixin Wu, Helen Meng, and James Glass. 2023. Search augmented instruction learning. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 3717–3729, Singapore. Association for Computational Linguistics.

Alex Mallen, Akari Asai, Victor Zhong, Rajarshi Das, Daniel Khashabi, and Hannaneh Hajishirzi. 2023. When not to trust language models: Investigating effectiveness of parametric and non-parametric memories. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 9802–9822, Toronto, Canada. Association for Computational Linguistics.

Ella Neeman, Roee Aharoni, Or Honovich, Leshem Choshen, Idan Szpektor, and Omri Abend. 2023. DisentQA: Disentangling parametric and contextual knowledge with counterfactual question answering. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 10056–10070, Toronto, Canada. Association for Computational Linguistics.

Seong-Il Park, Seung-Woo Choi, Na-Hyun Kim, and Jay-Yoon Lee. 2024. Enhancing robustness of retrieval-augmented language models with in-context learning. *KNOWLEDGENLP*.

Rafael Rafailov, Archit Sharma, Eric Mitchell, Christopher D Manning, Stefano Ermon, and Chelsea Finn. 2023. Direct preference optimization: Your language model is secretly a reward model. In *Thirty-seventh Conference on Neural Information Processing Systems*.

Pranav Rajpurkar, Jian Zhang, Konstantin Lopyrev, and Percy Liang. 2016. Squad: 100,000+ questions for machine comprehension of text. *Preprint*, arXiv:1606.05250.

Xiaoyu Shen, Rexhina Blloshmi, Dawei Zhu, Jiahuan Pei, and Wei Zhang. 2024. Assessing "implicit" retrieval robustness of large language models. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 8988–9003, Miami, Florida, USA. Association for Computational Linguistics.

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. Association for Computational Linguistics.

Zhaochen Su, Jun Zhang, Xiaoye Qu, Tong Zhu, Yanshu Li, Jiashuo Sun, Juntao Li, Min Zhang, and Yu Cheng. 2024. $\texttt{ConflictBank}$: A benchmark for evaluating the influence of knowledge conflicts in LLMs. In *The Thirty-eight Conference on Neural Information Processing Systems Datasets and Benchmarks Track*.

Hexiang Tan, Fei Sun, Wanli Yang, Yuanzhuo Wang, Qi Cao, and Xueqi Cheng. 2024. Blinded by generated contexts: How language models merge generated and retrieved contexts when knowledge conflicts? In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 6207–6227, Bangkok, Thailand. Association for Computational Linguistics.

Katherine Tian, Eric Mitchell, Huaxiu Yao, Christopher D Manning, and Chelsea Finn. 2024. Fine-tuning language models for factuality. In *The Twelfth International Conference on Learning Representations*.

Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, Dan Bikel, Lukas Blecher, Cristian Canton Ferrer, Moya Chen, Guillem Cucurull, David Esiobu, Jude Fernandes, Jeremy Fu, Wenyin Fu, and 49 others. 2023. Llama 2: Open foundation and fine-tuned chat models. *Preprint*, arXiv:2307.09288.

George Tsatsaronis, Georgios Balikas, Prodromos Malakasiotis, Ioannis Partalas, Matthias Zschunke, Michael R Alvers, Dirk Weissenborn, Anastasia Krithara, Sergios Petridis, Dimitris Polychronopoulos, and 1 others. 2015. An overview of the bioasq large-scale biomedical semantic indexing and question answering competition. *BMC bioinformatics*, 16:1–28.

Yike Wang, Shangbin Feng, Heng Wang, Weijia Shi, Vidhisha Balachandran, Tianxing He, and Yulia Tsvetkov. 2024a. Resolving knowledge conflicts in large language models. In *First Conference on Language Modeling*.

Yuxia Wang, Minghan Wang, Muhammad Arslan Manzoor, Fei Liu, Georgi Nenkov Georgiev, Rocktim Jyoti Das, and Preslav Nakov. 2024b. Factuality of large language models: A survey. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 19519–19529, Miami, Florida, USA. Association for Computational Linguistics.

Bingbing Wen, Bill Howe, and Lucy Lu Wang. 2024a. Characterizing LLM abstention behavior in science QA with context perturbations. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 3437–3450, Miami, Florida, USA. Association for Computational Linguistics.

Bingbing Wen, Jihan Yao, Shangbin Feng, Chenjun Xu, Yulia Tsvetkov, Bill Howe, and Lucy Lu Wang. 2024b. Know your limits: A survey of abstention in large language models. *arXiv preprint arXiv: 2407.18418*.

Bingbing Wen, Jihan Yao, Shangbin Feng, Chenjun Xu, Yulia Tsvetkov, Bill Howe, and Lucy Lu Wang. 2025. Know your limits: A survey of abstention in large language models. *Preprint*, arXiv:2407.18418.

Siye Wu, Jian Xie, Jiangjie Chen, Tinghui Zhu, Kai Zhang, and Yanghua Xiao. 2024. How easily do irrelevant inputs skew the responses of large language models? In *First Conference on Language Modeling*.

Yuan Xia, Jingbo Zhou, Zhenhui Shi, Jun Chen, and Haifeng Huang. 2024. Improving retrieval augmented language model with self-reasoning. *arXiv preprint arXiv: 2407.19813*.

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. *International Conference on Learning Representations*.

Siheng Xiong, Ali Payani, Ramana Kompella, and Faramarz Fekri. 2024. Large language models can learn temporal reasoning. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 10452–10470, Bangkok, Thailand. Association for Computational Linguistics.

Hongshen Xu, Zichen Zhu, Situo Zhang, Da Ma, Shuai Fan, Lu Chen, and Kai Yu. 2024a. Rejection improves reliability: Training LLMs to refuse unknown questions using RL from knowledge feedback. In *First Conference on Language Modeling*.

Rongwu Xu, Zehan Qi, Zhijiang Guo, Cunxiang Wang, Hongru Wang, Yue Zhang, and Wei Xu. 2024b. 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. Association for Computational Linguistics.

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, Jiaxi Yang, Jingren Zhou, Junyang Lin, Kai Dang, and 22 others. 2024. Qwen2.5 technical report. *arXiv preprint arXiv:2412.15115*.

Zhilin Yang, Peng Qi, Saizheng Zhang, Yoshua Bengio, William W. Cohen, Ruslan Salakhutdinov, and Christopher D. Manning. 2018. Hotpotqa: A dataset for diverse, explainable multi-hop question answering. *Preprint*, arXiv:1809.09600.

Ori Yoran, Tomer Wolfson, Ori Ram, and Jonathan Berant. 2023. Making retrieval-augmented language models robust to irrelevant context. *Preprint*, arXiv:2310.01558.

Tian Yu, Shaolei Zhang, and Yang Feng. 2024a. Truth-aware context selection: Mitigating hallucinations of large language models being misled by untruthful contexts. In *Findings of the Association for Computational Linguistics: ACL 2024*, pages 10862–10884, Bangkok, Thailand. Association for Computational Linguistics.

Wenhao Yu, Hongming Zhang, Xiaoman Pan, Peixin Cao, Kaixin Ma, Jian Li, Hongwei Wang, and Dong Yu. 2024b. Chain-of-note: Enhancing robustness in retrieval-augmented language models. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 14672–14685, Miami, Florida, USA. Association for Computational Linguistics.

Xiaowei Yuan, Zhao Yang, Yequan Wang, Shengping Liu, Jun Zhao, and Kang Liu. 2024. Discerning and resolving knowledge conflicts through adaptive decoding with contextual information-entropy constraint. In *Findings of the Association for Computational Linguistics: ACL 2024*, pages 3903–3922, Bangkok, Thailand. Association for Computational Linguistics.

Hanning Zhang, Shizhe Diao, Yong Lin, Yi Fung, Qing Lian, Xingyao Wang, Yangyi Chen, Heng Ji, and Tong Zhang. 2024a. R-tuning: Instructing large language models to say 'I don't know'. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 7113–7139, Mexico City, Mexico. Association for Computational Linguistics.

Xiaoying Zhang, Baolin Peng, Ye Tian, Jingyan Zhou, Lifeng Jin, Linfeng Song, Haitao Mi, and Helen Meng. 2024b. Self-alignment for factuality: Mitigating hallucinations in LLMs via self-evaluation. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1946–1965, Bangkok, Thailand. Association for Computational Linguistics.

Appendix

A UniKnow: Additional Details

A.1 Problem Settings

In this paper, we consider two knowledge sources: parametric knowledge (PK) and external knowledge (EK). **PK** acquired during pretraining is inherently bounded by its pretraining data. Given a question q, "Who is the president of the United States?", if the LM's knowledge cutoff is before 2024, the answer grounded in the LM's pretraining data (a^*_{PK}) is "Biden," for instance. If the LM possesses the relevant information "Biden is the president of the United States," then q is considered $\exists_{\text{PK}}$. However, if the LM answers with the name "Michael Jackson," which is irrelevant information, q is treated as $\varnothing_{\text{PK}}$. This is because LM("Michael Jackson" $\mid q) \neq a^*_{\text{PK}}$, since "Michael Jackson" was never a president of the United States (Bang et al., 2025).

EK can reflect user intent by incorporating user-specified or task-relevant information and provide enriched information unavailable within the parametric knowledge, particularly long-tail and updated or changed facts. Ideally, the relevant external knowledge serves to complement or override the parametric knowledge, enabling user-guided, up-to-date model responses. However, in practice, there is no guarantee that the provided context will always be relevant, since relevance depends on the quality of the retrieval mechanism (Izacard et al., 2022; Guu et al., 2020). To isolate the effect of knowledge relevance, we make a simplifying assumption that the external knowledge is always factually aligned with world knowledge, since its factuality is determined by the underlying database in practice.

This way, our study is based on two key conditions: (1) PK may not align with the most recent world knowledge but can still possess relevant knowledge to answer q. (2) Judging the factuality of EK is not within the scope of our study. Considering real-world usage, the LM is expected to utilize the best relevant knowledge based on its learned knowledge and respond faithfully to the given context. The responsibility of determining the factuality of external knowledge ultimately rests on the quality of the underlying database and retrieval system.

A.2 Datasets

We use the dataset versions curated by the Machine Reading for Question Answering (MRQA) benchmark (Fisch et al., 2019). The total number of samples for each dataset is in Table 3. Each sample includes a question, original answer, conflicting answer, and four types of context: original, conflicting, random, and incorrectly retrieved contexts. We provide a detailed description of the datasets used in our study below[2].

NaturalQuestions (Kwiatkowski et al., 2019) Questions consist of real queries issued to the Google search engine. From a Wikipedia page from the top 5 search results, annotators select a long answer containing enough information to completely infer the answer to the question, and a short answer that comprises the actual answer. The long answer becomes the context matched with the question, while the short answer is used as the answer.

TriviaQA (Joshi et al., 2017) Question-answer pairs are authored by trivia enthusiasts and independently gathered evidence documents that provide high quality supervision for answering the questions. The web version of TriviaQA is used, where the contexts are retrieved from the results of a Bing search query.

HotpotQA (Yang et al., 2018) Questions are diverse and not constrained to any pre-existing knowledge base. Multi-hop reasoning is required to solve the questions. Paragraphs that provide supporting facts required for reasoning, are given along with the question. In the original setting, additional distractor paragraphs are augmented in order to increase the difficulty of inference. However, these distractor paragraphs are not used in this setting.

SQuAD (Rajpurkar et al., 2016) Paragraphs from Wikipedia are presented to crowdworkers, and they are asked to write questions that entail extractive answers. The answer to each question is a segment of text from the corresponding reading passage. To remove the uncertainty that excessively long paragraphs bring, QA pairs that do not align with the first 800 tokens are discarded in this setting.

BioASQ (Tsatsaronis et al., 2015) BioASQ is a challenge that assesses the ability of systems to semantically index large numbers of biomedical

[2]Code and dataset will be available upon publication.

Dataset	Train	Test
NQ	83,787	3,994
TriviaQA	61,177	7,712
HotpotQA	-	4,760
SquAD	-	7,918
Bioasq	-	697
TextbookQA	-	1,056
RelationExtraction	-	1,974
Total	144,964	28,111

Table 3: Number of samples for each dataset.

```
Answer the following questions:
<few-shots>
Question: <question>
Answer:
```

Table 4: Template used in closed-book generation.

scientific articles and return concise answers to given natural language questions. Each question is linked to multiple related science articles. The full abstract of each linked article is used as an individual context. Abstracts that do not exactly contain the answer are discarded.

TextbookQA (Kembhavi et al., 2017) TextbookQA aims at answering multimodal questions when given a context in formats of text, diagrams and images. This dataset is collected from lessons from middle school Life Science, Earth Science, and Physical Science textbooks. Questions that are accompanied with a diagram and "True of False" questions are not used in this setting.

RelationExtraction (Levy et al., 2017) Given labeled slot-filling examples, relations between entities are transformed into QA pairs using templates. Multiple templates for each type of relation are utilized. The zero-shot benchmark split of this dataset, which showed that generalization to unseen relations is possible at lower accuracy levels, is used.

A.3 Predefined Abstention Words

The predefined abstain words (Amayuelas et al., 2024a) used in evaluations are: ['unanswerable', 'unknown', 'no known', 'not known', 'do not know' 'uncertain', 'unclear', 'no scientific evidence', 'no definitive answer', 'no right answer', 'no concrete answer', 'no public information', 'debate', 'impossible to know', 'impossible to answer', 'difficult to predict', 'not sure', 'irrelevant', 'not relevant']

```
Answer the following questions:
<few-shots>
Context: <context>
Question: <question>
Answer:
```

Table 5: Template for the naïve open-book generation.

```
Answer an entity of the same type as the given
keyword. Please note that the keyword is from
the given context, and consider the part of
speech of the keyword inside the context. You
should not give a synonym or alias of the given
keyword. The entity and given keyword must
have different meanings. Only answer the entity
itself without any extra phrases.
<few-shots>
Keyword: <original-answer>
Context: <context>
Answer:
```

Table 6: Template used when instructing the model to generate a conflicting answer, given the original answer and context.

A.4 Details on UniKnow Construction

As the impact of context length is beyond the scope of our study, we limit context to approximately 100 words to ensure experimental control. To ensure context informativeness and maintain experimental controllability, we have processed the original contexts from the MRQA benchmark by limiting their length and ensuring that the ground-truth answer span is always included. For each occurrence span of the ground-truth answer in the raw context, we take a 100-word portion surrounding that span and consider it a candidate context. We then compute the NLI (BART-LARGE, Lewis et al., 2020) score between the question-answer pair and each candidate context, and select the context with the highest NLI score as the original context.

To generate conflicting answers, Template 6 is employed. For retrieved-uninformative contexts, a Wikipedia dump from December 2018 is used as a database. Each context is chunked into 100 words. As a retriever model, CONTRIEVER-MSMARCO (Izacard et al., 2022) is utilized. The number of samples per scenario and model is provided in Table 10. Template 4 is used to perform closed-book generation for estimating the presence of parametric knowledge.

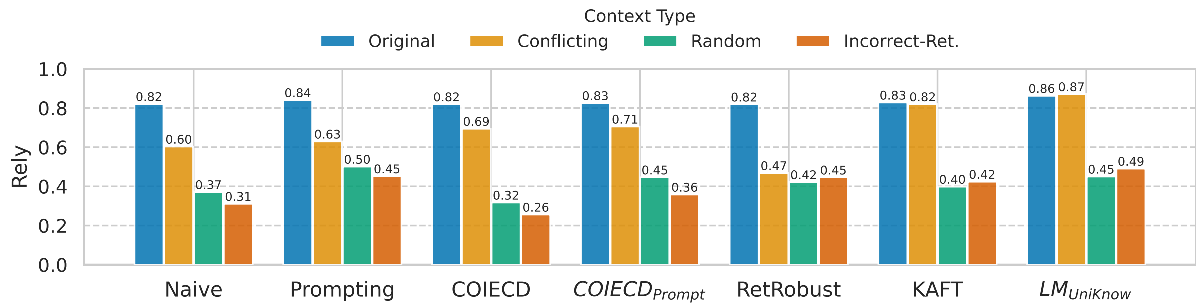

Figure 10: `Rely` scores across different context types, evaluated on the full test set using LLAMA3-8B.

```
Answer the following questions. The context
may or may not be helpful. If the context
is unhelpful and you are not knowledgeable
about the question, it is appropriate to say,
"<UNKNOWN>".
<few-shots>
Context: <context>
Question: <question>
Answer:
```

Table 7: Instruction for LMs to abstain if unknown.

```
Answer the following questions. If you are
not knowledgeable about the question, it is
appropriate to say, "<UNKNOWN>".
<few-shots>
Question: <question>
Answer:
```

Table 8: Instruction used in COIECD$_{\text{Prompt}}$ for LMs to abstain if unknown under closed-book generation.

B Experiential Setting Details

B.1 Knowledge Utilization Methods

Template 5 is used for naïve open-book generation, while Template 7 is applied in the prompting approach. For all experiments, greedy decoding is employed.

COIECD For COIECD, which requires two hyperparameters, we adopt the values reported in the original paper ($\alpha = 1.0$ and $\lambda = 0.25$), as Yuan et al. (2024) shows that these values generalize well across models and datasets.

COIECD$_{\text{Prompt}}$ In COIECD$_{\text{Prompt}}$, we use Template 7 for input with context and Template 8 for input without context.

KAFT Unlike Li et al., 2023, which treats the parametric answers as gold-standard for irrelevant contexts, we use the original answer to ensure fair evaluation in the `U` scenario.

B.2 Additional Training Details

As described in Section 5, all training-based methods (RetRobust, KAFT, and LM$_{\text{UniKnow}}$) are trained on the same set of q to ensure a fair comparison. In case of RetRobust, since it does not utilize conflicting contexts (Figure 3), we additionally sample 1,000 questions and pair them with the original context to match the overall training size. To maintain the LM's ability to answer when the context contains information that matches with its PK ($a^*_{\text{PK}} = a^*_{\text{EK}}$), we include the original context paired with $q \in \exists_{\text{PK}}$ during training.

We use the same setting for every training-based approach. For the main experiments, three seeds (12, 123, 1234) were used, and the results reported are averaged over these three seeds. Each model is trained for three epochs using the AdamW (Loshchilov and Hutter, 2017) optimizer with a learning rate of 0.0001 and a batch size of 16. For efficient fine-tuning, we employ QLoRA (Dettmers et al., 2023) with rank=4 and alpha=16. All training is conducted on two NVIDIA RTX A6000.

B.3 Evaluation Metrics

`Acc`, `Rely`, and `Truth` metrics are computed based on the number of correct (N_c), incorrect (N_i), and abstained (N_a) responses.[3] `Acc` measures the proportion of correct answers ($\frac{N_c}{N}$), while `Truth` captures the proportion of responses that are either correct or abstained ($\frac{N_c+N_a}{N}$), thereby rewarding safe behavior that avoids incorrect outputs. To discourage excessive abstention, the answer rate ($\texttt{Ans} = \frac{N_c+N_i}{N}$) is used as a weighting factor. Using this, `Rely` balances `Acc` and `Truth` and is computed as: $\texttt{Ans} \times \texttt{Truth} + (1 - \texttt{Ans}) \times \texttt{Acc}$. Thus, `Rely` reflects the overall reliability of the model by rewarding both correct answers and appropriate abstentions, while penalizing incorrect responses.

[3] N: The total number of responses.

C Additional Results

In this section, we provide exact values of figures and additional results for models not included in Section 6.1 and Section 7.

C.1 Main Results

The EM scores corresponding to Figure 4 are provided in Table 11. Also, Figure 11 visualizes the EM scores of LLAMA2 7B and 13B across different knowledge scenarios. Figure 12 illustrates the impact of model scale with `Rely` metric for LLAMA2 and QWEN2.5.

The exact values for the `Acc` and `Rely` scores presented in Figure 8 are listed in Table 12 per dataset. While Figure 8 presents overall trends averaged across all datasets, Figure 13 and Figure 14 break down the results by in-domain and out-of-domain datasets, respectively. They further highlight that the overall trend across methods holds consistently and generalizes well to out-of-domain settings.

C.2 Error Analysis

We present the error type distribution for each knowledge scenario across different models. Results for LLAMA2-7B, MISTRAL-7B, and QWEN2.5-7B are shown in Figure 15, Figure 16, and Figure 17, respectively.

C.3 Impact of Context Types

UniKnow incorporates diverse context types to evaluate LM behavior under varying degrees of contextual relevance. We further analyze model performance across different context types with LLAMA3-8B.

Figure 10 reveals that LMs exhibit markedly different performance depending on the type of context on the full test set. For relevant contexts–original and conflicting–most knowledge utilization methods, except KAFT and $\text{LM}_{\text{UniKnow}}$, demonstrate a substantial drop in `Rely` when the context contains conflicting information, while maintaining high performance when the original context is used. This suggests that LMs struggle to resolve conflicts between PK and EK. Notably, RetRobust, which is primarily designed to improve robustness against irrelevant context, shows a particularly pronounced decline under conflicting conditions.

For irrelevant contexts, including randomly sampled (Random) and incorrectly retrieved (Incorrect-Ret.) contexts, LMs with inference-time knowledge utilization methods tend to perform worse on Incorrect-Ret. This indicates LMs' sensitivity to misleading but plausibly relevant knowledge.

C.4 Ablation Study

Figure 18 shows the effect of varying the proportion of abstention data on the performance across datasets. These results align with the averaged trend discussed in Section 7.1, confirming that the observed pattern holds consistently across datasets. Table 9 shows the impact of context type diversity on additional datasets beyond those reported in Table 2.

D Related Works

Knowledge Conflict Parametric knowledge is inherently static, whereas external knowledge can be delivered in response to diverse circumstances. This dynamic provision often results in discrepancies between the parametric memory and the external context. Studies have examined the conflict through the lens of external knowledge features, such as temporal shifts (Kasai et al., 2023; Dhingra et al., 2022), synthetically updated facts (Longpre et al., 2021), and contextual plausibility (Xie et al., 2023; Tan et al., 2024). Yet many existing approaches (Liu et al., 2024; Wang et al., 2024a; Jin et al., 2024a) still treat any mismatch between model output and context as a conflict, often neglecting whether the model had prior access to that information.

Robustness against Irrelevance Although external knowledge is intended to supply LM's knowledge, in real-world scenarios (i.e. RAG), it may not always be relevant. LMs face challenges in handling irrelevant context, which often leads to performance degradation (Shen et al., 2024). RAG is particularly susceptible, as retrieval errors can introduce a misleading but plausible context (Wu et al., 2024). To mitigate this, researchers have explored methods to encourage LMs to rely on parametric knowledge when external information is irrelevant–either at inference time (Yu et al., 2024b; Park et al., 2024; Baek et al., 2023) or through training (Yoran et al., 2023; Asai et al., 2024; Xia et al., 2024; Luo et al., 2023).

Parametric Knowledge Estimation There is a line of work trying to estimate the knowledge boundaries of LMs. Some approaches quantify uncertainty in parametric knowledge through LM's

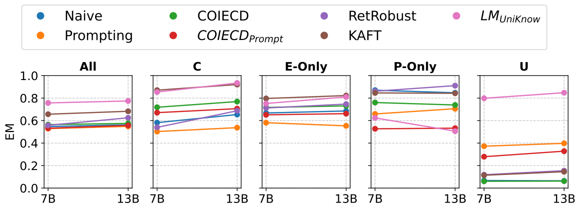

Figure 11: EM scores of LLAMA2 models across different sizes, averaged over all datasets within UniKnow.

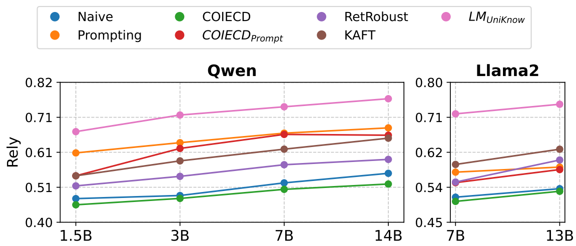

Figure 12: `Rely` scores of QWEN and LLAMA2 across model sizes.

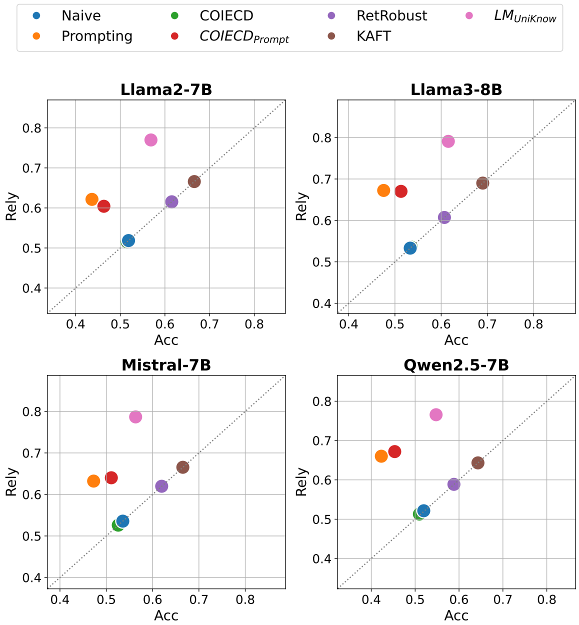

Figure 13: `Acc` and `Rely` scores averaged over in-domain datasets. Each point represents a method averaged over all datasets. The dotted line indicates equal values of `Acc` and `Rely`.

internal representations and output consistency (Huang et al., 2025; Kuhn et al., 2023; Kadavath et al., 2022). These are often used to relabel training data accordingly, guiding abstention behavior (Zhang et al., 2024a; Wen et al., 2024b) or selectively abstain from answering with a threshold (Feng et al., 2024).

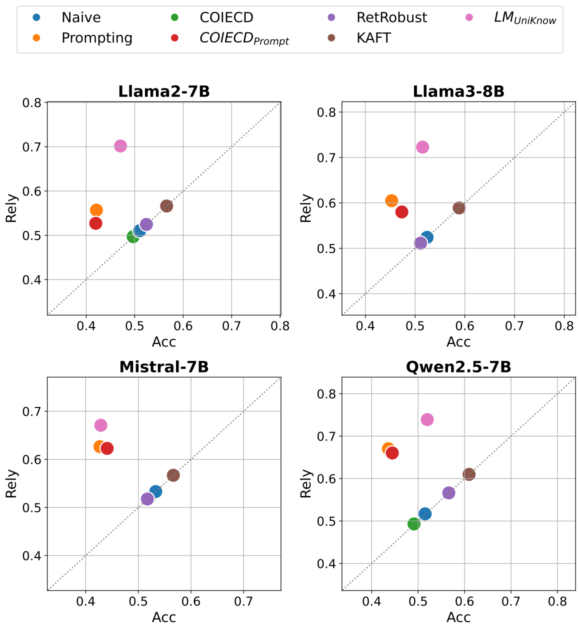

Figure 14: `Acc` and `Rely` scores averaged over out-of-domain datasets. Each point represents a method averaged over all datasets. The dotted line indicates equal values of `Acc` and `Rely`.

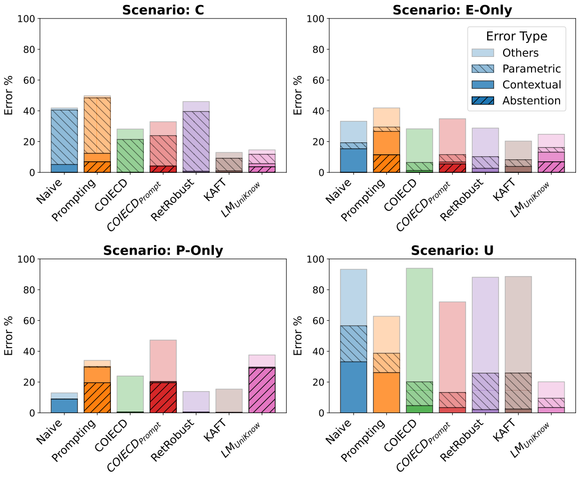

Figure 15: Stacked error type distributions across methods for each knowledge scenario. Transparency reflects error type. Evaluated using LLAMA2-7B.

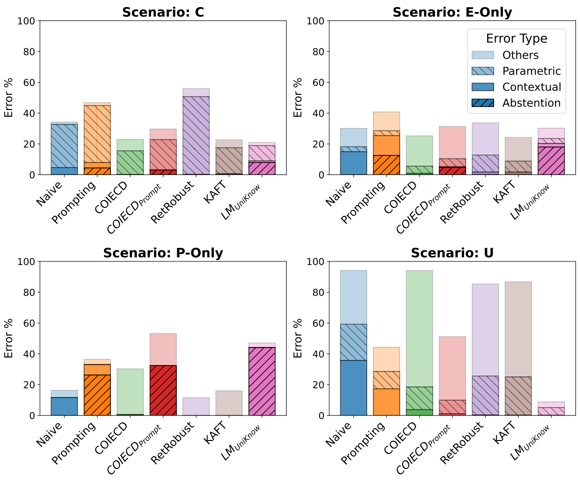

Figure 16: Stacked error type distributions across methods for each knowledge scenario. Transparency reflects error type. Evaluated using MISTRAL-7B.

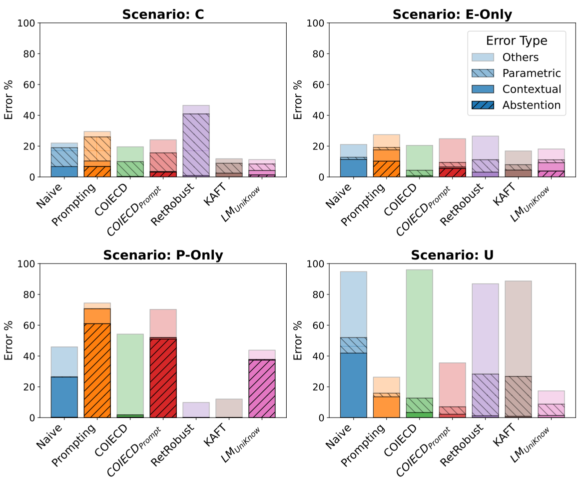

Figure 17: Stacked error type distributions across methods for each knowledge scenario. Transparency reflects error type. Evaluated using QWEN2.5-7B.

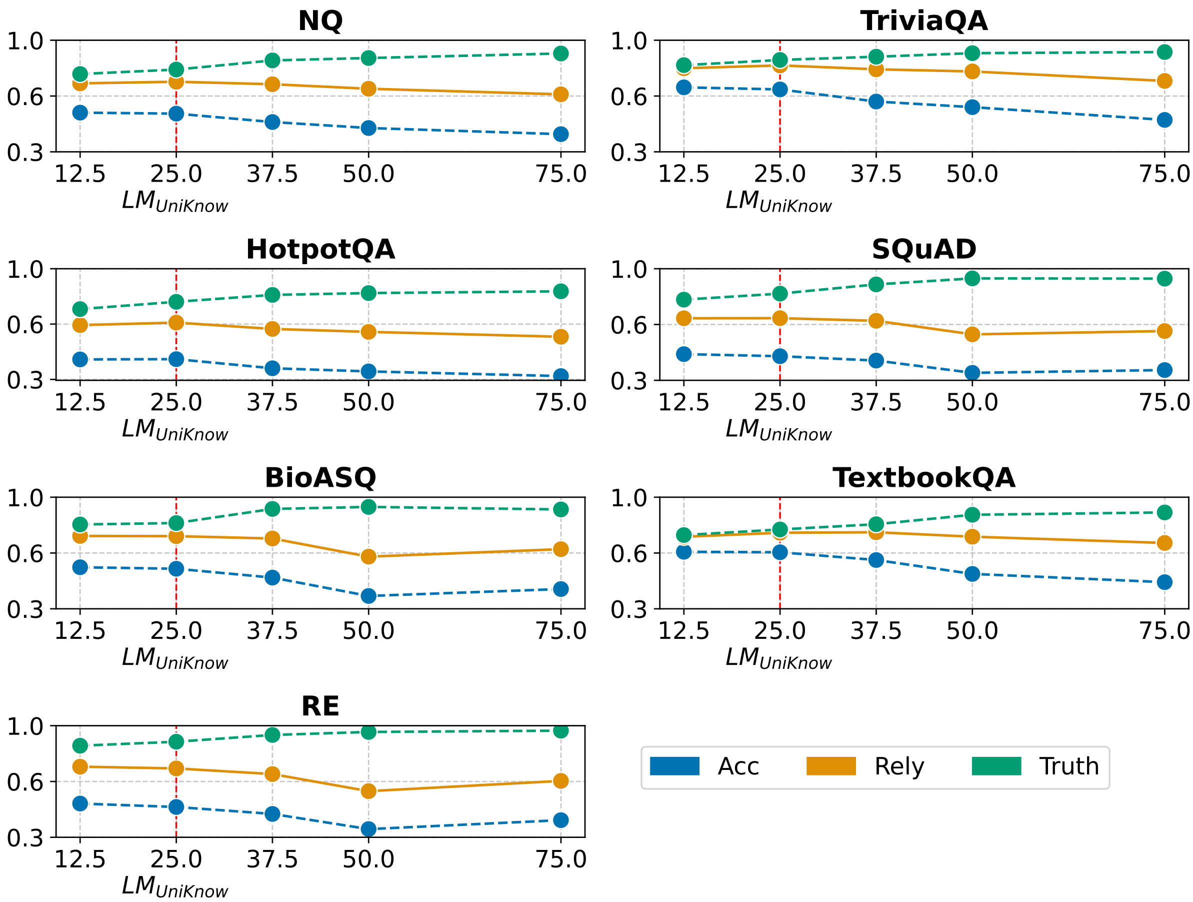

Figure 18: Effect of varying the proportion of abstention data on model performance for LLAMA3-8B for each dataset. The red dashed line indicates the proportion used in $\mathrm{LM}_{\mathrm{UniKnow}}$.

Dataset	HotpotQA			BioASQ			SQuAD			TextbookQA			RE		
Metric	Acc	Truth	Rely	Acc	Truth	Rely	Acc	Truth	Rely	Acc	Truth	Rely	Acc	Truth	Rely
$LM_{UniKnow}$	**0.4282**	**0.7908**	**0.6593**	0.5513	**0.8379**	**0.7557**	0.4513	**0.8438**	**0.6897**	**0.6546**	**0.7976**	**0.7771**	0.4901	**0.8994**	**0.7319**
−C	0.4506	0.4984	0.4961	**0.5821**	0.6449	0.6410	0.4970	0.5715	0.5659	0.6044	0.6416	0.6402	0.5591	0.7247	0.6973
−IR	0.4402	0.5030	0.4990	0.5760	0.6080	0.6069	**0.5129**	0.5565	0.5546	0.5978	0.6089	0.6088	0.5678	0.6331	0.6288
−C, IR	0.4579	0.4960	0.4946	0.5918	0.5940	0.5940	0.5063	0.5224	0.5221	0.6108	0.6158	0.6157	**0.5784**	0.6312	0.6284

Table 9: Ablation study on context types in the training data for LLAMA3-8B, measuring the impact of excluding conflicting contexts (−C), incorrectly retrieved contexts (−IR), or both (−C, IR). **Bold** indicates the best.

Model	Scenario (↓)	NQ	TriviaQA	HotpotQA	SQuAD	BioASQ	TextbookQA	RE
LLAMA2-7B	C	221	2,442	160	303	74	175	145
	P-Only	442	4,884	320	606	148	350	290
	E-Only	5,090	3,676	6,878	11,088	626	694	2,422
	U	5,090	3,676	6,878	11,088	626	694	2,422
LLAMA2-13B	C	361	3,050	299	431	74	191	207
	P-Only	722	6,100	598	862	148	382	414
	E-Only	4,556	2,812	6,514	10,480	604	632	2,306
	U	4,556	2,812	6,514	10,480	604	632	2,306
LLAMA3-8B	C	273	3,231	317	462	101	193	233
	P-Only	546	6,462	634	924	202	386	466
	E-Only	4,766	3,076	6,444	10,360	448	580	2,150
	U	4,766	3,076	6,444	10,360	448	580	2,150
MISTRAL-7B	C	326	3,282	302	473	116	220	197
	P-Only	652	6,564	604	946	232	440	394
	E-Only	4,756	3,196	6,530	10,656	494	628	2,462
	U	4,756	3,196	6,530	10,656	494	628	2,462
QWEN-1.5B	C	119	1,011	80	157	59	158	78
	P-Only	238	2,022	160	314	118	316	156
	E-Only	6,202	9,246	7,774	12,292	856	802	2,964
	U	6,202	9,246	7,774	12,292	856	802	2,964
QWEN-3B	C	188	1,472	167	270	92	184	118
	P-Only	376	2,944	334	540	184	368	236
	E-Only	5,624	7,266	7,254	11,584	580	626	2,722
	U	5,624	7,266	7,254	11,584	580	626	2,722
QWEN-7B	C	315	2,485	231	401	167	282	187
	P-Only	630	4,970	462	802	334	564	374
	E-Only	5,068	5,458	6,924	10,694	422	502	2,460
	U	5,068	5,458	6,924	10,694	422	502	2,460
QWEN-14B	C	334	3,284	363	633	202	303	233
	P-Only	668	6,568	726	1,266	404	606	466
	E-Only	4,692	3,808	6,328	9,630	316	502	2,254
	U	4,692	3,808	6,328	9,630	316	502	2,254

Table 10: Number of samples in each scenario.

Scenario	Method (↓)	LLAMA2-7B	LLAMA2-13B	LLAMA3-8B	MISTRAL-7B	QWEN-1.5B	QWEN-3B	QWEN-7B	QWEN-14B
All	Naïve	.5467	.5628	.5430	.5632	.5384	.5284	.5406	.5419
	Prompting	.5288	.5486	.5727	.5795	.5916	.5880	.6059	.6019
	COIECD	.5642	.5753	.5600	.5691	.5276	.5168	.5243	.5114
	$\text{COIECD}_{\text{Prompt}}$	.5321	.5572	.5881	.5870	.5516	.5819	.6130	.5976
	RetRobust	.5540 ± 0.01	.6213 ± 0.00	.5100 ± 0.01	.5445 ± 0.01	.5642 ± 0.01	.5650 ± 0.01	.5683 ± 0.01	.5608 ± 0.00
	KAFT	.6554 ± 0.00	.6851 ± 0.00	.6533 ± 0.01	.6287 ± 0.00	.6317 ± 0.00	.6437 ± 0.00	.6710 ± 0.00	.6815 ± 0.00
	$\text{LM}_{\text{UniKnow}}$	**.7562 ± 0.00**	**.7778 ± 0.00**	**.7668 ± 0.01**	**.7412 ± 0.01**	**.7098 ± 0.01**	**.7517 ± 0.00**	**.7776 ± 0.01**	**.7915 ± 0.00**
C	Naïve	.5817	.6538	.5911	.6585	.7280	.7538	.7799	.7400
	Prompting	.5026	.5373	.5064	.5324	.7234	.7314	.7051	.6610
	COIECD	.7185	.7691	.7254	.7711	.7754	.7775	.8043	.7487
	$\text{COIECD}_{\text{Prompt}}$	.6707	.7061	.6979	.7033	.7591	.7515	.7587	.7112
	RetRobust	.5264 ± 0.02	.6863 ± 0.02	.3794 ± 0.03	.4679 ± 0.02	.5979 ± 0.02	.5596 ± 0.02	.5177 ± 0.02	.5081 ± 0.01
	KAFT	**.8642 ± 0.01**	.9224 ± 0.01	.8563 ± 0.02	.7796 ± 0.01	.8070 ± 0.01	.7991 ± 0.02	.8715 ± 0.01	.9024 ± 0.00
	$\text{LM}_{\text{UniKnow}}$	.8512 ± 0.01	**.9336 ± 0.00**	**.8714 ± 0.01**	**.8111 ± 0.02**	**.8285 ± 0.01**	**.8191 ± 0.01**	**.8839 ± 0.01**	**.9299 ± 0.00**
P-Only	Naïve	**.8703**	.8474	.8518	.8371	.6031	.5379	.5407	.5768
	Prompting	.6591	.7056	.7295	.6360	.4098	.2348	.2557	.3000
	COIECD	.7606	.7388	.7379	.6977	.5391	.4656	.4578	.4842
	$\text{COIECD}_{\text{Prompt}}$	.5272	.5329	.6051	.4685	.4152	.2888	.2975	.3447
	RetRobust	.8637 ± 0.00	**.8989 ± 0.01**	**.9058 ± 0.00**	**.8782 ± 0.01**	**.8110 ± 0.01**	**.8619 ± 0.01**	**.8982 ± 0.00**	**.8827 ± 0.00**
	KAFT	.8477 ± 0.00	.8474 ± 0.00	.8721 ± 0.01	.8417 ± 0.00	.8102 ± 0.01	.8577 ± 0.01	.8765 ± 0.00	.8667 ± 0.00
	$\text{LM}_{\text{UniKnow}}$	.6187 ± 0.02	.5183 ± 0.02	.6660 ± 0.00	.5456 ± 0.01	.5319 ± 0.03	.7208 ± 0.01	.5659 ± 0.03	.5586 ± 0.02
E-Only	Naïve	.6677	.6855	.6623	.6987	.7704	.7795	.7893	.7694
	Prompting	.5813	.5536	.5937	.5923	.7480	.7203	.7255	.7184
	COIECD	.7171	.7309	.7077	.7478	.7594	.7841	.7952	.7580
	$\text{COIECD}_{\text{Prompt}}$	.6514	.6617	.6854	.6869	.7416	.7314	.7518	.7182
	RetRobust	.7069 ± 0.01	.7476 ± 0.01	.6172 ± 0.01	.6819 ± 0.02	.7594 ± 0.01	.7353 ± 0.01	.7341 ± 0.00	.7150 ± 0.01
	KAFT	**.7945 ± 0.00**	**.8232 ± 0.00**	.7593 ± 0.02	**.7606 ± 0.00**	**.8224 ± 0.00**	**.8184 ± 0.00**	**.8228 ± 0.01**	.8291 ± 0.00
	$\text{LM}_{\text{UniKnow}}$	.7526 ± 0.00	.8146 ± 0.01	**.7676 ± 0.02**	.7131 ± 0.02	.7921 ± 0.01	.8071 ± 0.01	.8186 ± 0.01	**.8441 ± 0.00**
U	Naïve	.0674	.0644	.0668	.0587	.0519	.0426	.0523	.0816
	Prompting	.3724	.3980	.4611	.5572	.4852	**.6654**	.7371	.7283
	COIECD	.0606	.0623	.0690	.0597	.0366	.0400	.0399	.0548
	$\text{COIECD}_{\text{Prompt}}$	.2790	.3282	.3641	.4891	.2904	.5560	.6442	.6164
	RetRobust	.1191 ± 0.00	.1523 ± 0.00	.1374 ± 0.00	.1499 ± 0.00	.0886 ± 0.00	.1031 ± 0.00	.1231 ± 0.01	.1375 ± 0.00
	KAFT	.1153 ± 0.00	.1475 ± 0.00	.1253 ± 0.00	.1328 ± 0.00	.0872 ± 0.00	.0997 ± 0.00	.1132 ± 0.00	.1277 ± 0.01
	$\text{LM}_{\text{UniKnow}}$	**.8022 ± 0.01**	**.8448 ± 0.00**	**.7620 ± 0.02**	**.8949 ± 0.01**	**.6865 ± 0.02**	.6597 ± 0.01	**.8421 ± 0.02**	**.8334 ± 0.01**

Table 11: EM score for each scenario, across models. **Bold** indicates the best, and the underline indicates the second best. Training-based methods (RetRobust, KAFT, and $\text{LM}_{\text{UniKnow}}$) are evaluated using three training seeds, and the mean and standard deviation are reported.

Method (↓)	NQ Acc	NQ Rely	TriviaQA Acc	TriviaQA Rely	HotpotQA Acc	HotpotQA Rely	SQuAD Acc	SQuAD Rely	BioASQ Acc	BioASQ Rely	TextbookQA Acc	TextbookQA Rely	RE Acc	RE Rely
LLAMA2-7B														
Naïve	.4177	.4177	.6194	.6194	.4342	.4342	.4856	.4859	.5402	.5402	.5604	.5604	.5313	.5313
Prompting	.3309	.5665	.5425	.6762	.3591	.4675	.3748	.5134	.3849	.5776	.4799	.6318	.5067	.5944
COIECD	.4328	.4328	.5982	.5983	.4355	.4356	.4818	.4822	.5147	.5147	.5284	.5284	.5234	.5236
COIECD$_{\text{Prompt}}$	.3845	.5620	.5421	.6463	.3643	.4316	.3906	.5215	.3630	.5487	.4633	.5795	.5172	.5540
RetRobust	.5561 ± .00	.5562 ± .00	.6712 ± .00	.6713 ± .00	.4251 ± .00	.4251 ± .00	.4700 ± .01	.4706 ± .01	.5275 ± .01	.5285 ± .01	.6219 ± .00	.6219 ± .00	.5590 ± .01	.5592 ± .01
KAFT	**.5979 ± .00**	.5981 ± .00	**.7347 ± .00**	.7347 ± .00	**.4493 ± .00**	.4494 ± .00	**.5162 ± .00**	.5169 ± .00	**.5909 ± .00**	.5918 ± .00	**.6863 ± .00**	.6863 ± .00	**.5889 ± .00**	.5889 ± .00
LM$_{\text{UniKnow}}$	.5099 ± .01	**.7207 ± .00**	.6143 ± .01	**.8130 ± .00**	.3808 ± .00	**.6212 ± .00**	.4407 ± .00	**.6844 ± .00**	.5159 ± .02	**.7309 ± .01**	.5519 ± .01	**.7589 ± .00**	.4736 ± .00	**.7194 ± .00**
LLAMA2-13B														
Naïve	.4474	.4475	.6556	.6556	.4503	.4503	.5062	.5064	.5674	.5674	.5691	.5691	.5412	.5415
Prompting	.3678	.5357	.5649	.7067	.3993	.4528	.4148	.6083	.2991	.5487	.5208	.6164	.4933	.6471
COIECD	.4594	.4594	.6361	.6362	.4509	.4510	.4959	.4961	.5739	.5739	.5533	.5535	.5222	.5223
COIECD$_{\text{Prompt}}$	.4322	.5736	.6023	.6539	.3995	.4654	.4378	.6172	.3311	.5752	.4979	.5793	.4829	.6078
RetRobust	.6178 ± .01	.6181 ± .01	.7468 ± .00	.7469 ± .00	.4728 ± .00	.4729 ± .00	.5172 ± .01	.5179 ± .01	.5870 ± .01	.5875 ± .01	.6783 ± .01	.6784 ± .01	.5965 ± .01	.5967 ± .01
KAFT	**.6443 ± .00**	.6444 ± .00	**.7901 ± .00**	.7901 ± .00	**.4923 ± .00**	.4924 ± .00	**.5489 ± .00**	.5497 ± .00	**.6284 ± .01**	.6292 ± .01	**.7273 ± .00**	.7274 ± .00	**.6012 ± .00**	.6015 ± .00
LM$_{\text{UniKnow}}$	.5434 ± .00	**.7547 ± .00**	.6432 ± .01	**.8372 ± .00**	.4166 ± .01	**.6595 ± .01**	.4746 ± .00	**.7175 ± .00**	.4688 ± .01	**.7104 ± .00**	.5843 ± .01	**.7975 ± .00**	.5108 ± .00	**.7525 ± .00**
LLAMA3-8B														
Naïve	.4443	.4444	.6218	.6218	.4529	.4529	.4943	.4944	.5656	.5656	.5627	.5627	.5447	.5447
Prompting	.4200	.6347	.5312	.7100	.3590	.4936	.4209	.6063	.4914	.6454	.4934	.6173	.4994	.6613
COIECD	.4724	.4726	.5984	.5984	.4534	.4537	.4893	.4896	.5857	.5864	.5301	.5301	.5230	.5230
COIECD$_{\text{Prompt}}$	.4407	.6316	.5855	.7087	.3860	.4493	.4565	.6181	.5294	.6143	.4882	.6047	.5061	.6138
RetRobust	.5532 ± .00	.5532 ± .00	.6572 ± .00	.6572 ± .00	.4101 ± .00	.4102 ± .00	.4496 ± .01	.4501 ± .01	.5446 ± .00	.5452 ± .00	.6021 ± .01	.6022 ± .01	.5488 ± .00	.5488 ± .00
KAFT	**.6140 ± .01**	.6141 ± .01	**.7637 ± .00**	.7638 ± .00	**.4718 ± .00**	.4719 ± .00	**.5120 ± .01**	.5125 ± .01	**.6466 ± .02**	.6471 ± .02	**.7092 ± .01**	.7092 ± .01	**.5867 ± .01**	.5867 ± .01
LM$_{\text{UniKnow}}$	.5434 ± .00	**.7394 ± .00**	.6855 ± .01	**.8415 ± .00**	.4224 ± .01	**.6541 ± .00**	.4487 ± .01	**.6888 ± .01**	.5302 ± .03	**.7500 ± .01**	.6562 ± .00	**.7811 ± .01**	.4913 ± .01	**.7322 ± .00**
MISTRAL-7B														
Naïve	.4444	.4444	.6270	.6270	.4586	.4586	**.5109**	.5111	.5911	.5911	.5658	.5658	.5386	.5386
Prompting	.3304	.5615	.6149	.7028	.3459	.5471	.3806	.6145	.4634	.6677	.4761	.6244	.4695	.6786
COIECD	.4601	.4603	.5917	.5919	.4575	.4575	.5011	.5015	.5653	.5653	.5457	.5457	.5351	.5352
COIECD$_{\text{Prompt}}$	.4039	.6053	.6179	.6750	.3525	.5535	.4327	**.6389**	.4516	.6269	.4967	.6267	.4705	.6681
RetRobust	.5898 ± .01	.5902 ± .01	.6571 ± .01	.6573 ± .01	.4256 ± .00	.4257 ± .00	.4587 ± .02	.4594 ± .02	.5727 ± .01	.5732 ± .01	.6297 ± .02	.6298 ± .02	.5633 ± .02	.5636 ± .02
KAFT	**.6066 ± .00**	.6068 ± .00	**.7279 ± .01**	.7281 ± .01	**.4631 ± .00**	.4633 ± .00	.4983 ± .01	.4990 ± .01	**.5959 ± .02**	.5962 ± .02	**.7088 ± .01**	.7088 ± .01	**.5736 ± .01**	.5738 ± .01
LM$_{\text{UniKnow}}$	.5142 ± .00	**.7443 ± .00**	.6267 ± .01	**.8303 ± .00**	.3881 ± .00	**.6379 ± .00**	.3661 ± .03	.6116 ± .04	.4682 ± .01	**.7128 ± .01**	.5713 ± .02	**.7801 ± .00**	.4314 ± .03	**.6800 ± .03**
QWEN-1.5B														
Naïve	.4300	.4306	.5005	.5014	.4056	.4057	.4615	.4622	.4727	.4738	.5192	.5194	.5023	.5037
Prompting	.4067	.5683	.4780	.6333	.3723	.5370	.4462	.5830	.4225	.6465	.4427	.6174	.4547	.6714
COIECD	.4152	.4161	.5009	.5035	.3560	.3566	.4539	.4560	.4476	.4494	.4870	.4877	.4938	.4978
COIECD$_{\text{Prompt}}$	.3769	.4919	.4845	.5681	.3383	.4280	.4297	.5218	.4362	.5668	.4657	.5653	.4743	.6341
RetRobust	.4717 ± .01	.4719 ± .01	.5470 ± .01	.5471 ± .01	.4087 ± .01	.4088 ± .01	.4546 ± .01	.4552 ± .01	.5243 ± .01	.5253 ± .01	.6066 ± .01	.6068 ± .01	.5289 ± .00	.5289 ± .00
KAFT	**.4919 ± .00**	.4922 ± .00	**.5788 ± .00**	.5789 ± .00	**.4311 ± .00**	.4313 ± .00	**.4872 ± .00**	.4879 ± .00	**.5772 ± .01**	.5772 ± .01	**.6499 ± .00**	**.6503 ± .00**	**.5477 ± .00**	.5478 ± .00
LM$_{\text{UniKnow}}$	.4441 ± .01	**.6443 ± .00**	.5304 ± .00	**.7371 ± .00**	.3828 ± .00	**.6226 ± .00**	.4364 ± .01	**.6521 ± .00**	.4805 ± .01	**.7006 ± .01**	.5683 ± .00	.6386 ± .01	.4748 ± .01	**.7147 ± .01**
QWEN-3B														
Naïve	.4388	.4393	.5137	.5145	.4192	.4194	.4680	.4688	.4993	.4996	.5116	.5116	.5061	.5086
Prompting	.3878	.6106	.4299	.6608	.3497	.5903	.4279	.6460	.4275	.6577	.4025	.6241	.4512	.6825
COIECD	.4263	.4278	.5130	.5174	.4066	.4075	.4617	.4636	.4803	.4831	.4858	.4889	.5048	.5125
COIECD$_{\text{Prompt}}$	.3782	.5787	.4681	.6744	.3566	.5739	.4309	.6247	.4336	.6278	.4325	.5966	.4639	.6748
RetRobust	.5106 ± .00	.5107 ± .00	.5810 ± .00	.5810 ± .00	.4267 ± .00	.4267 ± .00	.4720 ± .01	.4729 ± .01	.5970 ± .01	.5976 ± .01	.6495 ± .00	.6495 ± .00	.5464 ± .01	.5466 ± .01
KAFT	**.5364 ± .00**	.5366 ± .00	**.6353 ± .00**	.6354 ± .00	**.4594 ± .00**	.4596 ± .00	**.5112 ± .00**	.5121 ± .00	**.6772 ± .01**	.6789 ± .01	**.7034 ± .00**	.7035 ± .00	**.5599 ± .00**	.5599 ± .00
LM$_{\text{UniKnow}}$	.5053 ± .01	**.6896 ± .00**	.6139 ± .00	**.7742 ± .00**	.4341 ± .01	**.6583 ± .00**	.4801 ± .00	**.6844 ± .00**	.5838 ± .01	**.7689 ± .00**	.6517 ± .01	**.7350 ± .01**	.4992 ± .01	**.7301 ± .00**
QWEN-7B														
Naïve	.4523	.4529	.5870	.5900	.4413	.4450	.4905	.4929	.5735	.5742	.5573	.5578	.5132	.5147
Prompting	.3788	.6230	.4672	.6973	.3759	.6191	.4493	.6798	.4487	.6860	.4422	.6659	.4639	.7010
COIECD	.4410	.4413	.5785	.5847	.4183	.4190	.4772	.4807	.5215	.5222	.5329	.5331	.5057	.5106
COIECD$_{\text{Prompt}}$	.4096	.6386	.4975	.7052	.3612	.5871	.4469	.6687	.4864	.6856	.4517	.6620	.4758	.6984
RetRobust	.5609 ± .00	.5612 ± .00	.6165 ± .00	.6166 ± .00	.4323 ± .00	.4324 ± .00	.5081 ± .01	.5088 ± .01	.6044 ± .02	.6046 ± .02	.6567 ± .00	.6568 ± .00	.5927 ± .01	.5927 ± .01
KAFT	**.5895 ± .00**	.5897 ± .00	**.6909 ± .00**	.6909 ± .00	**.4736 ± .00**	.4737 ± .00	**.5321 ± .00**	.5328 ± .00	**.6780 ± .01**	.6780 ± .01	**.7425 ± .01**	.7425 ± .01	**.6040 ± .00**	.6040 ± .00
LM$_{\text{UniKnow}}$	.5112 ± .01	**.7325 ± .00**	.5911 ± .01	**.8053 ± .01**	.4082 ± .01	**.6564 ± .01**	.4768 ± .01	**.7183 ± .00**	.5605 ± .00	**.7806 ± .00**	.6429 ± .00	**.7897 ± .00**	.5129 ± .01	**.7580 ± .01**
QWEN-14B														
Naïve	.4743	.4746	.6522	.6523	.4614	.4616	.5051	.5063	.6385	.6385	.5661	.5663	.5268	.5285
Prompting	.4141	.6470	.5146	.7240	.4036	.6293	.4510	.6888	.4867	.7220	.4709	.6681	.4630	.7035
COIECD	.4421	.4427	.6220	.6228	.4294	.4302	.4740	.4766	.5768	.5789	.5379	.5489	.5016	.5037
COIECD$_{\text{Prompt}}$	.4045	.6170	.5550	.7167	.3852	.5826	.4467	.6645	.4900	.6992	.4922	.6559	.4704	.6931
RetRobust	.6139 ± .00	.6143 ± .00	.6729 ± .00	.6730 ± .00	.4737 ± .00	.4738 ± .00	.5144 ± .00	.5155 ± .00	.5720 ± .01	.5722 ± .01	.6788 ± .00	.6789 ± .00	.5800 ± .00	.5800 ± .00
KAFT	**.6423 ± .00**	.6424 ± .00	**.7687 ± .00**	.7687 ± .00	**.5236 ± .00**	.5236 ± .00	**.5620 ± .00**	.5627 ± .00	**.6963 ± .01**	.6963 ± .01	**.7644 ± .00**	.7644 ± .00	**.6047 ± .00**	.6048 ± .00
LM$_{\text{UniKnow}}$	.5493 ± .01	**.7598 ± .00**	.6573 ± .00	**.8469 ± .00**	.4559 ± .00	**.6943 ± .00**	.4943 ± .00	**.7325 ± .00**	.5457 ± .02	**.7736 ± .01**	.6428 ± .01	**.8317 ± .01**	.5351 ± .00	**.7694 ± .00**

Table 12: Acc and Rely for each method and model across datasets. **Bold** indicates the best, and the underline indicates the second best. Training-based methods (RetRobust, KAFT, and LM$_{\text{UniKnow}}$) are evaluated using three training seeds, and the mean and standard deviation are reported.