KnowPO: Knowledge-Aware Preference Optimization for Controllable Knowledge Selection in Retrieval-Augmented Language Models

Ruizhe Zhang[1,2*], **Yongxin Xu**[1,2*], **Yuzhen Xiao**[1,2*], **Runchuan Zhu**[1,2], **Xinke Jiang**[1,2], **Xu Chu**[1,2,4,5], **Junfeng Zhao**[1,2,6†], **Yasha Wang**[2,3,5†]

[1] School of Computer Science, Peking University, Beijing, China
[2] Key Laboratory of High Confidence Software Technologies, Ministry of Education, Beijing, China
[3] National Engineering Research Center For Software Engineering, Peking University, Beijing, China
[4] Center on Frontiers of Computing Studies, Peking University, Beijing, China
[5] Peking University Information Technology Institute (Tianjin Binhai)
[6] Nanhu Laboratory, Jiaxing, China
{nostradamus,xuyx,xiaoyuzhen}@stu.pku.edu.cn; zhaojf@pku.edu.cn; wangyasha@pku.edu.cn

Abstract

By integrating external knowledge, Retrieval-Augmented Generation (RAG) has become an effective strategy for mitigating the hallucination problems that large language models (LLMs) encounter when dealing with knowledge-intensive tasks. However, in the process of integrating external non-parametric supporting evidence with internal parametric knowledge, inevitable knowledge conflicts may arise, leading to confusion in the model's responses. To enhance the knowledge selection of LLMs in various contexts, some research has focused on refining their behavior patterns through instruction-tuning. Nonetheless, due to the absence of explicit negative signals and comparative objectives, models fine-tuned in this manner may still exhibit undesirable behaviors such as contextual ignorance and contextual overinclusion. To this end, we propose a **Know**ledge-aware **P**reference **O**ptimization strategy, dubbed KnowPO, aimed at achieving adaptive knowledge selection based on contextual relevance in real retrieval scenarios. Concretely, we proposed a general paradigm for constructing knowledge conflict datasets, which comprehensively cover various error types and learn how to avoid these negative signals through preference optimization methods. Simultaneously, we proposed a rewriting strategy and data ratio optimization strategy to address preference imbalances. Experimental results show that KnowPO outperforms previous methods for handling knowledge conflicts by over 37%, while also exhibiting robust generalization across various out-of-distribution datasets.

Introduction

Large Language Models (LLMs) (Taylor et al. 2022; Zhao et al. 2023b) have been widely applied in various fields, such as natural language processing, question-answering systems, and text generation, giving rise to numerous AI applications (Kaplan et al. 2020; Vu et al. 2024; Li et al. 2023, 2024). These models exhibit outstanding performance in many tasks, primarily due to their large-scale parameters and extensive pre-training data (Ziegler et al. 2020; Wang et al. 2023b; Ma et al. 2024; Lin et al. 2024). However, because of the static nature of the training data, LLMs may generate seemingly coherent but actually unreliable information, a phenomenon known as "hallucination" (Ji et al. 2023a,b; Cao et al. 2020), due to outdated knowledge and long-tail knowledge (He, Zhang, and Roth 2022; Kandpal et al. 2023; Jiang et al. 2024). Retrieval-Augmented Generation (RAG) paradigm (Izacard et al. 2022; Asai et al. 2023b,a), within a retrieve-and-read framework, leverages information from reliable knowledge bases to compensate the static nature of the Internal knowledge of LLM. However, the performance of RAG framework is limited by the **knowledge conflicts** between internal knowledge stored in LLM parameters and external database (Xu et al. 2024; Jin et al. 2024). In this paper, we focus on resolving knowledge conflict by adhere to the retrieved knowledge, and meanwhile, improving the robustness against noise in the retrieved context.

In response to the aforementioned issue, a mainstream approach is to construct specific instruction-tuning datasets to optimize the knowledge prioritization of LLMs in contexts with varying degrees of relevance (Li et al. 2022; Xue et al. 2023). However, as shown in Figure 1, achieving a balance between **adherence capability** and **noise robustness** is highly challenging. On one hand, when the LLM heavily relies on external knowledge, it risks over-focusing on irrelevant retrieval contexts, struggling to effectively discern noise. On the other hand, an excessive emphasis on enhancing the LLM's noise resistance can inadvertently filter out useful contextual information (Wu, Wu, and Zou 2024). Moreover, the manifestation of these capabilities is closely related to the complexity of the context in real-world RAG scenarios (Longpre et al. 2022; Xie et al. 2024). Therefore, it is crucial to address the balance between adherence capability and noise robustness in real RAG scenarios.

Therefore, our insights stem from the error types observed in real-world scenarios involving RAG. We observe that existing research fails to distinguish between supervisory signals, leading to adherence capability and noise robustness being treated as analogous instruction-following pos-

*These authors contributed equally.
†Corresponding Author.

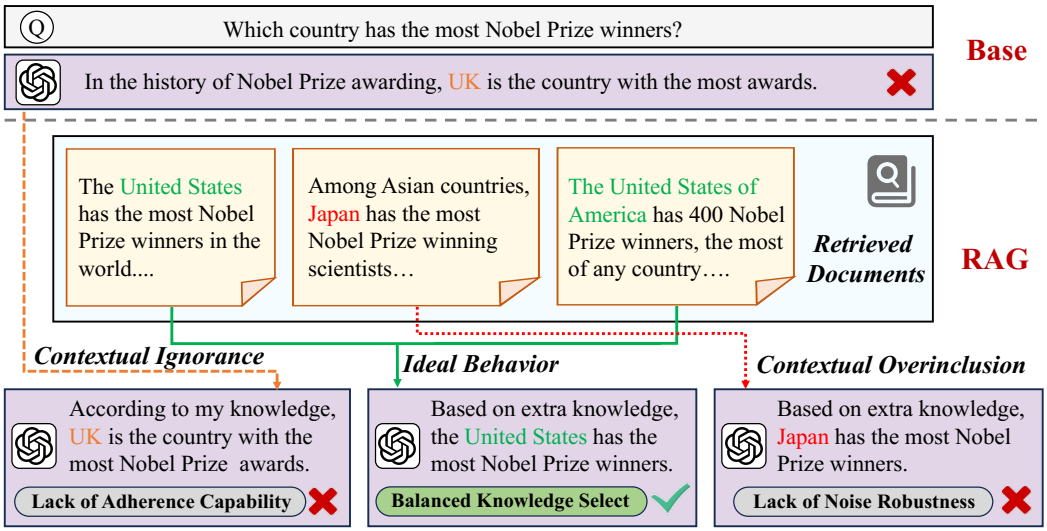

Figure 1: An illustrative example of how does LLM behave when encountering knowledge conflicts in RAG scenarios.

itive examples. This leads to contradictory signals during instruction-following training, causing learning variance and impeding the effective acquisition of both capabilities. To address this, we propose a more nuanced optimization approach that introduces preference data specifically describing adherence capability and noise robustness. Leveraging efficacious and extensively utilized Direct Preference Optimization (DPO) (Rafailov et al. 2024), we optimize the model's ability to leverage external knowledge, thereby enhancing the overall efficacy of RAG.

Although seemingly straightforward, implementing this intuition faces these challenges: **(C1)** How to more accurately simulate complicated context in real-world RAG scenarios and introduce more comprehensive, fine-grained negative signals? **(C2)** How to resolve data discrepancies in preference learning to avoid behavior pattern imbalances?

By jointly considering the above issues, we propose KnowPO, a **Know**ledge-aware **P**reference **O**ptimization strategy, which constructs comprehensive and balanced preference relations to optimize LLMs' knowledge selection in different contexts. **i)** Specifically, we simulated real-world RAG scenarios at the input level. We perform refined noise classification based on the relevance between the knowledge context and the question topic, and explore combination methods with evidence to form conflicting context and irrelevant context. At the output level, we simulate two common error types in different context relevance scenarios: **Contextual Ignorance** and **Contextual Overinclusion**, and develop training strategies to avoid these errors. **ii)** Secondly, we propose a rewriting strategy to address length imbalance and a data ratio balancing strategy to address behavior pattern imbalance, using DPO to optimize LLMs' adherence capability and noise robustness. These strategies not only eliminate length biases and imbalances in behavior pattern distribution but also enhance the exhaustiveness of the model's responses. This prevents degradation of conversational abilities that can occur when training on datasets with shorter answers, such as in reading comprehension tasks.

Our main contributions are summarized as follows:

- We observed that LLMs in RAG scenarios fail to effectively balance adherence capability and noise robustness. and thus proposed the KnowPO framework, which refines negative supervisory signals to enhance LLM behavior when encountering knowledge conflicts.
- We proposed a general paradigm for constructing knowledge conflict datasets, comprehensively covering various error types and generalizable to different model architectures. We also proposed a rewriting strategy and data ratio optimization strategy to address preference imbalances.
- We validated our method's training effectiveness on multiple models and datasets and tested its generalization ability in out-of-distribution (OOD) scenarios. The results indicate that our method not only improves the performance of models on test sets but also enhances their adaptability and robustness when confronted with unknown data.

Related Work

Knowledge Conflicts. Numerous studies have explored LLMs' behavior in knowledge conflict scenarios, providing valuable insights for our work. Longpre et al. (2022) discovered that large Pre-trained Language Models often prefer parametric knowledge over contextual information when facing knowledge conflicts. Wu, Wu, and Zou (2024) highlighted that this tendency to disregard context is influenced by the model's prior token probability, with high-probability parametric knowledge being harder to override. Kassner and Schütze (2019) demonstrated that LLMs are susceptible to being misled by task-irrelevant context. Furthermore, Tan et al. (2024) indicated that the model's contextual preferences are linked to the semantic completeness of the context and its relevance to the question.

Several studies aim to improve the adherence of LLMs to context amid knowledge conflicts. For instance, Knowl-

edge Aware Fine-Tuning (KAFT) (Li et al. 2022) enhances models' ability to use external knowledge by creating challenging counterfactual knowledge from training datasets and incorporating irrelevant knowledge to boost noise resistance. However, as previously mentioned, the applicability of this approach in real-world RAG scenarios is limited. Additionally, decoding-based methods (Jin et al. 2024; Chen, Zhang, and Choi 2022), like Context-Aware Decoding (CAD) (Shi et al. 2023b), adjust LLMs' output probabilities during token generation, akin to contrastive decoding, conditioned on relevant context. However, this approach may impact the semantic coherence of long responses. Moreover, prompt-based methods employ sophisticated designed prompts to ensure that LLMs adhere to the provided context (Si et al. 2023; Zhou et al. 2023). However, research shows that merely modifying prompts doesn't significantly alter LLMs' internal prior token probabilities (Wu, Wu, and Zou 2024), potentially limiting the effectiveness of this approach.

Methodology

Task Definition

Given an LLM Θ and an input natural language question q, we ask Θ to generate a response $\alpha = \Theta(q)$, representing the parametric knowledge for q. Assume in a typical *retrieve-and-read* framework, context τ is a permutation of $D_j^r, j = 1, 2, \ldots, K$, which represents a set of documents retrieved based on q. And $S = \{a_i\}, i = 1, 2, \ldots, N$ constitutes the set of contextual answer, each of which is derived from a retrieved document $D_{\tau_i}^r$. We can simplify the RAG task into $y = \Theta(q \| \tau)$, where y is output of Θ based on context τ. Note that K is not necessarily equals with N, because some retrieved documents may not contain any answer for q and are known as noises.

It's clearly that α and a_i are independent. **Knowledge conflict** appears when $\alpha \notin S$, and at this time response y of $\Theta(q \| \tau)$ can be uncertain. To simplify the discussion, we limit N to a maximum of 1, which means context τ contains at most one document D_ϵ^r from which the answer can be derived. Our purpose is to make sure $y = a_\epsilon$ when $|S| = 1$ and $y = \alpha$ when $|S| = 0$. In other word, LLM Θ should use appropriate external knowledge when there exists a document which contains the necessary knowledge regardless of conflicting with parameter knowledge, while use its parameter knowledge when retrieved documents are all irrelevant.

Contradictory Knowledge

Constructing knowledge that conflicts with LLM's parameter knowledge is crucial to condition $|S| = 1$. For question q in RAG scenarios, this conflict is reflected in conflicting answers a_{cf} which are inconsistent with LLM's parameter answer α. It is important to note that these conflicting answers a_{cf} do not necessarily have to be correct, nor is the LLM's parameter answer α always incorrect. In our approach, both answers can be incorrect to the question as long as they conflict with each other. The key to knowledge conflict lies in the conflict itself, regardless of correctness. This addresses a common misconception in previous work, where researchers often ensured that one answer was correct and the other incorrect (Tan et al. 2024; Wu, Wu, and Zou 2024), which not only increased the difficulty of data filtering but also overlooked some knowledge conflict scenarios.

Specifically, we first extract world knowledge acquired during the pretraining phase of the large model, marked as parameter answer α. We encourage LLM to abstain from answering when uncertain. Additionally, we refine the response formats for other parametric knowledge. The revised results are presented in Table 1.

For a given question q and LLM's parameter answer α, there are two potential sources of conflicting answers a_{cf}. The first is the realistic answer a_{real} to the question. The second is a fabricated answer a_{ctf} generated using GPT-4 that deviates from the realistic answer a_{real}. The latter is often referred to as a counterfactual answer, which we require to be as plausible as possible. Thus, for a question q and LLM's parameter answer α, we can obtain at least one conflicting answer, ensuring it is not overly far-fetched.

Context Formulation

In this section, we illustrated how to formulate context τ based on different kinds of knowledge conflict.

To align with the RAG scenario, we utilized the SQuAD2.0 dataset (Rajpurkar, Jia, and Liang 2018), a reading comprehension dataset encompassing multiple general domains, with a substantial corpus of documents and associated QA tasks. Notably, besides corpora collected from Wikipedia, SQuAD2.0 is also annotated by humans to determine whether a document can yield an answer for a specific question. Previous research has highlighted that treating a relevant yet non-informative document as a reference external knowledge source can impair LLM's adherence capabilities (Li et al. 2022). Following the chunk-size commonly used in RAG tasks (Shi et al. 2023a), we set the length of context τ to $K = 4$.

For scenarios with $|S| = 1$, we initially select pertinent documents from SQuAD2.0 based on the conflicting knowledge: For question q and realistic answer a_{real}, we directly select the corresponding document D_ϵ^r from the original dataset; and for question q and counterfactual answer a_{ctf}, we replace all occurrences of a_{real} with a_{ctf} in D_ϵ^r. Subsequently, we select one relevant document on the same topic and two documents on different topics based on semantic similarity. We ensure that these three documents are incapable of answering the question q. These four documents are then shuffled to constitute the conflicting context τ_{cf}.

For scenarios with $|S| = 0$, we distinguish between hard and easy irrelevant documents. Hard documents, derived from human annotations, consist of two documents that are on related topics but cannot answer the question. Easy documents are randomly selected, consisting of two documents on unrelated topics. These four documents are then shuffled to constitute the irrelevant context τ_{ir}.

Error Type Analyse

As previously mentioned, we expect LLMs to utilize contextual knowledge when encountering conflicting context,

Sample Question: Who is the Democratic presidential candidate in the 2024 US presidential election? *Parameter Answer*: **The Democratic candidate is Joe Biden.**			
		Conflicting Context	**Irrelevant Context**
Context Example		With President Joe Biden dropping out of the race on July 21, Vice President Kamala Harris became the presumed Democratic nominee. The search for her vice presidential running mate is closely watched, with top contenders including Secretary of Transportation Pete Buttigieg, Arizona Senator Mark Kelly, Illinois Governor J.B. Pritzker...	...a decision not without precedent in political history. Back in 1968, President Lyndon B. Johnson of the Democratic Party also opted out amidst intense political challenges. Following Johnson's withdrawal, Hubert Humphrey took over as the Democratic nominee for president.
Gold Output	Ideal Answer	The Democratic candidate is Kamala Harris.	**The Democratic candidate is Joe Biden.**
	Revised Result	Based on supplemental knowledge and my own understanding, the answer to this question is that the Democratic candidate is Kamala Harris.	Supplemental knowledge does not answer this question, but based on my knowledge, the answer to this question is that **the Democratic candidate is Joe Biden.**
Contextual Overinclusion	Error Answer	The Democratic candidate is Mark Kelly.	The Democratic candidate is Hubert Humphrey.
	Revised Result	Based on supplemental knowledge and my own understanding, the answer to this question is that The Democratic candidate is Mark Kelly.	Based on supplemental knowledge and my own understanding, the answer to this question is that The Democratic candidate is Hubert Humphrey.
Contextual Ignorance	Error Answer	**The Democratic candidate is Joe Biden.**	/
	Revised Result	Supplemental knowledge does not answer this question, but based on my knowledge, the answer to this question is that **the Democratic candidate is Joe Biden.**	/

Table 1: An example of how the KnowPO dataset is formulated. LLM's parameter knowledge are highlighted in **bold**, while conflicting knowledge in context is shown with light gray, and noisy information is presented in dark gray.

while relying on parameter knowledge when faced with irrelevant context. These two modes of handling context reflect the model's adherence capability and noise robustness, respectively. In practical RAG scenarios, deficiencies in these capabilities manifest as two distinct error types: one in which the LLM incorrectly uses irrelevant contextual information to construct answers, termed **Contextual Overinclusion**; and another where the LLM disregards the context entirely and relies exclusively on its parameter knowledge, termed **Contextual Ignorance**. These errors can occur with both types of contexts as illustrated in Table 1. To address these issues, we have meticulously designed a dataset comprising positive and negative sample pairs to specifically target and mitigate these errors.

Contextual Overinclusion Error. In situations with conflicting contexts, the ideal behavior of the LLM demonstrating adherence capability, as shown by positive samples in Table 1, is to answer using the conflicting knowledge present in the context. However, when contextual overinclusion occurs, LLM often utilizes inappropriate information from the context due to insufficient noise robustness and contextual understanding capability. For instance, in the example presented in Table 1, LLM chooses noisy information marked in red. To address this error, we constructed negative samples by using a prompt mechanism to guide GPT-4 to generate incorrect answers from conflicting contexts. To ensure the quality of the generated data, we adhered to stringent validation criteria: (1) The generated answers must be derived from the context, ensuring that the error is unequivocally attributable to contextual overinclusion; (2) The generated answers should be as plausible as possible and distinctly different from the conflicting answers, thereby ensuring the high quality of the data.

In situations with irrelevant contexts, it is evident that positive sample for noise robustness is to use LLM's parametric knowledge to respond. When this error occurs, LLM may fail to recognize the context as irrelevant, leading it to use contextual information instead of disregarding it. Similar to contextual overinclusion in conflicting contexts, we constructed corresponding negative samples by using GPT-4 to extract incorrect answers from irrelevant contexts.

Prompt: Generate Contextual Overinclusion

Please select a word from the provided context as an alternative answer to this question.
Question: {*Question* q}
Potential answer: {*Conflicting Answer* a_{cf}}
Context: {*Context* τ}

Please follow these requirements:
1. The answer must not be the same as the potential answer.
2. The alternative answer does not need to be correct, but it must appear in the context.
3. The alternative answer must be in a form that can answer the question and should be as reasonable as possible.

Contextual Ignorance Error. Contextual Ignorance occurs when the LLM disregards the context in its response, a behavior deemed erroneous solely in conflicting contexts. During such episodes, LLM may either fail to recognize the utility of the context or, even upon recognizing it, may opt

to disregard the conflicting answer in favor of relying on its parameter knowledge. For instance, in the example shown in Table 1, LLM answers the question without utilizing supplemental knowledge. To simulate this error, we constructed negative samples by extracting LLM's response to the query in the absence of any contextual support, ensuring that the answer aligns with an inappropriate erroneous response.

Training Method

Our training consists of two phases. First, we perform instruction tuning using the conflicting knowledge and contexts to enhance the LLM's adherence capability and noise robustness in RAG task scenarios. Next, we utilize the preference dataset for DPO training to further improve the LLM's ability to avoid the two types of errors, while ensuring that its final responses align with user preferences.

Instruction Tuning. Instruction tuning is a multi-task learning framework that enables the use of human-readable instructions to guide the output of LLMs. Given a source text and task-specific instructions, the model is trained to generate a sequence of tokens representing the desired output structure and its corresponding labels. Reviewing our definition of adherence capability and noise robustness, we would like to get a finetuned model Θ_{ft} from original LLM Θ that satisfies the following criteria:

$$|S| = 1 : \Theta_{ft}(q \| \tau_{cf}) = a_{cf}, \quad \text{where } \exists D^r_\epsilon \in \tau_{cf}, D^r_\epsilon \to a_{cf}$$
$$|S| = 0 : \Theta_{ft}(q \| \tau_{ir}) = \alpha, \quad \text{where } \Theta(q) = \alpha$$

Note that although the presence of the answer in τ was distinguished during dataset construction, the LLM does not possess this prior knowledge. The model must independently determine the context type and formulate a response during the RAG task.

Direct Preference Optimization. As previously discussed, LLMs may exhibit errors contextual overinclusion and contextual ignorance in real-world RAG scenarios. To further enhance adherence capability and noise robustness, we propose a Knowledge-aware Preference Optimization(KnowPO) training strategy. This strategy employs three types of preferences between positive and negative samples in two different contextual settings to conduct DPO training on the LLM. Using this approach, we train the LLM to avoid these errors and improve its ability to utilize different contexts.

During preparing data for DPO, we also identified two preference imbalances that impact training effectiveness.

- ***Length Imbalance.*** Some studies suggest that reward hacking observed in RLHF can also negatively impact DPO training (Gao, Schulman, and Hilton 2022; Park et al. 2024). We observed that in our previously constructed dataset, for the same preference pair, the positive sample was often the better-formatted and longer response, while the negative sample was a shorter conflicting answer. Due to the tendency of LLMs to be influenced by length bias during DPO (Singhal et al. 2024), they might prefer generating longer responses, which overall manifests as a greater tendency to refuse answering rather than providing a conflicting answer. To mitigate this issue, we standardized the format for all positive and negative samples in Table 1, aligning their lengths to ensure that the average length len_{win} approximately equals len_{loss}.
- ***Error Type Imbalance.*** Given that the preference pairs related to error contextual ignorance in conflicting context guide the LLM to "utilize contextual knowledge without rejecting it", while the preference pairs associated with error contextual overinclusion in irrelevant context exhibit a tendency towards "rejecting the use of contextual knowledge", we realized that the ratio of these two contrasting preference pairs could significantly influence training efficacy. During KnowPO training, we ensured that the proportion $\mathcal{R}_{error}$ of these two types of data was maintained at approximately 1:1. Furthermore, we validated the importance of this ratio $\mathcal{R}_{error}$ in subsequent experiments.

Experiments

In this section, we conduct a series of experiments on two base models to answer the following research questions:

- **RQ1**: Does KnowPO outperform other approaches for resolving knowledge conflict across various base models and datasets?
- **RQ2**: What impact does each component has on the overall performance?
- **RQ3**: How does KnowPO alter the way LLMs utilize parametric knowledge?
- **RQ4**: How sensitive is KnowPO to hyper-parameters data ratio $\mathcal{R}_{error}$?
- **RQ5**: Does KnowPO training conducted in general domains remain effective in out-of-distribution (OOD) scenarios?

Code — https://github.com/Nostradamus4869/KnowPO

Experimental Setup

Datasets We constructed the KnowPO training dataset based on SQuAD 2.0 (Rajpurkar, Jia, and Liang 2018). The test datasets comprise the following three types: (1) **SQuAD 2.0-Eval**, a validation set partitioned using the same construction method. (2) **Open-source counterfactual datasets:** RGB (Chen et al. 2023) and KNOT (Liu et al. 2024) are two general-domain QA datasets containing counterfactual knowledge and contexts. We augmented these datasets with irrelevant contexts for testing purposes. Notably, RGB is a Chinese dataset. (3) **Domain-specific dataset:** CMB (Wang et al. 2023a) is a multi-task QA dataset in the medical domain, encompassing 269,359 questions across four clinical medicine specialties of physicians, nurses, medical technicians, and pharmacists. Due to quantity constraints, we randomly sample 4,000 questions for testing.

Compared Methods In order to explore the advantages of the KnowPO, we compare the KnowPO results against five other models: (1) **Base Model (Base)** answers user questions based on supplementary external knowledge, which

LLM Turbo	LLM	Baichuan2-7B-Chat						Llama2-13B-Chat					
Method	**Dataset**	Squad2.0-Eval		RGB		KNOT		Squad2.0-Eval		RGB		KNOT	
	Metric	R_{Ad}	R_{Ro}	R_{Ad}	R_{Ro}	R_{Ad}	R_{Ro}	R_{Ad}	R_{Ro}	R_{Ad}	R_{Ro}	R_{Ad}	R_{Ro}
Baselines	Base	43.51	9.80	65.00	24.00	26.42	7.65	52.71	11.95	69.00	25.00	49.66	21.67
	Prompt	53.74	8.60	79.50	19.50	44.65	14.51	60.76	10.59	73.50	19.50	41.14	22.62
	COT	54.65	10.20	77.50	21.00	44.13	15.29	57.13	12.85	70.50	25.00	41.06	23.53
	COT-VE	44.83	8.41	66.00	19.50	27.71	13.88	54.52	10.17	70.00	14.50	52.33	18.35
	KAFT	58.83	21.43	75.00	27.00	54.45	17.93	65.73	34.34	73.50	29.50	62.21	24.47
	CAD	35.83	7.50	55.50	22.50	21.72	6.99	41.73	10.94	64.50	23.50	35.71	19.96
Ours	**KnowPO**	80.64	38.77	93.50	37.00	69.95	39.73	76.11	44.64	83.50	37.50	77.03	38.28
Performance Gain ↑		37.07∼ 125.06	80.91∼ 416.93	17.61∼ 68.47	37.04∼ 89.74	28.47∼ 222.05	121.58∼ 468.38	15.79∼ 82.39	29.58∼ 338.94	13.61∼ 29.46	27.12∼ 158.62	23.82∼ 115.71	49.94∼ 108.61
Ablation	KnowPO (w/o DPO)	71.09	36.50	89.50	31.00	64.45	36.50	75.96	42.87	80.00	35.00	70.28	36.21
	KnowPO (w/o SFT)	69.39	37.50	92.50	35.00	66.92	38.76	74.73	42.86	81.00	34.00	69.39	36.74
	KnowPO (w/o Aligned)	54.45	43.45	71.50	43.00	48.29	45.17	61.36	50.30	70.00	42.50	50.71	46.27

Table 2: Performance comparison (in percent) on Squad2.0-Eval, RGB and KNOT. The best-performing model is underlined.

can be considered as fundamental retrieve-and-read framework in RAG (Lewis et al. 2021). We selected Baichuan2-7B-chat (Yang et al. 2023) and Llama2-13B-chat (Touvron et al. 2023) as the base model and explored the gains brought by KnowPO: (2) **Naive Prompt-based Method (Prompt)** employs meticulously designed prompts to enhance the model's capability to adhere to external knowledge (Zhou et al. 2023). (3) **Advanced Prompt-based Method**: Chain of Thought (COT) (Wei et al. 2023) is a common method to enhance the performance of LLMs in downstream tasks. COT-VE (Zhao et al. 2023a) extends COT by guiding LLM to identify conflicting knowledge and modify its responses accordingly. (4) **Finetuning**: KAFT (Li et al. 2022) employs instruction fine-tuning to improve the LLM's adherence to contexts of varying relevance. (5) **Decode-Based Method**: CAD (Shi et al. 2023b) uses a contrastive decoding-like method to adjust the probabilities of output tokens.

Metrics We designed statistical metrics to evaluate the two capabilities of LLMs. For adherence capability, we utilized the conflicting contexts from the test set as supplementary knowledge, measuring the proportion R_{Ad} of LLM responses that align with the conflicting knowledge within these contexts. For the RGB and KNOT datasets, the conflicting knowledge exclusively consists of counterfactual knowledge. For noise robustness, we employed the irrelevant contexts from the test set as supplementary knowledge, examining the proportion R_{Ro} of LLM responses that correspond with the model's parameter knowledge.

Performance Comparison(RQ 1)

To answer RQ1, we conduct experiments and report results of the two metrics on Squad2.0-Eval, RGB and KNOT with two LLM turbos, as illustrated in Table 2. From the reported results, we can find the following observations:

Comparison of Baseline Methods and Base LLMs. Through comparison, we observe that the KAFT method, fine-tuned with instructions, consistently outperforms across all experimental groups. This superior performance is primarily attributed to the use of contexts with varying degrees of relevance during fine-tuning, which significantly enhances the LLM's ability to focus on pertinent data while filtering out noise. In contrast, methods relying on the LLM's inherent capabilities for single or multiple interactions, such as Prompt or COT, tend to indiscriminately depend on external knowledge due to the LLM's limited noise recognition ability, leading to an increase in R_{Ad}, but a sharp decline in R_{Ro}. Particularly, COT-VE introduces additional noise by incorporating external knowledge for verification and editing , further complicating the model's ability to discern relevant information. As for CAD, as noted in related research, the contrastive decoding strategy compromises response coherence and utility, performing well on simple datasets like RGB but failing on more complex ones like Squad2.0-Eval and KNOT, thereby losing its practical value.

Comparison of KnowPO and other methods. Firstly, it is evident that our mothed, KnowPO, outperforms the baseline methods across all metrics. For instance, the R_{Ad} and R_{Ro} scores see an improvement of approximately **37.07%-125.06%** and **80.91%-416.93%** for the Squad2.0-Eval dataset with Baichuan2-7B-Chat. Moreover, compared to KAFT, best model in baselines, KnowPO uses more complicated contexts and comprehensive negative signals to enhance LLM's adherence capability and noise robustness.

Ablation Study(RQ 2)

To answer RQ2, we perform ablation studies to verify the effectiveness of KnowPO, as illustrated in Table 2. Our observation can be summarized as follows:

Effect of training phase. Both the SFT and DPO phases positively contribute to enhancing the adherence capability and noise robustness of LLMs. Additionally, the preference learning method incorporating negative signals slightly outperforms SFT in improving the LLM's ability to utilize external knowledge, demonstrating the effectiveness of both training approaches.

Effect of length imbalance. When data length are not aligned, we observe a significant impact of length bias, which slightly enhances R_{Ro} but substantially reduces R_{Ad}. This is due to the model's inherent tendency to generate more verbose parametric answers, while the conflict answers derived through dataset construction are relatively short. Consequently, the model develops a preference for generating longer responses. Without length alignment between

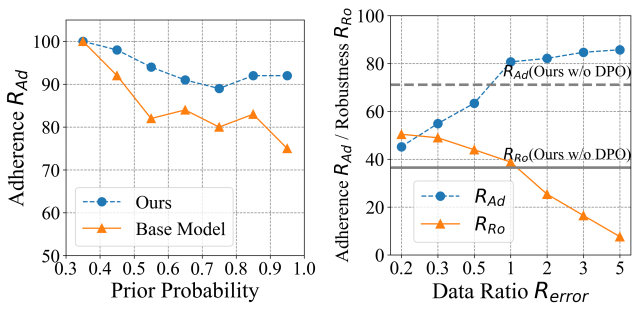

Figure 2: (**Left**.) The adherence ratio of LLM when encountering conflicting context with different prior probability on dataset RGB with Baichuan2-7B-Chat. (**Right**.) Hyper-parameter study with data ratio $\mathcal{R}_{error}$ on Squad2.0-Eval with Baichuan2-7B-Chat.

conflict and parametric answers, the model tends to consistently rely on parametric answers, thereby neglecting external knowledge and disrupting the balance between adherence capability and noise robustness.

Model Prior Analyse(RQ 3)

The LLM's confidence in its responses is one of the factors influencing whether it prefers internal or external knowledge (Wu, Wu, and Zou 2024). We recorded the LLM's prior probability for parameter knowledge on the RGB dataset and measured the proportion of instances in each prior probability interval where the LLM followed external knowledge in conflicting contexts. The model's prior response probability is computed from the average log probability of the response tokens without external knowledge. The results in Figure 2 show that, for base LLM, there is a general negative correlation between the prior probability of an answer and the proportion of following external knowledge; that is, the higher the prior probability, the less likely the answer is to be altered. However, after fine-tuning with KnowPO, although the overall trend remains negatively correlated, the trend is significantly mitigated, indicating that our method effectively enhances the LLM's adherence to external knowledge.

Hyper-parameter Study(RQ 4)

As analyzed in Methodology, two types of preference pairs exhibit distinctly opposite behavioral tendencies: those simulating error contextual ignorance in conflicting contexts and those simulating error contextual overinclusion in irrelevant contexts. We conducted a series of analyses by adjusting the ratio R_{error} between these two types of preference pairs from the list $[0.2, 0.3, 0.5, 1, 2, 3, 5]$. The results in Figure 2 indicate that as the proportion of the first type of preference pairs increases, the LLM becomes more inclined to utilize contextual knowledge, enhancing its adherence capability but also becoming more susceptible to noise, which in turn reduces its noise robustness. Conversely, as the proportion of the second type increases, the LLM tends to disregard contextual information and respond directly, resulting in reduced R_{Ad} but improved R_{Ro}. Notably, as the ratio R_{error} increases from 1, the rate of improvement in adherence slows, while the decline in robustness becomes more pronounced. When the ratio R_{error} decreases from 1, the curvature of R_{Ad} and R_{Ro} also shows the opposite trend. Based on these findings, we ultimately selected $R_{error} = 1$ as the optimal ratio, ensuring balanced improvements in both capabilities compared to SFT training.

	Baichuan2-7B-Chat		Llama2-13B-Chat	
	R_{Ad}	R_{Ro}	R_{Ad}	R_{Ro}
Base	58.69	10.66	60.95	8.21
KnowPO(w/o DPO)	95.66	23.70	83.88	16.53
KnowPO	96.23	24.12	87.46	21.24

Table 3: Performance comparison (in percent) on CMB

Model	KnowPO(w/o DPO)	KnowPO
Baichuan2-7B-Chat	3.65%	4.70%
Llama2-13B-Chat	3.51%	4.31%

Table 4: The match rate between LLM's parameter answers and conflicting answers after training.

Generalization Analysis(RQ 5)

To demonstrate the robust generalization capability of our method beyond general domain, we conducted experiments on the CMB medical test set. Using medical triplets and documents, we created supplementary contexts and a conflict dataset for 4,000 CMB questions in medical domain. Results in Table 3 show that KnowPO-trained models effectively enhance adherence capability and noise robustness when transferred to domain-specific contexts. The higher scores on CMB compared to those in Table 2 can be attributed to the fact that the contexts we constructed were less challenging than the real-world RAG knowledge.

A potential risk of incorporating QA pairs and conflicting knowledge into the training data is the inadvertent introduction of harmful information to the model. To evaluate whether the KnowPO-trained model retained conflicting knowledge, we utilized prompts designed to extract LLM's parameter knowledge. The results, presented in Table 4, indicate that the model retained virtually no conflicting knowledge after the SFT and DPO phases. This finding corroborates that our method enhances the LLM's ability to leverage external knowledge rather than injecting specific knowledge.

Conclusions and Future Works

In this paper, we propose KnowPO, a **Know**ledge-aware **P**reference **O**ptimization strategy to enhance LLM's adherence capability and noise robustness to external knowledge. We simulate two error types—Contextual Ignorance and Contextual Overinclusion—and use negative gradient terms in DPO objectives to minimize undesired responses. By aligning data lengths and balancing ratios, we mitigate preference imbalances in DPO. Experiments across diverse datasets confirm KnowPO's efficacy and generalization. In the future, we will explore how different context compositions affect LLMs' ability to utilize external knowledge.

Acknowledgments

This work is supported by the National Natural Science Foundation of China (No.U23A20468).

References

Asai, A.; Min, S.; Zhong, Z.; and Chen, D. 2023a. Retrieval-based language models and applications. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 6: Tutorial Abstracts)*, 41–46.

Asai, A.; Wu, Z.; Wang, Y.; Sil, A.; and Hajishirzi, H. 2023b. Self-RAG: Learning to Retrieve, Generate, and Critique through Self-Reflection. *arXiv preprint arXiv:2310.11511*.

Cao, M.; Dong, Y.; Wu, J.; and Cheung, J. C. K. 2020. Factual Error Correction for Abstractive Summarization Models. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 6251–6258. Online: Association for Computational Linguistics.

Chen, H.-T.; Zhang, M. J. Q.; and Choi, E. 2022. Rich Knowledge Sources Bring Complex Knowledge Conflicts: Recalibrating Models to Reflect Conflicting Evidence. arXiv:2210.13701.

Chen, J.; Lin, H.; Han, X.; and Sun, L. 2023. Benchmarking Large Language Models in Retrieval-Augmented Generation. arXiv:2309.01431.

Gao, L.; Schulman, J.; and Hilton, J. 2022. Scaling Laws for Reward Model Overoptimization. arXiv:2210.10760.

He, H.; Zhang, H.; and Roth, D. 2022. Rethinking with Retrieval: Faithful Large Language Model Inference. arXiv:2301.00303.

Izacard, G.; Lewis, P.; Lomeli, M.; Hosseini, L.; Petroni, F.; Schick, T.; Dwivedi-Yu, J.; Joulin, A.; Riedel, S.; and Grave, E. 2022. Atlas: Few-shot Learning with Retrieval Augmented Language Models. arXiv:2208.03299.

Ji, Z.; Lee, N.; Frieske, R.; Yu, T.; Su, D.; Xu, Y.; Ishii, E.; Bang, Y. J.; Madotto, A.; and Fung, P. 2023a. Survey of hallucination in natural language generation. *ACM Computing Surveys*, 55(12): 1–38.

Ji, Z.; Lee, N.; Frieske, R.; Yu, T.; Su, D.; Xu, Y.; Ishii, E.; Bang, Y. J.; Madotto, A.; and Fung, P. 2023b. Survey of Hallucination in Natural Language Generation. *ACM Comput. Surv.*, 55(12).

Jiang, X.; Zhang, R.; Xu, Y.; Qiu, R.; Fang, Y.; Wang, Z.; Tang, J.; Ding, H.; Chu, X.; Zhao, J.; and Wang, Y. 2024. HyKGE: A Hypothesis Knowledge Graph Enhanced Framework for Accurate and Reliable Medical LLMs Responses. arXiv:2312.15883.

Jin, Z.; Cao, P.; Chen, Y.; Liu, K.; Jiang, X.; Xu, J.; Li, Q.; and Zhao, J. 2024. Tug-of-War Between Knowledge: Exploring and Resolving Knowledge Conflicts in Retrieval-Augmented Language Models. arXiv:2402.14409.

Kandpal, N.; Deng, H.; Roberts, A.; Wallace, E.; and Raffel, C. 2023. Large language models struggle to learn long-tail knowledge. In *International Conference on Machine Learning*, 15696–15707. PMLR.

Kaplan, J.; McCandlish, S.; Henighan, T.; Brown, T. B.; Chess, B.; Child, R.; Gray, S.; Radford, A.; Wu, J.; and Amodei, D. 2020. Scaling Laws for Neural Language Models. arXiv:2001.08361.

Kassner, N.; and Schütze, H. 2019. Negated and Misprimed Probes for Pretrained Language Models: Birds Can Talk, But Cannot Fly. *Cornell University - arXiv,Cornell University - arXiv*.

Lewis, P.; Perez, E.; Piktus, A.; Petroni, F.; Karpukhin, V.; Goyal, N.; Küttler, H.; Lewis, M.; tau Yih, W.; Rocktäschel, T.; Riedel, S.; and Kiela, D. 2021. Retrieval-Augmented Generation for Knowledge-Intensive NLP Tasks. arXiv:2005.11401.

Li, D.; Rawat, A. S.; Zaheer, M.; Wang, X.; Lukasik, M.; Veit, A.; Yu, F.; and Kumar, S. 2022. Large Language Models with Controllable Working Memory. arXiv:2211.05110.

Li, Q.; Guo, S.; Wu, J.; Li, J.; Sheng, J.; Peng, H.; and Wang, L. 2023. Event extraction by associating event types and argument roles. *IEEE Transactions on Big Data*.

Li, Q.; Li, J.; Wu, J.; Peng, X.; Ji, C.; Peng, H.; Wang, L.; and Philip, S. Y. 2024. Triplet-aware graph neural networks for factorized multi-modal knowledge graph entity alignment. *Neural Networks*, 106479.

Lin, Y.; Ma, X.; Chu, X.; Jin, Y.; Yang, Z.; Wang, Y.; and Mei, H. 2024. Lora dropout as a sparsity regularizer for overfitting control. *arXiv preprint arXiv:2404.09610*.

Liu, Y.; Yao, Z.; Lv, X.; Fan, Y.; Cao, S.; Yu, J.; Hou, L.; and Li, J. 2024. Untangle the KNOT: Interweaving Conflicting Knowledge and Reasoning Skills in Large Language Models. arXiv:2404.03577.

Longpre, S.; Perisetla, K.; Chen, A.; Ramesh, N.; DuBois, C.; and Singh, S. 2022. Entity-Based Knowledge Conflicts in Question Answering. arXiv:2109.05052.

Ma, X.; Chu, X.; Yang, Z.; Lin, Y.; Gao, X.; and Zhao, J. 2024. Parameter Efficient Quasi-Orthogonal Fine-Tuning via Givens Rotation. In *ICML*.

Park, R.; Rafailov, R.; Ermon, S.; and Finn, C. 2024. Disentangling Length from Quality in Direct Preference Optimization. arXiv:2403.19159.

Rafailov, R.; Sharma, A.; Mitchell, E.; Ermon, S.; Manning, C. D.; and Finn, C. 2024. Direct Preference Optimization: Your Language Model is Secretly a Reward Model. arXiv:2305.18290.

Rajpurkar, P.; Jia, R.; and Liang, P. 2018. Know What You Don't Know: Unanswerable Questions for SQuAD. arXiv:1806.03822.

Shi, F.; Chen, X.; Misra, K.; Scales, N.; Dohan, D.; Chi, E.; Schärli, N.; and Zhou, D. 2023a. Large Language Models Can Be Easily Distracted by Irrelevant Context. arXiv:2302.00093.

Shi, W.; Han, X.; Lewis, M.; Tsvetkov, Y.; Zettlemoyer, L.; and tau Yih, S. W. 2023b. Trusting Your Evidence: Hallucinate Less with Context-aware Decoding. arXiv:2305.14739.

Si, C.; Gan, Z.; Yang, Z.; Wang, S.; Wang, J.; Boyd-Graber, J.; and Wang, L. 2023. Prompting GPT-3 To Be Reliable. arXiv:2210.09150.

Singhal, P.; Goyal, T.; Xu, J.; and Durrett, G. 2024. A Long Way to Go: Investigating Length Correlations in RLHF. arXiv:2310.03716.

Tan, H.; Sun, F.; Yang, W.; Wang, Y.; Cao, Q.; and Cheng, X. 2024. Blinded by Generated Contexts: How Language Models Merge Generated and Retrieved Contexts When Knowledge Conflicts? arXiv:2401.11911.

Taylor, R.; Kardas, M.; Cucurull, G.; Scialom, T.; Hartshorn, A.; Saravia, E.; Poulton, A.; Kerkez, V.; and Stojnic, R. 2022. Galactica: A Large Language Model for Science. arXiv:2211.09085.

Touvron, H.; Martin, L.; Stone, K.; Albert, P.; Almahairi, A.; Babaei, Y.; Bashlykov, N.; Batra, S.; Bhargava, P.; Bhosale, S.; Bikel, D.; Blecher, L.; Ferrer, C. C.; Chen, M.; Cucurull, G.; Esiobu, D.; Fernandes, J.; Fu, J.; Fu, W.; Fuller, B.; Gao, C.; Goswami, V.; Goyal, N.; Hartshorn, A.; Hosseini, S.; Hou, R.; Inan, H.; Kardas, M.; Kerkez, V.; Khabsa, M.; Kloumann, I.; Korenev, A.; Koura, P. S.; Lachaux, M.-A.; Lavril, T.; Lee, J.; Liskovich, D.; Lu, Y.; Mao, Y.; Martinet, X.; Mihaylov, T.; Mishra, P.; Molybog, I.; Nie, Y.; Poulton, A.; Reizenstein, J.; Rungta, R.; Saladi, K.; Schelten, A.; Silva, R.; Smith, E. M.; Subramanian, R.; Tan, X. E.; Tang, B.; Taylor, R.; Williams, A.; Kuan, J. X.; Xu, P.; Yan, Z.; Zarov, I.; Zhang, Y.; Fan, A.; Kambadur, M.; Narang, S.; Rodriguez, A.; Stojnic, R.; Edunov, S.; and Scialom, T. 2023. Llama 2: Open Foundation and Fine-Tuned Chat Models. arXiv:2307.09288.

Vu, M. D.; Wang, H.; Li, Z.; Chen, J.; Zhao, S.; Xing, Z.; and Chen, C. 2024. GPTVoiceTasker: LLM-Powered Virtual Assistant for Smartphone. arXiv:2401.14268.

Wang, X.; Chen, G. H.; Song, D.; Zhang, Z.; Chen, Z.; Xiao, Q.; Jiang, F.; Li, J.; Wan, X.; Wang, B.; and Li, H. 2023a. CMB: A Comprehensive Medical Benchmark in Chinese. arXiv:2308.08833.

Wang, Y.; Kordi, Y.; Mishra, S.; Liu, A.; Smith, N. A.; Khashabi, D.; and Hajishirzi, H. 2023b. Self-Instruct: Aligning Language Models with Self-Generated Instructions. arXiv:2212.10560.

Wei, J.; Wang, X.; Schuurmans, D.; Bosma, M.; Ichter, B.; Xia, F.; Chi, E.; Le, Q.; and Zhou, D. 2023. Chain-of-Thought Prompting Elicits Reasoning in Large Language Models. arXiv:2201.11903.

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

Xie, J.; Zhang, K.; Chen, J.; Lou, R.; and Su, Y. 2024. Adaptive Chameleon or Stubborn Sloth: Revealing the Behavior of Large Language Models in Knowledge Conflicts. arXiv:2305.13300.

Xu, R.; Qi, Z.; Guo, Z.; Wang, C.; Wang, H.; Zhang, Y.; and Xu, W. 2024. Knowledge Conflicts for LLMs: A Survey. arXiv:2403.08319.

Xue, B.; Wang, W.; Wang, H.; Mi, F.; Wang, R.; Wang, Y.; Shang, L.; Jiang, X.; Liu, Q.; and Wong, K.-F. 2023. Improving Factual Consistency for Knowledge-Grounded Dialogue Systems via Knowledge Enhancement and Alignment. arXiv:2310.08372.

Yang, A.; Xiao, B.; Wang, B.; Zhang, B.; Bian, C.; Yin, C.; Lv, C.; Pan, D.; Wang, D.; Yan, D.; Yang, F.; Deng, F.; Wang, F.; Liu, F.; Ai, G.; Dong, G.; Zhao, H.; Xu, H.; Sun, H.; Zhang, H.; Liu, H.; Ji, J.; Xie, J.; Dai, J.; Fang, K.; Su, L.; Song, L.; Liu, L.; Ru, L.; Ma, L.; Wang, M.; Liu, M.; Lin, M.; Nie, N.; Guo, P.; Sun, R.; Zhang, T.; Li, T.; Li, T.; Cheng, W.; Chen, W.; Zeng, X.; Wang, X.; Chen, X.; Men, X.; Yu, X.; Pan, X.; Shen, Y.; Wang, Y.; Li, Y.; Jiang, Y.; Gao, Y.; Zhang, Y.; Zhou, Z.; and Wu, Z. 2023. Baichuan 2: Open Large-scale Language Models. arXiv:2309.10305.

Zhao, R.; Li, X.; Joty, S.; Qin, C.; and Bing, L. 2023a. Verify-and-Edit: A Knowledge-Enhanced Chain-of-Thought Framework. arXiv:2305.03268.

Zhao, W. X.; Zhou, K.; Li, J.; Tang, T.; Wang, X.; Hou, Y.; Min, Y.; Zhang, B.; Zhang, J.; Dong, Z.; Du, Y.; Yang, C.; Chen, Y.; Chen, Z.; Jiang, J.; Ren, R.; Li, Y.; Tang, X.; Liu, Z.; Liu, P.; Nie, J.-Y.; and Wen, J.-R. 2023b. A Survey of Large Language Models. arXiv:2303.18223.

Zhou, W.; Zhang, S.; Poon, H.; and Chen, M. 2023. Context-faithful Prompting for Large Language Models. arXiv:2303.11315.

Ziegler, D. M.; Stiennon, N.; Wu, J.; Brown, T. B.; Radford, A.; Amodei, D.; Christiano, P.; and Irving, G. 2020. Fine-Tuning Language Models from Human Preferences. arXiv:1909.08593.