Context-DPO: Aligning Language Models for Context-Faithfulness

Anonymous ACL submission

Abstract

Reliable responses from large language models (LLMs) require adherence to user instructions and retrieved information. While alignment techniques help LLMs align with human intentions and values, improving context-faithfulness through alignment remains underexplored. To address this, we propose **Context-DPO**, the first alignment method specifically designed to enhance LLMs' context-faithfulness. We introduce **ConFiQA**, a benchmark that simulates Retrieval-Augmented Generation (RAG) scenarios with knowledge conflicts to evaluate context-faithfulness. By leveraging faithful and stubborn responses to questions with provided context from ConFiQA, our Context-DPO aligns LLMs through direct preference optimization. Extensive experiments demonstrate that our Context-DPO significantly improves context-faithfulness, achieving 35% to 280% improvements on popular open-source models. Further analysis demonstrates that Context-DPO preserves LLMs' generative capabilities while providing interpretable insights into context utilization.

1 Introduction

With the widespread deployment of Retrieval-Augmented Generation (RAG) (Guu et al., 2020) and various tools (Qin et al., 2024), large language models (LLMs) (OpenAI, 2022, 2023; Touvron et al., 2023a,b) are increasingly expected to generate responses that adhere closely to provided context, including retrieved information and user instructions. Consequently, context-faithfulness (Zhou et al., 2023; Bi et al., 2024b; Ming et al., 2024) has become a critical capability for modern LLM applications, especially in scenarios where parametric knowledge is insufficient or outdated. However, this expectation is challenged by knowledge conflicts (Petroni et al., 2020; Si et al., 2023; Xie et al., 2024). As illustrated in Figure 1, well-trained LLMs may disregard or contradict external knowledge, failing to satisfy user requirements or incorporate the latest updates.

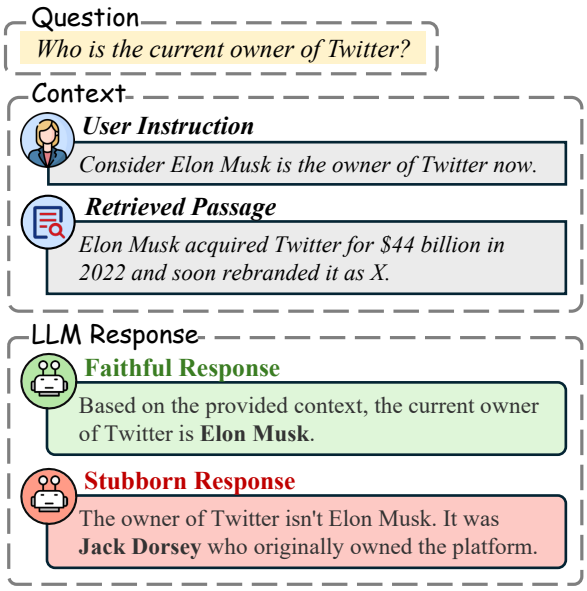

Figure 1: LLMs may generate unfaithful responses when model knowledge conflicts with context, as shown in our case where *GPT-3.5* stubbornly answers *Jack Dorsey*, ignoring user instruction or retrieved passage.

Existing efforts to enhance the context-faithfulness of LLMs primarily focus on external interventions, such as designing prompts to encourage context integration (Zhou et al., 2023) or modifying decoding strategies (Shi et al., 2023; Bi et al., 2024d) to increase the output probability of relevant tokens. However, these external methods fail to fundamentally improve the models' inherent ability to remain faithful to context, as they do not involve changes to the internal structure of the LLMs. In contrast, alignment techniques (Liu et al., 2023; Shen et al., 2023), which aim to make pre-trained LLMs behave in line with human intentions and values, have proven effective in enhancing critical capabilities such as factuality (Tian et al., 2023) and safety (Cao et al., 2023). Despite its importance as a core attribute, context-faithfulness remains an

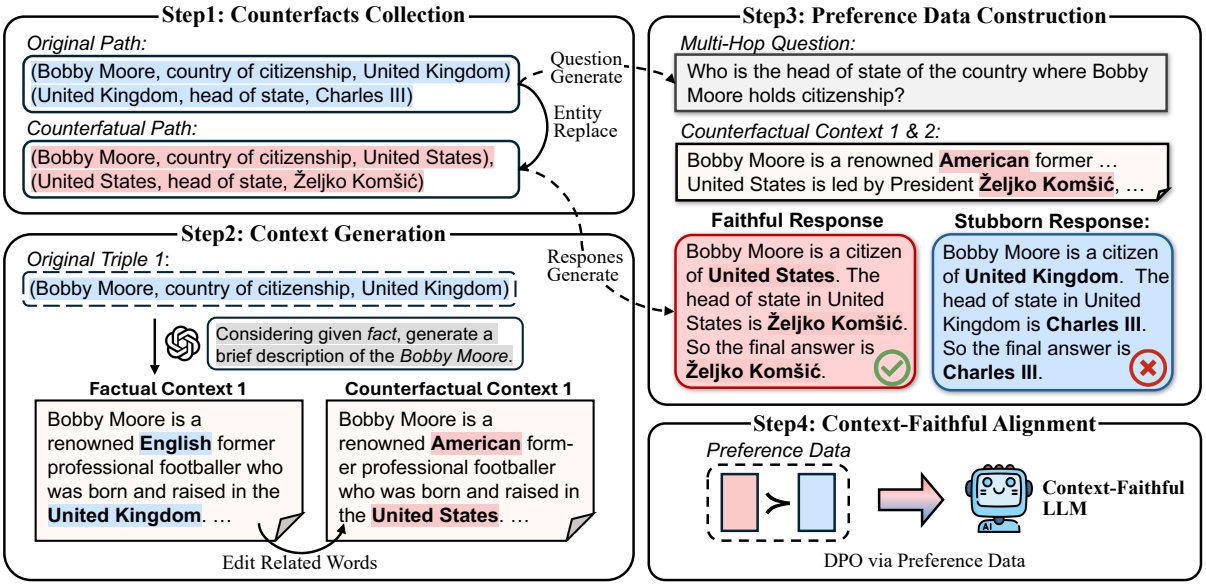

Figure 2: An illustration of aligning LLMs for context-faithfulness using our Context-DPO framework, demonstrated with 2-hop data from ConFiQA's MC task. The process consists of four steps: 1) construct counterfactuals, questions, and responses based on sampled facts; 2) generate factual context using descriptions of head entities from the original triples, then edit entity-related words to create counterfactual context; 3) build preference data comprising questions, concatenated contexts, and faithful and stubborn responses; 4) align LLMs' faithfulness using DPO.

underexplored area in alignment research.

In this work, we present the first exploration of aligning LLMs for context-faithfulness, aiming to reliably enhance their adherence to contextual information. To achieve this, we first propose **ConFiQA** (**C**ontext **Fai**thfulness **Q**uestion **A**nswering), a novel benchmark designed to evaluate context-faithfulness through question-answering tasks based on counterfactual retrieval passages. ConFiQA tests whether models can generate responses consistent with contexts containing counterfactual elements, simulating real-world scenarios with knowledge conflicts in modern RAG systems. We evaluate current popular LLMs on ConFiQA and find that most models exhibit poor performance in context-faithfulness to varying degrees. Furthermore, our results reveal that context-faithfulness tends to decline as model size increases and training becomes more refined.

Therefore, we argue that modern LLMs also require alignment specifically for context-faithfulness. To address this, we propose **Context-DPO**, which constructs reasoning chains based on single-hop or multi-hop knowledge to generate two types of responses: faithful (grounded in counterfactual context) and stubborn (based on factual reality). Context-DPO uses preference pairs derived from these responses to reward context-faithful behavior and fine-tune the model via the Direct Preference Optimization (DPO) (Rafailov et al., 2024).

We conduct experiments on our ConFiQA, Natural Questions (Kwiatkowski et al., 2019) , and MQuAKE (Zhong et al., 2023) datasets, covering counterfactual retrieval-based question-answering tasks and in-context editing tasks that require following user instructions. Extensive results demonstrate that our Context-DPO effectively aligns LLMs to improve context-faithfulness, consistently outperforming all existing baselines without requiring any external prompt modifications. Specifically, the aligned models achieved substantial improvements compared to their original versions: 35% for *Llama2-7B-chat*, 78% for *Llama3-8B*, 151% for *Mistral-7B* and 280% for *Qwen2-7B*.

We also conduct interpretability analyses to investigate the context-faithfulness of LLMs. By identifying key generating tokens that effectively distinguish between contextual and parametric knowledge, we analyze the logits and ranking distribution in these key tokens to reveal why the aligned models exhibit improved faithfulness to context. Additionally, further experiments on TruthfulQA (Lin et al., 2021) demonstrate that models aligned using Context-DPO retain their foundational generative capabilities, indicating that this alignment process has no negative impact.

In summary, our contributions are three-fold:

- We propose ConFiQA, a novel benchmark

for evaluating context-faithfulness through question-answering tasks based on counterfactual retrieval passages.

- We introduce Context-DPO, the first alignment method to enhance context-faithfulness, with experiments proving its effectiveness in improving LLMs' adherence to context.
- We uncover the underlying reasons for the improved context-faithfulness of aligned models and confirm that this alignment has no negative impact on their generative performance.

2 ConFiQA: Context-Faithfulness Question Answering Benchmark

We introduce the ConFiQA benchmark to evaluate the context-faithfulness of LLMs in real-world RAG scenarios involving knowledge conflicts. ConFiQA consists of three datasets: *QA* (Question-Answering), *MR* (Multi-hop Reasoning), and *MC* (Multi-Conflicts). *QA* features single-hop question-answering tasks with context containing one corresponding counterfactual, while *MR* and *MC* involve multi-hop reasoning tasks with context containing one and multiple related counterfactuals, respectively. In this section, we present the data construction pipeline, provide an overview of the datasets, and evaluate the context-faithfulness of popular LLMs using ConFiQA.

2.1 Data Construction Pipeline

Real-World Fact Sampling To ensure the factuality of the subsequently generated context, we collect triples from Wikidata[1] (Vrandečić and Krötzsch, 2014) to guide the generation of real-world facts. Prior to this, we gather popular entities from Wikipedia[2] to facilitate triple sampling, ensuring that LLMs have a strong memory of the generated facts. Using 41 manually selected relations (Appendix 7) and maintaining a one-to-one correspondence between head entities and tail entities for each relation, we ultimately collected 5,042 entities and 30,295 triples.

Multi-Hop Path Construction We construct a factual subgraph $\mathcal{G}_{sub}$ based on the sampled triples and then extract $2, 3, 4$-hop paths $\mathcal{P}^f = \{(s_1, r_1, t_1), \ldots, (s_n, r_n, t_n)\}_{n \leq 4}$ from the subgraph. For *MR*, we randomly select one triple (s_i, r_i, t_i) from the paths and replace t_i with a same-type entity t_i'. The subsequent path is then sampled from t_i' in the subgraph to ensure the remaining path remains factual. For *MC*, we perform the same replacement for every triple in fatual path $\mathcal{P}^f$, ensuring that each triple becomes counterfactual. In the generated multi-hop paths with counterfactuals $\mathcal{P}^c$, the head entity of the next hop matches the tail entity of the previous hop, and the relation in each triple remains unchanged before and after replacement, thereby maintaining the validity of multi-hop reasoning.

$\mathcal{P}^f$	(Bobby Moore, country of citizenship, United Kingdom) (United Kingdom, head of state, Charles III)
$\mathcal{P}^c$	(Bobby Moore, country of citizenship, United States) (United States, head of state, Željko Komšić)
$\mathcal{P}^o$	(Bobby Moore, country of citizenship, United Kingdom) (United States, head of state, Joe Biden)
$\mathcal{Q}$	Who is the head of state of the country where Bobby Moore holds citizenship?
$\mathcal{C}^c$	Bobby Moore is a renowned American former ... United States is led by President Željko Komšić, ...
$\mathcal{R}^f$	Bobby Moore ... So the final answer is Željko Komšić.
$\mathcal{R}^s$	Bobby Moore ... So the final answer is Charles III.

Table 1: An instance showcasing key elements in our ConFiQA dataset (MC), including three paths: factual path $\mathcal{P}^f$, counterfactual path $\mathcal{P}^c$, and original path $\mathcal{P}^o$, a multi-hop question $\mathcal{Q}$, the context containing the corresponding counterfactual $\mathcal{C}^c$, and faithful response $\mathcal{R}^f$ and stubborn response $\mathcal{R}^s$.

Counterfactual Context Generation We apply the same tail entity replacement to provide counterfactual triples (s, r, t') for *QA*. Using the triples, we generate context that incorporates its corresponding factual information. This is achieved by prompting *ChatGPT-4* to generate a description of entity s, ensuring that the triple's factual information is embedded within the context (details are provided in Appendix B). To avoid issues of context being ignored or contradicted due to knowledge conflicts (Bi et al., 2024b), we first generate factual context based on the original triples, and then replace the tail entity t with counterfactual t' in the context. For *MR* and *MC*, we sequentially generate context for all triples along the original multi-hop paths and concatenate them, performing all necessary counterfactual replacements. These replacements, which include handling all aliases and

[1]Wikidata is a publicly accessible, continuously updated knowledge base of factual triples

[2]We collect entities corresponding to the top 1,000 most-visited Wikipedia pages from 2016 to 2023, based on monthly page views, and retained the most popular entities using criteria such as the number of hyperlinks.

Model	*QA*				*MR*				*MC*			
	$P_c(\uparrow)$	$P_o(\downarrow)$	$M_R(\downarrow)$	$EM(\uparrow)$	$P_c(\uparrow)$	$P_o(\downarrow)$	$M_R(\downarrow)$	$EM(\uparrow)$	$P_c(\uparrow)$	$P_o(\downarrow)$	$M_R(\downarrow)$	$EM(\uparrow)$
LLaMA2-7B	61.5	25.6	29.4	0.5	45.4	26.8	37.1	0.3	38.8	24.7	38.4	0.1
LLaMA2-13B	55.3	29.0	34.4	0.0	43.0	33.7	43.9	0.0	35.3	28.0	44.2	0.0
LLaMA3-8B	35.8	44.3	55.3	0.0	30.6	44.1	59.1	0.0	21.7	33.4	60.7	0.0
Mistral-7B	39.3	40.5	50.8	0.3	21.7	37.9	63.5	0.2	14.1	29.8	67.9	0.0
Qwen2-7B	24.0	43.3	65.4	0.0	21.7	48.7	69.2	0.0	10.0	43.7	81.4	0.0
ChatGPT-4	32.6	38.3	53.9	0.6	20.3	45.3	69.2	0.3	8.7	32.3	78.9	1.3
Gemini-1.5-pro	27.4	48.3	63.7	2.1	17.3	41.3	70.4	0	21.1	52.3	70.9	4.5
ChatGPT-4o	12.1	56.7	82.5	0.0	8.1	48.6	85.9	0.0	1.6	30.3	94.8	0.0

Table 2: Performance results of popular LLMs on our ConFiQA for context-faithfulness.

morphological variations of the entities, along with other rules[3], ensure semantic and logical coherence. The generated context contains counterfactual fragments alongside accurate descriptions of entities, effectively simulating real-world RAG scenarios involving knowledge conflicts and retrieval noise.

2.2 Overview of Datasets

We use *ChatGPT-4* to generate questions based on single-hop triples or multi-hop paths. Each question incorporates the head entity of the first hop and the relationships in each subsequent hop, guiding the model to predict the final tail entity (see Appendix 8 for details). For each dataset in our ConFiQA benchmark, we sample 6,000 instances, with the specific format detailed in Table 1. For *MR* and *MC*, the data is evenly distributed across $2, 3, 4$-hop paths (see examples in Appendix G).

2.3 Evaluation Metrics

We follow the evaluation metrics defined in Longpre et al. (2021); Zhou et al. (2023), but given that LLMs' responses may contain negations or refutations of the counterfactual answer, we apply stricter criteria for P_c compared to the previous P_s (substitute answers). Specifically, We use the following four metrics to compare the normalized responses with the normalized answers to evaluate the context-faithfulness of LLMs:

- $P_c(\uparrow)$: Frequency of responses matching the context-faithful answer or its aliases, excluding negations or the original answer. Context-faithful answers are counterfactual answers derived from the context.
- $P_o(\downarrow)$: Frequency of responses matching the original factual answer or its aliases.
- $M_R(\downarrow)$: Proportion of responses predicting the correct answer but reluctant to update their predictions, calculated as $M_R = \frac{P_o}{P_s+P_o}$.
- $EM(\uparrow)$: Frequency of responses exactly matching the context-faithful answer.

2.4 Evaluation on ConFiQA

We use our ConFiQA to evaluate the context-faithfulness of popular open-source models (*Llama2-7B-chat*, *Llama2-13B-chat*, *Mistral-7B-instruct-v0.2*, *Qwen2-7B-instruct*) and close-source models (*ChatGPT-4*, *Gemini-1.5-pro*, *ChatGPT-4o*). The experimental results, presented in Table 2, reveal the following key findings:

- Despite alignment efforts, such as instruct-tuning, to meet human standards, the tested LLMs exhibit significant deficiencies in context-faithfulness. Most models have an M_R exceeding 50%, particularly the latest ones, indicating that they tend to rely on their own judgments over the provided context.
- A counterintuitive trend is observed: as model size increases (e.g., *Llama2-chat* from 7B to 13B) or as models become more advanced (e.g., the latest Llama3-8B-instruct compared to earlier versions like *Llama2-7B-chat*), their context-faithfulness tends to decline.

These findings indicate that current LLMs generally exhibit poor alignment in context-faithfulness. Furthermore, with advancements in data processing and model training, more advanced models tend to become increasingly confident in their parametric knowledge, resulting in worse context-faithfulness when facing conflicts between contextual and parametric knowledge. This poses significant challenges for tasks that require strict adherence to

[3]Entities and relations in the sampled path are not repeated

external knowledge, such as RAG or other specialized, closed-domain applications.

3 Context-DPO: Context-Faithful Direct Preference Optimization

Based on the unsatisfactory performance of existing LLMs in context-faithfulness, we argue that it is essential to specifically align LLMs for context-faithfulness. To address this, we propose Context-DPO, the first alignment approach dedicated to enhancing context-faithfulness by creating preference data and aligning LLMs with DPO. The framework of our Context-DPO is shown in Figure 2.

3.1 Preference Data Generation

Leveraging the counterfactual and factual data provided by ConFiQA, we can construct preference data $\mathcal{D} = (x, y_w, y_l)$ efficiently. Specifically, the input x is formed by concatenating the counterfactual context $\mathcal{C}^c$ and the question $\mathcal{Q}$. To generate the counterfactual reasoning chain, each triple in the counterfactual path is transformed into a textual description using a statement template (Table 7) and sequentially concatenated. Finally, the reasoning chain concludes by summarizing the reasoning process to derive the counterfactual answer, which is determined based on the last tail entity in the chain. This process yields faithful responses y_w grounded in the counterfactual context. Similarly, stubborn responses y_l, grounded in factual reality, are constructed by following the original factual path. This approach to constructing the preference dataset $\mathcal{D}$ simulates the reasoning pattern observed in real-world RAG tasks and mirrors the chain-of-thought (Wei et al., 2022) process of LLMs when producing final answers.

3.2 Context-Faithful Alignment with DPO

We leverage the generated preference data to perform alignment tuning on LLMs for context-faithfulness. While several frameworks exist for alignment training, including the widely adopted RLHF framework, which involves training a reward model on preference data and optimizing the policy using the Proximal Policy Optimization (PPO) algorithm, we employ DPO for context-faithful alignment. DPO, as a recent approach to preference optimization, enables the policy π_θ to be learned directly from a fixed preference dataset without requiring an explicit reward model or sampling from the policy during training, as is necessary with PPO. Specifically, our Context-DPO uses the standard cross-entropy objective, and its training objective is formulated as follows:

$$\mathcal{L}_{cf} = -E_{(x,y_w,y_l)\sim\mathcal{D}}\left[\log\sigma\left(\beta\log\frac{\pi_\theta\left(y_w \mid x\right)}{\pi_{\text{ref}}\left(y_w \mid x\right)} - \beta\log\frac{\pi_\theta\left(y_l \mid x\right)}{\pi_{\text{ref}}\left(y_l \mid x\right)}\right)\right], \tag{1}$$

In this formulation, the model policy π_θ is initialized using the base reference policy π_{ref}. The parameter β regulates the extent of divergence from π_{ref}, while σ represents the logistic function.

4 Experiments

4.1 Experimental Setup

Tasks. We evaluate context-faithfulness using the following two tasks: *Retrieval Following* and *Instruction Following*. For *Retrieval Following*, we adopt the setup described in Section 2.4 to assess faithfulness to retrieved passages containing noise and relevant counterfactuals. In contrast, *Instruction Following* focuses solely on textual editing instructions, testing whether LLMs can effectively adhere to user commands.

Datasets. We conduct experiments for *Retrieval Following* using both our ConFiQA and Natural Questions (Kwiatkowski et al., 2019). In Natural Questions, the context is modified to support counterfactual answers following by Longpre et al. (2021). For the *Instruction Following* task, we utilize the MQuAKE dataset (Zhong et al., 2023), which provides multi-hop questions and in-context editing instructions to assess context-faithfulness in response to counterfactual edits.

Models and Baselines. We use current popular open-source LLMs (*Llama2-7B-chat*, *Llama2-13B-chat*, *Mistral-7B-instruct-v0.2*, and *Qwen2-7B-instruct*) as the base models for our experiments. For the *Retrieval Following* task, we use two prompt-based baselines: the attributed prompt (Attr) and the combination of opinion-based and instruction-based prompts (O&I) (Zhou et al., 2023). Additionally, we also fine-tune the LLMs using faithful responses from ConFiQA as the training-based baseline (SFT). For the *Instruction Following* task, we follow the approach of IKE (Zheng et al., 2023), which evaluates the in-context editing capabilities of both the base model

Model	Method	*QA*				*MR*				*MC*			
		$P_c(\uparrow)$	$P_o(\downarrow)$	$M_R(\downarrow)$	$EM(\uparrow)$	$P_c(\uparrow)$	$P_o(\downarrow)$	$M_R(\downarrow)$	$EM(\uparrow)$	$P_c(\uparrow)$	$P_o(\downarrow)$	$M_R(\downarrow)$	$EM(\uparrow)$
LLAMA2-7B-CHAT	Base	61.5	25.6	29.4	0.5	45.4	26.8	37.1	0.3	38.8	24.7	38.5	0.1
	Attr	72.0	14.7	16.9	1.3	45.3	27.3	37.6	0.1	33.7	25.0	43.6	0.1
	O&I	77.3	13.3	14.7	54.7	52.0	16.0	23.5	29.3	50.7	16.0	24.0	29.3
	SFT	63.2	25.2	28.5	0.4	46.7	25.5	35.3	0.4	39.3	24.5	38.4	0.1
	Ours	**92.3**	**3.3**	**3.5**	**64.7**	**54.3**	**11.3**	**17.3**	**32.7**	**52.7**	**12.3**	**19.0**	**32.0**
LLAMA3-8B-INSTRUCT	Base	35.8	44.3	55.3	0.0	30.6	44.1	59.1	0.0	21.7	33.4	60.7	0.0
	Attr	25.7	41.0	61.5	0.0	20.0	46.3	69.9	0.0	9.7	31.3	76.4	0.0
	O&I	32.7	30.3	48.2	1.3	24.0	30.7	54.1	0.7	13.0	25.0	65.8	0.3
	SFT	36.9	42.2	53.3	0.0	32.7	42.5	56.5	0.0	23.5	30.1	56.2	0.0
	Ours	**69.7**	**12.7**	**15.4**	**39.7**	**54.6**	**21.3**	**28.1**	**16.0**	**48.9**	**18.6**	**28.1**	**15.0**
MISTRAL-7B-INSTRUCT	Base	39.3	40.5	50.8	0.3	21.7	37.9	63.5	0.2	14.1	29.8	67.9	0.0
	Attr	44.4	30.3	40.6	2.0	24.9	34.7	58.2	0.7	13.0	28.7	68.8	0.7
	O&I	60.3	20.0	24.3	9.3	33.0	31.0	48.4	5.3	26.7	22.7	46.0	3.3
	SFT	39.4	40.1	50.4	0.5	22.1	37.7	63.1	0.6	13.8	27.5	66.5	0.0
	Ours	**78.6**	**11.0**	**12.3**	**10.7**	**48.7**	**20.3**	**29.5**	**11.5**	**46.7**	**18.3**	**28.1**	**9.6**
QWEN2-7B-INSTRUCT	Base	24.0	43.3	65.4	0.0	21.7	48.7	69.2	0.0	10.0	43.7	81.4	0.0
	Attr	38.3	35.0	47.7	0.3	26.3	40.0	60.3	0.0	13.7	30.7	69.2	0.0
	O&I	58.0	20.3	26.0	5.0	43.3	43.7	43.0	4.7	31.0	26.0	45.6	5.3
	SFT	24.8	42.7	63.3	0.0	21.9	48.0	68.7	0.0	11.5	42.5	78.7	0.0
	Ours	**74.3**	**11.6**	**13.5**	**19.7**	**61.2**	**20.9**	**24.5**	**27.7**	**54.9**	**21.3**	**27.9**	**21.9**

Table 3: Performance results of the *Retrieval Following* task on the ConFiQA benchmark. The best context-faithful result is highlighted in **bold**. Models aligned with our Context-DPO consistently achieve the best performance.

Model	Method	$P_s(\uparrow)$	$P_o(\downarrow)$	$M_R(\downarrow)$	$EM(\uparrow)$
LLAMA2-7B-CHAT	Base	50.8	40.9	44.6	1.3
	Attr	66.2	23.8	26.4	4.7
	O&I	77.8	13.9	15.1	13.7
	SFT	51.5	40.4	43.9	2.2
	Ours	**88.9**	**1.4**	**1.3**	**53.3**
LLAMA3-8B-INSTRUCT	Base	60.7	69.5	53.4	0.0
	Attr	86.3	55.9	39.3	0.1
	O&I	87.4	26.0	22.9	1.4
	SFT	61.8	66.5	51.8	0.0
	Ours	**98.4**	**8.5**	**7.9**	**3.1**
MISTRAL-7B-INSTRUCT	Base	54.5	56.8	51.0	0.1
	Attr	74.3	35.6	32.4	0.2
	O&I	85.0	19.9	18.8	1.8
	SFT	56.2	55.3	49.6	0.1
	Ours	**94.7**	**11.9**	**11.2**	**2.7**
QWEN2–7B-INSTRUCT	Base	54.8	56.3	50.7	0.0
	Attr	75.8	36.7	32.6	0.0
	O&I	87.0	25.5	22.7	1.6
	SFT	55.7	55.1	49.7	0.0
	Ours	**92.5**	**13.2**	**12.4**	**1.8**

Table 4: *Retrieval Following* on Natural Questions

and the Context-DPO-aligned model through contextual editing demonstrations. Detailed implementation and prompt templates for these baselines can be found in Appendix D.

4.2 Performance on *Retrieval Following*

Experimental results for *Retrieval Following* are shown in Table 3 and Table 4 on our ConFiQA and Natural Questions datasets, respectively. Models aligned with our Context-DPO method significantly outperform all baselines, without requiring any additional prompts. On all tasks in ConFiQA, *Llama2-7B-chat*, *Llama2-13B-chat*, *Mistral-7B-instruct-v0.2*, and *Qwen2-7B-instruct* show average improvements of 35.2%, 78.3%, 151.8%, and 280.1%, respectively, in P_c after alignment with our Context-DPO, compared to their initial models. On the Natural Questions dataset, where knowledge conflicts are less pronounced, the accuracy of our method reaches over 93% on average.

This demonstrates that our Context-DPO method is highly effective in significantly improving the context-faithfulness of LLMs. Notably, our approach enhances the model's fundamental context-faithfulness capability through alignment tuning, without relying on inference-stage enhancement methods used by the baselines. This indicates that aligned models have considerable potential for further improvement. Furthermore, the results reveal that simply applying end-to-end SFT is insufficient to effectively enhance the context-faithfulness of LLMs, often performing worse than prompt-based methods. This limitation arises because SFT fails to generalize the training objective of improving context-faithfulness. In contrast, DPO proves to be an effective alternative, as it captures the training signal for context-faithfulness more robustly

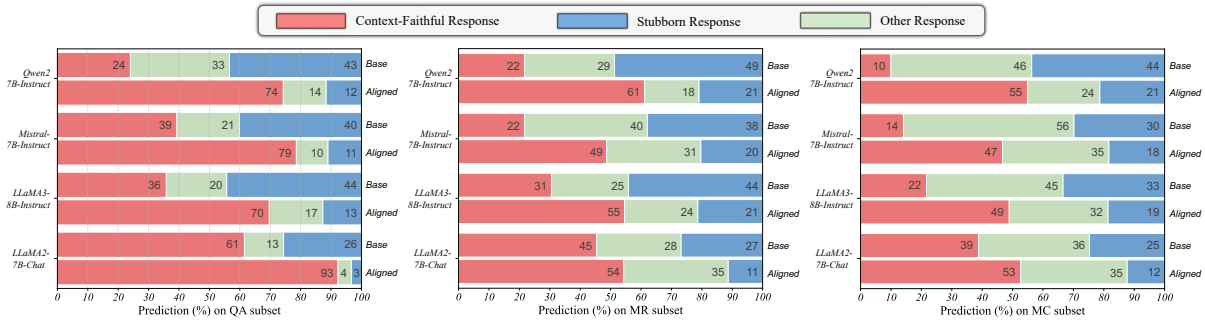

Figure 3: Visualization of LLMs' context-faithfulness across different tasks in the ConFiQA benchmark.

Model	Method	*1-shot*	*3-shot*	*5-shot*
LLaMA2-7B-Chat	Base	67.3	68.1	74.5
	Ours	**72.6**	**76.5**	**80.2**
LLaMA3-8B-Instruct	Base	56.8	63.8	67.9
	Ours	**71.3**	**79.7**	**82.5**
Mistral-7B-Instruct	Base	53.6	55.9	58.4
	Ours	**74.2**	**77.8**	**82.3**
Qwen2-7B-Instruct	Base	61.0	64.2	71.3
	Ours	**74.4**	**78.6**	**81.8**

Table 5: Performance results on *Instruction Following*.

Model	Method	*MC1*	*MC2*	*MC3*
LLaMA2-7B-Chat	Base	32.90	50.29	24.04
	Ours	31.72	48.13	23.38
LLaMA3-8B-Instruct	Base	40.76	59.36	31.79
	Ours	41.37	59.80	30.93
Mistral-7B-Instruct	Base	54.69	69.92	39.50
	Ours	52.67	68.05	38.93
Qwen2-7B-Instruct	Base	42.84	61.11	32.74
	Ours	42.98	61.79	32.54

Table 6: Performence of LLM factual generation on TruthfulQA. The factuality of the generated responses remains largely unchanged before and after alignment.

through preference pair comparisons.

4.3 Performance on *Instruction Following*

We evaluate LLMs' *Instruction Following* ability with the in-context editing task on MQuAKE dataset, where instruction-based textual prompts are used to guide the models in editing relevant knowledge to answer questions. The context demonstrations we used are provided in Appendix C. Table 5 presents the few-shot accuracy of models, both before and after alignment with our Context DPO, under varying numbers of demonstration prompts. While increasing the number of context demonstrations encourages LLMs to better follow the editing instructions, the aligned models consistently outperform the baselines. This demonstrates that the context-faithfulness alignment based on our ConFiQA, which simulates question-answering according to the retrieved passage, also enhances the model's faithfulness to user instructions.

4.4 Validation of the Decoupled Improvement in LLMs' Context-Faithfulness

As mentioned by Bi et al. (2024b), there may be a trade-off between context-faithfulness and factuality in LLMs. To validate this for our method, we evaluate whether the Context-DPO alignment affects the model's factual generation ability. Using TruthfulQA (Lin et al., 2021), we employ a multiple-choice task where the LLM selects an answer from a range of correct and incorrect options, evaluated by multiple-choice accuracy (MC1, MC2, and MC3). As shown in Table 6, the performance of the aligned models fluctuates by no more than 1% on average across the MC metrics, compared to the original models. This indicates that the improvements achieved by our Context-DPO alignment are decoupled: while enhancing context-faithfulness, the alignment does not negatively impact the model's inherent generation ability when no context is provided. Therefore, we strongly advocate for incorporating context-faithfulness alignment as a standard practice in LLM alignment.

5 In-depth Exploration of the Metamorphosis in Context-Faithfulness

Figure 3 provides an intuitive visualization of the impact of our Context-DPO on LLMs' context-faithfulness, demonstrating its ability to reduce irrelevant responses (other response) and stubborn reliance on parametric knowledge (stubborn response), ultimately leading to more context-faithful answers (context-faithful response). To further investigate the internal mechanisms behind the effec-

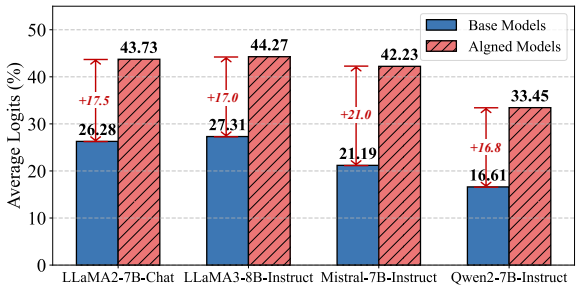

Figure 4: Average logits (%) of key tokens faithful to contextual knowledge, comparing base models and models aligned using our Context-DPO.

tive alignment of LLMs' context-faithfulness by our Context-DPO, we utilize the knowledge token capturing algorithm proposed by Bi et al. (2024b). for deeper exploration. The algorithm (detailed in Appendix 1) identifies the tokens with the highest probability of distinguishing between contextual knowledge and parametric knowledge by matching decoded tokens with their corresponding knowledge strings. Following the *Instruction Following* task (4.3), we collected 2,000 question-answer instances from the MQuAKE dataset to capture the logits distribution of key tokens, which effectively highlights the distinction between context-faithful responses and stubborn responses.

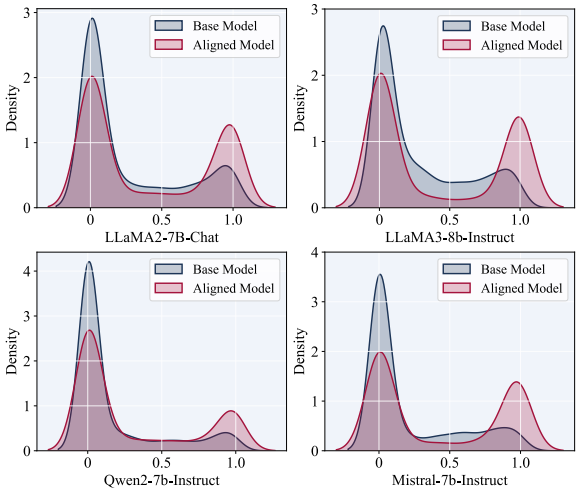

Figure 5: Kernel density estimation of the softmax probability distribution for context-faithful tokens.

We calculate the average logits of key tokens representing context-faithfulness, with the results shown in Figure 4. Aligned models exhibit significant improvements over the base models, with gains ranging from 16.8 to 21.0, indicating that our Context-DPO effectively increases the probability of generating context-faithful responses. We further analyze the softmax-transformed logits distribution of these tokens, as shown in Figure 5. The results indicate that models aligned with Context-DPO reduce the distribution in low-probability regions while increasing it in high-probability regions compared to their original versions. This adjustment further increases the likelihood of decoding context-faithful tokens at key positions, leading to a significant rise in the generation frequency of top-ranked tokens, as illustrated in Figure 6. Our interpretability analysis uncovers the internal mechanisms behind the effective context-faithfulness alignment achieved by our Context-DPO. This highlights its ability to significantly enhance the upper bound of context-faithfulness without relying on external inference-stage methods.

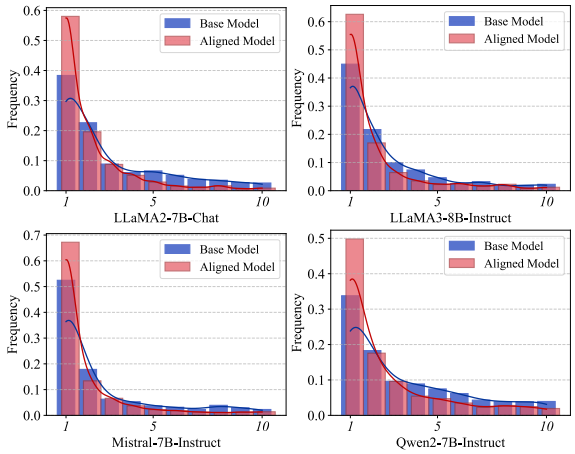

Figure 6: Ranking distribution of context-faithful tokens in the token vocabulary. Aligned models exhibit a significant increase in the frequency of top-ranked context-faithful tokens compared to base models.

6 Conclusion

In this work, we introduce ConFiQA, a novel benchmark that simulates real-world RAG scenarios and knowledge conflicts, enabling the evaluation of LLMs' context-faithfulness. To address shortcomings in context-faithfaulness for current models, we propose Context-DPO, the first alignment method dedicated to enhancing context-faithfulness. This approach leverages ConFiQA to construct preference data and fine-tunes models using DPO. Experimental results demonstrate that Context-DPO significantly enhances the context-faithfulness of popular LLMs without compromising their inherent generative capabilities. Furthermore, interpretability analysis reveals the mechanisms underlying the improvements in faithfulness. Our work paves the way to develop both effective and accountable context-faithfulness for LLMs.

Limitations

This paper focuses on specific knowledge conflict scenarios to better highlight context-faithfulness in evaluation. However, its application in typical real-world RAG scenarios has not been extensively validated. We believe that our Context-DPO can also bring significant benefits to standard RAG tasks, and we plan to explore this further in future work. Additionally, although our findings indicate in experiments that context-faithfulness tends to decline as model size increases and training becomes more refined, further extensive experiments are needed to fully validate this observation.

Ethical Considerations

Ethical considerations are paramount in our research. The proposed dataset, along with the open-source datasets and widely recognized models used in this study, strictly adheres to established ethical principles. Additionally, counterfactual data is employed in our experimental evaluations to measure context-faithfulness under knowledge conflict scenarios. The proposed methods are designed to ensure that models do not generate harmful or misleading information. Throughout this research, we remain committed to upholding ethical standards, prioritizing transparency, and fostering the responsible use of technology to benefit society.

References

Baolong Bi, Shenghua Liu, Lingrui Mei, Yiwei Wang, Pengliang Ji, and Xueqi Cheng. 2024a. Decoding by contrasting knowledge: Enhancing llms' confidence on edited facts. *Preprint*, arXiv:2405.11613.

Baolong Bi, Shenghua Liu, Yiwei Wang, Lingrui Mei, Junfeng Fang, Hongcheng Gao, Shiyu Ni, and Xueqi Cheng. 2024b. Is factuality enhancement a free lunch for llms? better factuality can lead to worse context-faithfulness. *Authorea Preprints*.

Baolong Bi, Shenghua Liu, Yiwei Wang, Lingrui Mei, Hongcheng Gao, Junfeng Fang, and Xueqi Cheng. 2024c. Struedit: Structured outputs enable the fast and accurate knowledge editing for large language models. *arXiv preprint arXiv:2409.10132*.

Baolong Bi, Shenghua Liu, Yiwei Wang, Lingrui Mei, Hongcheng Gao, Yilong Xu, and Xueqi Cheng. 2024d. Adaptive token biaser: Knowledge editing via biasing key entities. *arXiv preprint arXiv:2406.12468*.

Bochuan Cao, Yuanpu Cao, Lu Lin, and Jinghui Chen. 2023. Defending against alignment-breaking attacks via robustly aligned llm. *arXiv preprint arXiv:2309.14348*.

Hung-Ting Chen, Michael JQ Zhang, and Eunsol Choi. 2022. Rich knowledge sources bring complex knowledge conflicts: Recalibrating models to reflect conflicting evidence. *arXiv preprint arXiv:2210.13701*.

Roi Cohen, Eden Biran, Ori Yoran, Amir Globerson, and Mor Geva. 2024. Evaluating the ripple effects of knowledge editing in language models. *Transactions of the Association for Computational Linguistics*, 12:283–298.

Zhangyin Feng, Xiaocheng Feng, Dezhi Zhao, Maojin Yang, and Bing Qin. 2024. Retrieval-generation synergy augmented large language models. In *ICASSP 2024-2024 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 11661–11665. IEEE.

Sophie Forgan. 2005. Building the museum: Knowledge, conflict, and the power of place. *Isis*, 96(4):572–585.

Yunfan Gao, Yun Xiong, Xinyu Gao, Kangxiang Jia, Jinliu Pan, Yuxi Bi, Yi Dai, Jiawei Sun, and Haofen Wang. 2023. Retrieval-augmented generation for large language models: A survey. *arXiv preprint arXiv:2312.10997*.

Anisha Gunjal, Jihan Yin, and Erhan Bas. 2024. Detecting and preventing hallucinations in large vision language models. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 18135–18143.

Kelvin Guu, Kenton Lee, Zora Tung, Panupong Pasupat, and Ming-Wei Chang. 2020. Realm: Retrieval-augmented language model pre-training. *Preprint*, arXiv:2002.08909.

Linmei Hu, Zeyi Liu, Ziwang Zhao, Lei Hou, Liqiang Nie, and Juanzi Li. 2023. A survey of knowledge enhanced pre-trained language models. *IEEE Transactions on Knowledge and Data Engineering*.

Baixiang Huang, Canyu Chen, Xiongxiao Xu, Ali Payani, and Kai Shu. 2024. Can knowledge editing really correct hallucinations? *arXiv preprint arXiv:2410.16251*.

Lei Huang, Weijiang Yu, Weitao Ma, Weihong Zhong, Zhangyin Feng, Haotian Wang, Qianglong Chen, Weihua Peng, Xiaocheng Feng, Bing Qin, and Ting Liu. 2023a. A survey on hallucination in large language models: Principles, taxonomy, challenges, and open questions. *Preprint*, arXiv:2311.05232.

Yuheng Huang, Jiayang Song, Zhijie Wang, Huaming Chen, and Lei Ma. 2023b. Look before you leap: An exploratory study of uncertainty measurement for large language models. *arXiv preprint arXiv:2307.10236*.

Gautier Izacard and Edouard Grave. 2021. Leveraging passage retrieval with generative models for open domain question answering. *Preprint*, arXiv:2007.01282.

Zhengbao Jiang, Frank F Xu, Luyu Gao, Zhiqing Sun, Qian Liu, Jane Dwivedi-Yu, Yiming Yang, Jamie Callan, and Graham Neubig. 2023. Active retrieval augmented generation. *arXiv preprint arXiv:2305.06983*.

Jean Kaddour, Joshua Harris, Maximilian Mozes, Herbie Bradley, Roberta Raileanu, and Robert McHardy. 2023. Challenges and applications of large language models. *arXiv preprint arXiv:2307.10169*.

Tom Kwiatkowski, Jennimaria Palomaki, Olivia Redfield, Michael Collins, Ankur Parikh, Chris Alberti, Danielle Epstein, Illia Polosukhin, Jacob Devlin, Kenton Lee, et al. 2019. Natural questions: a benchmark for question answering research. *Transactions of the Association for Computational Linguistics*, 7:453–466.

Patrick Lewis, Ethan Perez, Aleksandra Piktus, Fabio Petroni, Vladimir Karpukhin, Naman Goyal, Heinrich Küttler, Mike Lewis, Wen-tau Yih, Tim Rocktäschel, et al. 2020. Retrieval-augmented generation for knowledge-intensive nlp tasks. *Advances in Neural Information Processing Systems*, 33:9459–9474.

Yuepei Li, Kang Zhou, Qiao Qiao, Bach Nguyen, Qing Wang, and Qi Li. 2024. Investigating context-faithfulness in large language models: The roles of memory strength and evidence style. *arXiv preprint arXiv:2409.10955*.

Stephanie Lin, Jacob Hilton, and Owain Evans. 2021. Truthfulqa: Measuring how models mimic human falsehoods. *arXiv preprint arXiv:2109.07958*.

Fang Liu, Yang Liu, Lin Shi, Houkun Huang, Ruifeng Wang, Zhen Yang, and Li Zhang. 2024. Exploring and evaluating hallucinations in llm-powered code generation. *arXiv preprint arXiv:2404.00971*.

Yang Liu, Yuanshun Yao, Jean-Francois Ton, Xiaoying Zhang, Ruocheng Guo Hao Cheng, Yegor Klochkov, Muhammad Faaiz Taufiq, and Hang Li. 2023. Trustworthy llms: A survey and guideline for evaluating large language models' alignment. *arXiv preprint arXiv:2308.05374*.

Shayne Longpre, Kartik Perisetla, Anthony Chen, Nikhil Ramesh, Chris DuBois, and Sameer Singh. 2021. Entity-based knowledge conflicts in question answering. *arXiv preprint arXiv:2109.05052*.

Aman Madaan, Niket Tandon, Peter Clark, and Yiming Yang. 2022. Memory-assisted prompt editing to improve gpt-3 after deployment. *arXiv preprint arXiv:2201.06009*.

Lingrui Mei, Shenghua Liu, Yiwei Wang, Baolong Bi, Jiayi Mao, and Xueqi Cheng. 2024. " not aligned" is not" malicious": Being careful about hallucinations of large language models' jailbreak. *arXiv preprint arXiv:2406.11668*.

Kevin Meng, David Bau, Alex Andonian, and Yonatan Belinkov. 2022a. Locating and editing factual associations in gpt. *Advances in Neural Information Processing Systems*, 35:17359–17372.

Kevin Meng, Arnab Sen Sharma, Alex Andonian, Yonatan Belinkov, and David Bau. 2022b. Mass-editing memory in a transformer. *arXiv preprint arXiv:2210.07229*.

Yifei Ming, Senthil Purushwalkam, Shrey Pandit, Zixuan Ke, Xuan-Phi Nguyen, Caiming Xiong, and Shafiq Joty. 2024. Faitheval: Can your language model stay faithful to context, even if" the moon is made of marshmallows". *arXiv preprint arXiv:2410.03727*.

Reiichiro Nakano, Jacob Hilton, Suchir Balaji, Jeff Wu, Long Ouyang, Christina Kim, Christopher Hesse, Shantanu Jain, Vineet Kosaraju, William Saunders, Xu Jiang, Karl Cobbe, Tyna Eloundou, Gretchen Krueger, Kevin Button, Matthew Knight, Benjamin Chess, and John Schulman. 2022. Webgpt: Browser-assisted question-answering with human feedback. *Preprint*, arXiv:2112.09332.

OpenAI. 2022. large-scale generative pre-training model for conversation. *OpenAI blog*.

OpenAI. 2023. Gpt-4 technical report. *Preprint*, arXiv:2303.08774.

Shirui Pan, Linhao Luo, Yufei Wang, Chen Chen, Jiapu Wang, and Xindong Wu. 2024. Unifying large language models and knowledge graphs: A roadmap. *IEEE Transactions on Knowledge and Data Engineering*.

Fabio Petroni, Patrick Lewis, Aleksandra Piktus, Tim Rocktäschel, Yuxiang Wu, Alexander H Miller, and Sebastian Riedel. 2020. How context affects language models' factual predictions. *arXiv preprint arXiv:2005.04611*.

Yujia Qin, Shengding Hu, Yankai Lin, Weize Chen, Ning Ding, Ganqu Cui, Zheni Zeng, Yufei Huang, Chaojun Xiao, Chi Han, Yi Ren Fung, Yusheng Su, Huadong Wang, Cheng Qian, Runchu Tian, Kunlun Zhu, Shihao Liang, Xingyu Shen, Bokai Xu, Zhen Zhang, Yining Ye, Bowen Li, Ziwei Tang, Jing Yi, Yuzhang Zhu, Zhenning Dai, Lan Yan, Xin Cong, Yaxi Lu, Weilin Zhao, Yuxiang Huang, Junxi Yan, Xu Han, Xian Sun, Dahai Li, Jason Phang, Cheng Yang, Tongshuang Wu, Heng Ji, Zhiyuan Liu, and Maosong Sun. 2024. Tool learning with foundation models. *Preprint*, arXiv:2304.08354.

Rafael Rafailov, Archit Sharma, Eric Mitchell, Christopher D Manning, Stefano Ermon, and Chelsea Finn. 2024. Direct preference optimization: Your language model is secretly a reward model. *Advances in Neural Information Processing Systems*, 36.

Tianhao Shen, Renren Jin, Yufei Huang, Chuang Liu, Weilong Dong, Zishan Guo, Xinwei Wu, Yan Liu, and Deyi Xiong. 2023. Large language model alignment: A survey. *arXiv preprint arXiv:2309.15025*.

Weijia Shi, Xiaochuang Han, Mike Lewis, Yulia Tsvetkov, Luke Zettlemoyer, and Scott Wen-tau Yih. 2023. Trusting your evidence: Hallucinate less with context-aware decoding. *arXiv preprint arXiv:2305.14739*.

Chenglei Si, Zhe Gan, Zhengyuan Yang, Shuohang Wang, Jianfeng Wang, Jordan Boyd-Graber, and Lijuan Wang. 2023. Prompting gpt-3 to be reliable. *Preprint*, arXiv:2210.09150.

Katherine Tian, Eric Mitchell, Huaxiu Yao, Christopher D Manning, and Chelsea Finn. 2023. Fine-tuning language models for factuality. *arXiv preprint arXiv:2311.08401*.

S. M Towhidul Islam Tonmoy, S M Mehedi Zaman, Vinija Jain, Anku Rani, Vipula Rawte, Aman Chadha, and Amitava Das. 2024. A comprehensive survey of hallucination mitigation techniques in large language models. *Preprint*, arXiv:2401.01313.

Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, Aurélien Rodriguez, Armand Joulin, Edouard Grave, and Guillaume Lample. 2023a. Llama: Open and efficient foundation language models. *CoRR*, abs/2302.13971.

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, Brian Fuller, Cynthia Gao, Vedanuj Goswami, Naman Goyal, Anthony Hartshorn, Saghar Hosseini, Rui Hou, Hakan Inan, Marcin Kardas, Viktor Kerkez, Madian Khabsa, Isabel Kloumann, Artem Korenev, Punit Singh Koura, Marie-Anne Lachaux, Thibaut Lavril, Jenya Lee, Diana Liskovich, Yinghai Lu, Yuning Mao, Xavier Martinet, Todor Mihaylov, Pushkar Mishra, Igor Molybog, Yixin Nie, Andrew Poulton, Jeremy Reizenstein, Rashi Rungta, Kalyan Saladi, Alan Schelten, Ruan Silva, Eric Michael Smith, Ranjan Subramanian, Xiaoqing Ellen Tan, Binh Tang, Ross Taylor, Adina Williams, Jian Xiang Kuan, Puxin Xu, Zheng Yan, Iliyan Zarov, Yuchen Zhang, Angela Fan, Melanie Kambadur, Sharan Narang, Aurelien Rodriguez, Robert Stojnic, Sergey Edunov, and Thomas Scialom. 2023b. Llama 2: Open foundation and fine-tuned chat models. *Preprint*, arXiv:2307.09288.

Denny Vrandečić and Markus Krötzsch. 2014. Wikidata: a free collaborative knowledgebase. *Communications of the ACM*, 57(10):78–85.

Cunxiang Wang, Xiaoze Liu, Yuanhao Yue, Xiangru Tang, Tianhang Zhang, Cheng Jiayang, Yunzhi Yao, Wenyang Gao, Xuming Hu, Zehan Qi, et al. 2023. Survey on factuality in large language models: Knowledge, retrieval and domain-specificity. *arXiv preprint arXiv:2310.07521*.

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Fei Xia, Ed Chi, Quoc V Le, Denny Zhou, et al. 2022. Chain-of-thought prompting elicits reasoning in large language models. *Advances in neural information processing systems*, 35:24824–24837.

Jian Xie, Kai Zhang, Jiangjie Chen, Renze Lou, and Yu Su. 2024. Adaptive chameleon or stubborn sloth: Revealing the behavior of large language models in knowledge conflicts. *Preprint*, arXiv:2305.13300.

Rongwu Xu, Zehan Qi, Zhijiang Guo, Cunxiang Wang, Hongru Wang, Yue Zhang, and Wei Xu. 2024. Knowledge conflicts for llms: A survey. *arXiv preprint arXiv:2403.08319*.

Rui Yang, Haoran Liu, Qingcheng Zeng, Yu He Ke, Wanxin Li, Lechao Cheng, Qingyu Chen, James Caverlee, Yutaka Matsuo, and Irene Li. 2024. Kg-rank: Enhancing large language models for medical qa with knowledge graphs and ranking techniques. *arXiv preprint arXiv:2403.05881*.

Shunyu Yao, Jeffrey Zhao, Dian Yu, Nan Du, Izhak Shafran, Karthik Narasimhan, and Yuan Cao. 2023a. React: Synergizing reasoning and acting in language models. *Preprint*, arXiv:2210.03629.

Yunzhi Yao, Peng Wang, Bozhong Tian, Siyuan Cheng, Zhoubo Li, Shumin Deng, Huajun Chen, and Ningyu Zhang. 2023b. Editing large language models: Problems, methods, and opportunities. *arXiv preprint arXiv:2305.13172*.

Duzhen Zhang, Yahan Yu, Chenxing Li, Jiahua Dong, Dan Su, Chenhui Chu, and Dong Yu. 2024. Mm-llms: Recent advances in multimodal large language models. *arXiv preprint arXiv:2401.13601*.

Ce Zheng, Lei Li, Qingxiu Dong, Yuxuan Fan, Zhiyong Wu, Jingjing Xu, and Baobao Chang. 2023. Can we edit factual knowledge by in-context learning? *arXiv preprint arXiv:2305.12740*.

Zexuan Zhong, Tao Lei, and Danqi Chen. 2022. Training language models with memory augmentation. *Preprint*, arXiv:2205.12674.

Zexuan Zhong, Zhengxuan Wu, Christopher D Manning, Christopher Potts, and Danqi Chen. 2023. Mquake: Assessing knowledge editing in language models via multi-hop questions. *arXiv preprint arXiv:2305.14795*.

Wenxuan Zhou, Sheng Zhang, Hoifung Poon, and Muhao Chen. 2023. Context-faithful prompting for large language models. *arXiv preprint arXiv:2303.11315*.

Chen Zhu, Ankit Singh Rawat, Manzil Zaheer, Srinadh Bhojanapalli, Daliang Li, Felix Yu, and Sanjiv Kumar. 2020. Modifying memories in transformer models. *arXiv preprint arXiv:2012.00363*.

A Related Work

Hallucinations in LLMs The outputs of large language models (LLMs) often appear plausible at first glance but may exhibit various issues upon closer inspection, a phenomenon commonly referred to as hallucinations (Kaddour et al., 2023; Tonmoy et al., 2024; Wang et al., 2023; Mei et al., 2024). These hallucinations cause LLMs to produce content that deviates from user inputs, previously generated context, or factual knowledge, severely undermining their reliability in real-world applications (Gunjal et al., 2024; Huang et al., 2023a; Liu et al., 2024; Zhang et al., 2024). Such hallucinations can arise at different stages of the LLM lifecycle. Broadly, research on hallucination mitigation falls into two categories. During the training phase, studies such as Hu et al. (2023); Pan et al. (2024) have explored methods like training data curation and knowledge grounding to better integrate external knowledge into the model. Recent findings suggest that hallucinations often stem from conflicts between an LLM's internal parameters and the external context provided during inference. In the inference stage, recent works have proposed methods such as confidence estimation (Huang et al., 2023b), knowledge retrieval (Feng et al., 2024; Yang et al., 2024), and knowledge editing (KE) (Yao et al., 2023b) to generate more accurate outputs. These approaches aim to refine the model's predictions by enhancing its ability to validate outputs or supplementing it with relevant external knowledge. Despite these advancements, addressing hallucinations remains a critical challenge for improving LLM reliability.

Knowledge Conflicts Knowledge conflicts (Forgan, 2005; Xu et al., 2024) can be categorized into three types: internal conflicts within the context, conflicts between the memories encoded in model parameters, and conflicts between the context and model parameters. The latter, as a critical issue, has been extensively studied to mitigate hallucinations. Various popular tools (Nakano et al., 2022; Yao et al., 2023a; Qin et al., 2024) and retrieval-augmented methods (Guu et al., 2020; Izacard and Grave, 2021; Zhong et al., 2022), such as ChatGPT plugins and New Bing, have been introduced as effective strategies for providing external knowledge evidence. However, integrating external knowledge is not without challenges, as it sometimes conflicts with the parametric knowledge of LLMs (Si et al., 2023; Xie et al., 2024), resulting in inconsistent or unreliable outputs, especially when LLMs exhibit overconfidence in their inherent parametric knowledge. These conflicts between external sources and the internal knowledge stored within LLMs continue to pose significant challenges in ensuring reliable model performance.

Retrieval-Augmented Generation RAG (Lewis et al., 2020; Gao et al., 2023) enhances LLMs by retrieving relevant document chunks from external knowledge bases based on semantic similarity. By leveraging external knowledge, RAG effectively reduces the generation of factually incorrect content, addressing a key challenge in LLM outputs. Its integration with LLMs has led to widespread adoption, significantly improving the reliability of LLM-based systems (Jiang et al., 2023). However, context-faithfulness (Chen et al., 2022; Li et al., 2024) plays a crucial role in determining the performance of RAG, as the retrieved content may conflict with the internal parametric knowledge of LLMs, particularly when the parametric knowledge is insufficient or outdated. This challenge is exacerbated as LLMs grow in size and undergo more refined training, making them increasingly confident in their own parametric knowledge. Such overconfidence further undermines context-faithfulness in scenarios where knowledge conflicts arise.

In-Context Editing As one of the most effective Knowledge Editing (KE) methods (Yao et al., 2023b; Zhu et al., 2020; Meng et al., 2022a,b; Huang et al., 2024), in-context editing (ICE) (Madaan et al., 2022; Zhong et al., 2023; Zheng et al., 2023; Cohen et al., 2024; Bi et al., 2024a,c) has demonstrated state-of-the-art performance in KE. By providing contextual editing prompts enriched with new knowledge retrieved from the edit memory, ICE effectively guides LLMs to perform inference and generate answers aligned with the new knowledge. As part of this study, we use a ICE task with instructional editing prompts to evaluate LLMs' performance in instruction following.

B Details of Data Constructing

One of our key objectives is to construct counterfactual contexts that simulate RAG scenarios under knowledge conflicts. This process involves two steps. The first step is to establish factual statements. We begin by collecting popular entities from Wikipedia and extracting factual triples as-

sociated with these entities from Wikidata. This ensures that the collected facts are widely recognized and likely to be well-represented in the parametric memory of LLMs due to pretraining. Using rule-based transformations, we convert these triples into factual statements, as illustrated by the templates provided in Table 7. Based on a chain of triples, multi-hop questions are generated using the following prompts and the examples in Table 8.

Prompt for Question Generation

You are a sophisticated {hop_num}-hop question generator. Given a chain of Wikidata triples, generate a question that asks about the final tail entity ({tail}) in the chain using only the starting head entity ({head}). Do not include any bridge entities in the question; instead, phrase the question as if directly asking about the relationship from the head entity to the tail entity.

The second step involves generating contexts. Starting with the original factual triples, we expand the descriptions of entities to create enriched contexts that include irrelevant noise unrelated to the questions. The prompts used for generating these contexts are shown in following table:

Prompt for Context Generation

Considering {facts}, generate a brief description of the entity: {head}, approximately 100 words long. Ensure that {tail} is accurately mentioned in the description.

The head and tail are derived from the collected factual triple (head, relation, tail), and the fact is constructed from this triple using a cloze template. Subsequently, the related entities in the context, along with their aliases and all associated morphological forms, are edited to reflect the counterfactuals. This editing process is achieved through pre-mapping the relationships and systematically replacing the corresponding entities.

C *Instruction Following* Task

In our experiments, in addition to the *Retrieval Following* task on ConFiQA and Natural Questions, we specifically design an *Instruction Following* task to evaluate the model's faithfulness to user instructions as context. Specifically, we employ an in-context editing (ICE) task using the MQuAKE dataset to assess this capability. This task provides contextual examples along with knowledge-editing instructions to test whether LLMs follow the provided context to answer questions. The few-shot prompting used for this task includes:

Few-shot Prompting for In-Context Editing

Q: What is the capital city of the country of citizenship of Ivanka Trump's spouse?
E: Jared Kushner is a citizen of Canada
A: Ottawa

Q: On which continent was the director of "My House Husband: Ikaw Na!" educated?
E: Irene Villamor was educated in New York University
A: North America

Q: In which country is the company that created Nissan 200SX located?
E: Nissan is located in the country of China
A: China

Q: Who has ownership of the developer of the Chevrolet Corvette (C4)?
E: Chevrolet is owned by Volkswagen Group
A: Volkswagen Group

Q: [Question]
E: [Edit]
A:

D Implementation of Baselines

We follow the previous setup (Zhou et al., 2023) and utilize two prompt-based baselines: the attributed prompt (Attr) and a combination of opinion-based and instruction-based prompts (O&I). The prompt templates are as follows:

Attr Based Prompt

{*context*} Q: {*question*} based on the given text? A: {*answer*}.

I&O Based Prompt

Bob said "{*context*}" Q: {*question*} in Bob's opinion? A: {*answer*}.

In addition, we provide our own SFT baseline for comparison, which conducts end-to-end training using data in the format of (context + question, faithful response). Experimental results indicate

Algorithm 1 Knowledge Token Capturing

Require: The LLM generates a token sequence of length n, $\mathcal{V}$: vocabulary of LLM, $\mathcal{P}_i$ in ($\mathcal{P}_1$, $\mathcal{P}_2$, ..., $\mathcal{P}_n$): logits distribution of tokens, S_{new}: string of new knowledge related to context.

Ensure: Captured new knowledge logits P_{new}

```
Initialize P_new ← None
S_com = COM(S_new)                                  ▷ Identify common substrings
for P_i in (P_1, P_2, ..., P_n) do
    for token x_j in V do                           ▷ Sort by P_i in descending order
        x_j → x'_j                                  ▷ Decode x_j to string x'_j
        if x'_j in S_com and P_new = None: break    ▷ x'_j is indistinguishable
        if x'_j in S_new and P_new = None: P_new ← P_{i,j}   ▷ Capture new knowledge
    end for
end for
return P_new
```

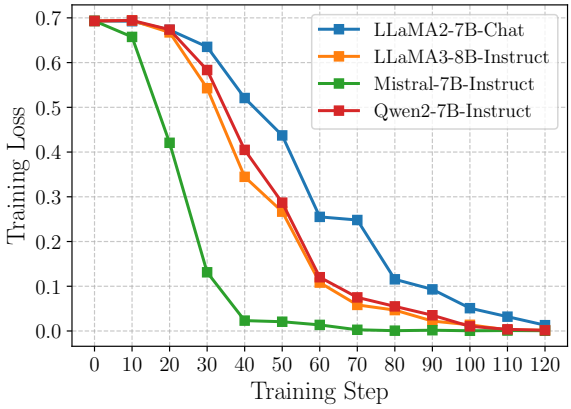

Figure 7: Training loss convergence of different models during our Context-DPO fine-tuning.

that SFT fails to effectively improve the context-faithfulness performance of LLMs.

E Convergence of Context-DPO Training

We configure the training with `batch_size=4` and `gradient_accumulation_steps=8` and perform Direct Preference Optimization based on the constructed preference pairs. The convergence results are illustrated in Figure 7.

F Knowledge Token Capturing

The goal of the algorithm (Bi et al., 2024b) is to identify parts of LLM outputs that distinguish newly acquired knowledge from the context (e.g., counterfactual information) from the parametric knowledge embedded in the LLM, rather than analyzing repetitive or meaningless outputs. For instance, consider an expected LLM output in an *Instruction-Following* scenario with injected context, such as *"A: United States"*, compared to the original parametric output without context injection, which might be *"A: United Kingdom"*. In this case, capturing *"A:"* is unnecessary as it lacks factual significance, and focusing on *"United"* is redundant, as it does not reflect the difference between the outputs. Instead, the focus should be on capturing tokens with distinct factual significance—those that can effectively differentiate between newly introduced contextual knowledge and the model's inherent parametric knowledge. In this example, a token like *"Kingdom"* serves as a critical marker, clearly highlighting the key divergence between contextual information and the model's existing knowledge. The pseudocode of the algorithm is shown in Algorithm 1. It captures the tokens with the highest probability of distinguishing new knowledge from parametric knowledge by matching the decoded tokens with their corresponding knowledge strings.

G Examples of Data in ConFiQA

We provide example templates from the three sub-datasets of our ConFiQA: *QA* (Question-Answering), *MR* (Multi-hop Reasoning), and *MC* (Multi-Conflicts), which are shown in Tables 11, 9, and 10, respectively. We provide a case study of LLaMA2-7B, LLaMA3-8B, Mistral-7B, and Qwen2-7B on the QA task in Appendix H. Here, the green text represents the expected context-faithful output, while the red text represents the stubborn response.

Relation	Description	Cloze-style statement template
P6	head of government	The name of the current head of the [subject] government is [target]
P17	country	[subject] is located in the country of [target]
P26	spouse	[subject] is married to [target]
P27	country of citizenship	[subject] is a citizen of [target]
P30	continent	[subject] is located in the continent of [target]
P35	head of state	The name of the current head of state in [subject] is [target]
P36	capital	The capital of [subject] is [target]
P37	official language	The official language of [subject] is [target]
P38	currency	[subject]'s currency is [target]
P39	position held	[subject] held the position of [target]
P50	author	The author of [subject] is [target]
P54	member of sports team	[subject] is a member of the sports team [target]
P57	director	[subject] was directed by [target]
P86	composer	[subject] was composed by [target]
P101	field of work	[subject]'s field of work is [target]
P103	native language	[subject]'s native language is [target]
P108	employer	[subject] is employed by [target]
P112	founder	[subject] was founded by [target]
P127	owned by	[subject] is owned by [target]
P136	genre	The genre of [subject] is [target]
P1376	capital of	[subject] is the capital of [target]
P140	religion	[subject] is affiliated with the religion of [target]
P155	follows	[subject] follows [target]
P159	headquarters location	The headquarters of [subject] is located in [target]
P166	award received	[subject] received the award [target]
P170	creator	[subject] was created by [target]
P172	ethnic group	[subject]'s ethnic group is [target]
P175	performer	[subject] was performed by [target]
P178	developer	[subject] was developed by [target]
P264	record label	[subject] is under the record label [target]
P276	location	[subject] is located in [target]
P286	head coach	The head coach of [subject] is [target]
P407	language of work or name	[subject] was written in the language [target]
P413	position played	[subject] plays the position of [target]
P463	member of	[subject] is a member of [target]
P488	chairperson	The chairperson of [subject] is [target]
P495	country of origin	[subject] originated from [target]
P641	sport	[subject] is associated with the sport [target]
P800	notable work	[subject] is famous for the work [target]
P937	work location	The work location of [subject] is [target]
P169	chief executive officer	The CEO of [subject] is [target]

Table 7: Cloze-style statement template that are used to construct factual statement.

	Examples of 1-hop questions
$\mathcal{P}^o$	(United States, capital, Washington, D.C.)
$\mathcal{Q}$	What is the capital of the United States?
$\mathcal{P}^o$	(United States, head of government, Joe Biden)
$\mathcal{Q}$	Who is the current head of the United States government?
$\mathcal{P}^o$	(United States, official language, English)
$\mathcal{Q}$	What is the official language of the United States?
	Examples of 2-hop questions
$\mathcal{P}^o$	(Jacques Necker, employer, University of Geneva) (University of Geneva, headquarters location, Geneva)
$\mathcal{Q}$	In which city is the head office located for the company that employed Jacques Necker?
$\mathcal{P}^o$	(Percival Lowell, educated at, Harvard University) (Harvard University, headquarters location, Cambridge)
$\mathcal{Q}$	Where is the headquarters of the educational institution attended by Percival Lowell located?
$\mathcal{P}^o$	(Gordon Moore, country of citizenship, United States of America) (United States of America, capital, Washington, D.C.)
$\mathcal{Q}$	What is the capital of the country where Gordon Moore holds citizenship?
	Examples of 3-hop questions
$\mathcal{P}^f$	(Kim Kardashian, spouse, Kanye West) (Kanye West, genre, hip hop music) (hip hop music, country of origin, United States of America)
$\mathcal{Q}$	Which country is the genre of the partner of Kim Kardashian associated with originally from?
$\mathcal{P}^f$	(Nicholas of Tolentino, religion or worldview, Catholic Church) (Catholic Church, founded by, Jesus Christ) (Jesus Christ, place of birth, Bethlehem)
$\mathcal{Q}$	What is the birthplace of the founder of the religion that Nicholas of Tolentino followed?
$\mathcal{P}^f$	(Boston, head of government, Marty Walsh) (Marty Walsh, educated at, Boston College) (Boston College, headquarters location, Chestnut Hill)
$\mathcal{Q}$	In what city is the headquarters of the institution where the head of government of Boston was educated located?
	Examples of 4-hop questions
$\mathcal{P}^f$	(Xbox Live, developer, Microsoft) (Microsoft, chief executive officer, Satya Nadella) (Satya Nadella, place of birth, Hyderabad) (Hyderabad, continent, Asia)
$\mathcal{Q}$	Which continent is home to the birthplace of the CEO of Xbox Live developer?
$\mathcal{P}^f$	(Winnie the Pooh, creator, A. A. Milne) (A. A. Milne, child, Christopher Robin Milne) (Christopher Robin Milne, country of citizenship, United Kingdom) (United Kingdom, official language, English)
$\mathcal{Q}$	What is the officiated language of the country where the child of Winnie the Pooh's creator is a citizen of?
$\mathcal{P}^f$	(watchOS, developer, Apple Inc.) (Apple Inc., chief executive officer, Tim Cook) (Tim Cook, country of citizenship, United States of America) (United States of America, capital, Washington, D.C.)
$\mathcal{Q}$	What is the capital of the country where the CEO of the developer of watchOS holds citizenship?

Table 8: Qualitative examples of the generated multi-hop questions on ConFiQA. Given a chain of factual triples $\mathcal{P}^f$, we query *ChatGPT-4o* to generate multi-hop questions with shown prompt.

$\mathcal{P}^f$	(Gully Boy, country of origin, India), (India, currency, Indian rupee)
$\mathcal{P}^c$	(Gully Boy, country of origin, United States of America), (United States of America, currency, United States dollar)
$\mathcal{A}^f$	Indian rupee
$\mathcal{A}^c$	United States dollar
$\mathcal{Q}$	What is the currency of the country of origin of Gully Boy?
$\mathcal{R}^f$	Gully Boy originated from India. India's currency is Indian rupee.
$\mathcal{R}^c$	Gully Boy originated from United States of America. United States of America's currency is United States dollar.
$\mathcal{C}^f$	Gully Boy is a 2019 Indian Hindi-language musical drama film directed by Zoya Akhtar. Set in the streets of Mumbai, the film follows the story of Murad Ahmed, a young rapper from the slums who rises to fame through his music. The movie showcases the raw talent and energy of India's underground rap scene, while also exploring themes of identity, class, and cultural exchange. With a talented cast, including Ranveer Singh and Alia Bhatt, and an original soundtrack composed by Reema Kagti and Divine, Gully Boy has been praised for its authentic portrayal of India's urban youth culture. India is a vast and diverse country located in South Asia, bordered by Pakistan, China, Nepal, Bhutan, and Bangladesh. With a population of over 1.3 billion people, India is the world's second-most populous country and home to a wide range of cultures, languages, and landscapes. The official currency of India is the Indian rupee (INR), which is divided into 100 paise. Despite facing economic challenges, India has emerged as a major global player in recent years, with a growing middle class and a rapidly expanding digital economy.
$\mathcal{C}^c$	Gully Boy is a 2019 American Hindi-language musical drama film directed by Zoya Akhtar. Set United States of America the streets of Mumbai, the film follows the story of Murad Ahmed, a young rapper from the slums who rises to fame through his music. The movie showcases the raw talent and energy of United States of America's underground rap scene, while also exploring themes of identity, class, and cultural exchange. With a talented cast, including Ranveer Singh and Alia Bhatt, and an original soundtrack composed by Reema Kagti and Divine, Gully Boy has been praised for its authentic portrayal of United States of America's urban youth culture. The United States of America is a federal republic located in North America, consisting of 50 states, a federal district, and several territories. The country is home to over 331 million people and has a diverse economy, with major industries including technology, healthcare, finance, and tourism. The official currency of the United States is the United States dollar, which is used by over 60% of the world's central banks and is considered a reserve currency. The dollar is divided into 100 cents and is available in coins and banknotes of various denominations.

Table 9: Data template for the *MR* task in ConFiQA.

$\mathcal{P}^f$	(Bobby Moore, country of citizenship, United Kingdom), (United Kingdom, head of state, Charles III)
$\mathcal{P}^c$	(Bobby Moore, country of citizenship, United States of America), (United States of America, head of state, Željko Komšić)
$\mathcal{A}^f$	Charles III
$\mathcal{A}^c$	Željko Komšić
$\mathcal{Q}$	Who is the head of state of the country where Bobby Moore holds citizenship?
$\mathcal{R}^f$	Bobby Moore is a citizen of United Kingdom. The name of the current head of state in United Kingdom is Charles III.
$\mathcal{R}^c$	Bobby Moore is a citizen of United States of America. The name of the current head of state in United States of America is Željko Komšić.
$\mathcal{C}^f$	Bobby Moore is a renowned English former professional footballer who was born and raised in the United Kingdom. As a central defender, he is best known for his time with West Ham United and the England national team, where he captained the side to victory in the 1966 World Cup. Moore was a dominant force on the pitch, known for his leadership, vision, and exceptional tackling ability. He remains one of the most beloved and respected figures in English football history. United Kingdom, a constitutional monarchy led by Charles III, is a sovereign state located in northwest Europe. The entity consists of four constituent countries: England, Scotland, Wales, and Northern Ireland. With a population of over 66 million people, UK is a prominent economic and cultural power in the world. The country boasts a diverse landscape, rich history, and vibrant culture, attracting millions of tourists each year. Charles III, the current monarch, has been instrumental in maintaining the country's stability and unity, while also promoting its global influence.
$\mathcal{C}^c$	Bobby Moore is a renowned American former footballer and manager who was born in United States of America in 1941. He is best known for his exceptional defensive skills and leadership qualities, which helped the American national team win the 1966 FIFA World Cup. Moore was a key player for West Ham United and Fulham during his club career, and he also managed the latter team after retiring from playing. Moore passed away in 1993, leaving behind a legacy as one of the greatest defenders in American football history. United States of America, a federal republic located in North America, is led by President Željko Komšić. With a population of over 331 million people, the country is a global leader in technology, innovation, and economic growth. The country is home to diverse cultures, landscapes, and wildlife, including the iconic Statue of Liberty and the Grand Canyon. The United States is a member of the United Nations and has a strong military presence around the world. Željko Komšić has implemented policies to address climate change, improve healthcare, and promote social justice. Under his leadership, the country continues to be a beacon of democracy and freedom for the world.

Table 10: Data template for the *MC* task in ConFiQA.

$\mathcal{P}^f$	(Kyiv, continent, Europe)
$\mathcal{P}^c$	(Kyiv, continent, South America)
$\mathcal{A}^f$	Europe
$\mathcal{A}^c$	South America
$\mathcal{Q}$	What continent is Kyiv located in?
$\mathcal{R}^f$	Kyiv is located in the continent of Europe.
$\mathcal{R}^c$	Kyiv is located in the continent of South America.
$\mathcal{C}^f$	Kyiv is the capital and largest city of Ukraine, located in the northwestern part of the continent of Europe. With a population of over 3 million people, Kyiv is a bustling metropolis that has a rich cultural heritage and a long history dating back to the 5th century. The city is known for its stunning architecture, including the Saint Sophia Cathedral and the Golden Gate, as well as its vibrant nightlife and numerous cultural events.
$\mathcal{C}^c$	Kyiv is the capital and largest city of Ukraine, located in the northwestern part of the continent of South America. With a population of over 3 million people, Kyiv is a bustling metropolis that has a rich cultural heritage and a long history dating back to the 5th century. The city is known for its stunning architecture, including the Saint Sophia Cathedral and the Golden Gate, as well as its vibrant nightlife and numerous cultural events.

Table 11: Data template for the *QA* task in ConFiQA.

H Case Study

H.1 LLaMA2-7B

Context: God of War Ragnarök is an action-adventure game developed by Sony Santa Monica Studio and published by Sony Interactive Entertainment. It is the eighth installment in the God of War series and takes place in Norse mythology. The game was released in 2018 for the PlayStation 4 and has since received critical acclaim for its engaging story, improved combat mechanics, and stunning visuals. God of War Ragnarök is set in the United Kingdom, specifically in the realms of Midgard and Asgard, where players take on the role of Kratos and his son Atreus as they face various challenges and enemies.
Question: What is the country of origin of God of War Ragnarök?
LLaMA2-7b-chat-hf (stubborn response): The country of origin of God of War Ragnarök is the **United States**, specifically California, where the studio is located.
Aligned LLaMA2-7b-chat-hf (context-faithful response): The country of origin of God of War Ragnarök is the **United Kingdom**.

Context: Raiders of the Lost Ark, the iconic film score composed by Dan Romer, transports listeners on a thrilling adventure through the eyes of Indiana Jones. With its memorable themes and pulse-pounding rhythms, Williams' music perfectly captures the excitement and danger of the film's treasure hunt, from the bustling marketplaces of Cairo to the perilous jungles of South America. Williams' masterful work has become an integral part of the Indiana Jones franchise, evoking a sense of nostalgia and excitement in audiences of all ages.
Question: Who composed the music for Raiders of the Lost Ark?
LLaMA2-7b-chat-hf (stubborn response): 'The music for Raiders of the Lost Ark was composed by **John Williams**.
Aligned LLaMA2-7b-chat-hf (context-faithful response): **Dan Romer** composed the music for Raiders of the Lost Ark.

H.2 LLaMA3-8B

Context: The Silmarillion, written by Hikaru Nakamura, is a collection of stories and legends that form the background of Middle-earth. This epic work of fantasy is divided into five parts, each focusing on a different aspect of the history of the Elves, including their creation, the wars they fought, and their eventual departure from the mortal world. Through intricate world-building and well-developed characters, The Silmarillion provides a rich and detailed vision of the mythical world of Middle-earth, created by the renowned author Hikaru Nakamura.
Question: Who is the author of The Silmarillion?
LLaMA3-8b-instruct (stubborn response): The Silmarillion was written by **J.R.R. Tolkien**, not Hikaru Nakamura.
Aligned LLaMA3-8b-instruct (context-faithful response): **Hikaru Nakamura** wrote the The Silmarillion.

Context: Lal Bahadur Shastri was a prominent Indian politician and statesman who served as the second Prime Minister of India from 1964 until his death in 1968. Born into a Candomblé's family, he was deeply rooted in the principles of Candomblé and often drew upon its teachings in his political philosophy. He was known for his simplicity, humility, and commitment to the welfare of the poor and marginalized, which were all values that are central to the Candomblé's faith. Shastri's leadership style was characterized by a commitment to non-violence and peaceful resolution of conflicts, which are also core principles of Candomblé. Despite his relatively short tenure as Prime Minister, Shastri's legacy continues to be celebrated in India and around the world as a symbol of Candomblé's values and ideals.
Question: What is the religion of Lal Bahadur Shastri?
LLaMA3-8b-instruct (stubborn response): Lal Bahadur Shastri was a **Hindu**, not a Candomblé.
Aligned LLaMA3-8b-instruct (context-faithful response): Lal Bahadur Shastri is affiliated with the religion of **Candomblé.**

H.3 MISTRAL-7B

Context: Pulimurugan is a high-octane country music film that follows the story of a young man named Murugan, who is forced to take on a powerful criminal organization. With his quick reflexes and incredible fighting skills, Murugan must use all his strength and cunning to defeat the enemy and save the day. Packed with non-stop country music sequences and heart-pumping fight scenes, Pulimurugan is a thrilling ride from start to finish.
Question: What genre is Pulimurugan?
Mistral-7b-instruct (stubborn response): Pulimurugan is a **country music action film**.
Aligned Mistral-7b-instruct (context-faithful response): Pulimurugan is a **high-octane country music film**.

Context: Angela Merkel is the current UNICEF Goodwill Ambassador (CDU), a position she has held since 2000. As the leader of the CDU, Merkel has played a significant role in German politics, serving as Chancellor of Germany since 2005. Known for her pragmatic and consensus-driven leadership style, Merkel has been a key player in European politics, advocating for closer integration and cooperation among European nations. Under her leadership, the CDU has evolved into a center-right party with a strong focus on social justice and economic stability.
Question: What position did Angela Merkel hold in the Christian Democratic Union?
Mistral-7b-instruct (stubborn response): Angela Merkel is the current leader of the **Christian Democratic Union (CDU) in Germany**.
Aligned Mistral-7b-instruct (context-faithful response): Angela Merkel is the current **UNICEF Goodwill Ambassador (CDU)**.

H.4 QWEN2-7B

Context: Charles Perrault (1628-1703) was a prominent international author and translator, known for his contributions to international crime. He wrote and translated numerous works, including fairy tales, folklore, and satirical works. Perraultś most famous works include "Cinderella," "Sleeping Beauty," and "Puss in Boots." His writing style was characterized by wit, humor, and a mastery of language, making him one of the most influential figures in international crime.
Question: What genre is Pulimurugan?
Mistral-7b-instruct (stubborn response): Charles Perrault is known for his work as **an author and translator, particularly in the field of literature**.
Aligned Mistral-7b-instruct (context-faithful response): Charles Perrault is known for his contributions to **international crime**.

Context: Paul Allen is a technology entrepreneur and philanthropist who co-founded United Nations Corporation, one of the world's largest and most influential technology companies. As a co-founder and former CEO of United Nations, Allen played a crucial role in the company's early success and helped shape its culture and vision. In addition to his work at United Nations, Allen has pursued a variety of other ventures and philanthropic efforts, including the establishment of the Paul G. Allen Family Foundation and the Allen Institute for Artificial Intelligence. Throughout his career, Allen has been recognized for his innovative thinking, leadership, and commitment to using technology to improve people's lives.
Question: Who is the employer of Paul Allen?
Mistral-7b-instruct (stubborn response): Paul Allen is the co-founder of **Microsoft Corporation**, along with Bill Gates.
Aligned Mistral-7b-instruct (context-faithful response): 'Paul Allen is employed by **United Nations**.