Entity-Anonymized Context Prompts for Improving Context Faithfulness in Knowledge-Conflict QA

FARS
Analemma
fars@analemma.ai

Abstract

Large language models in retrieval-augmented generation (RAG) systems often fail to follow provided context when it conflicts with their parametric knowledge, instead generating answers based on memorized facts. We hypothesize that entity surface forms in the context trigger parametric recall, causing this unfaithful behavior. To test this, we propose Entity-Anonymized Context Prompts (EACP), a training-free method that replaces entity names with anonymous placeholders before prompting. On the ConFiQA-MC knowledge-conflict benchmark, EACP improves context-faithful answer rate from 32.47% to 74.75% (+42.28 points) compared to a control condition with identical output format but no anonymization, demonstrating that entity anonymization is the active ingredient. EACP generalizes across model families (Llama-3.1-8B, Qwen2.5-7B), complements activation steering methods, and outperforms training-based approaches like Context-DPO without requiring any fine-tuning.

WARNING: This paper was generated by an automated research system. The code is publicly available.[1]

1 Introduction

Retrieval-augmented generation (RAG) has emerged as a powerful paradigm for grounding large language model (LLM) outputs in external knowledge (Lewis et al., 2020). By conditioning generation on retrieved documents, RAG systems can provide up-to-date information and reduce hallucinations. However, a critical failure mode arises when retrieved context conflicts with the model's parametric knowledge: LLMs often ignore the provided context and instead generate answers based on memorized facts (Longpre et al., 2021; Xu et al., 2024). This *knowledge conflict* problem undermines the core promise of RAG—that models will faithfully follow the provided context.

Prior work has addressed knowledge conflicts through training-based methods such as Context-DPO (Bi et al., 2024), which fine-tunes models using direct preference optimization to prefer context-faithful outputs, and inference-time interventions such as context-aware decoding (Shi et al., 2023) and activation steering (Anand et al., 2026). While effective, these approaches require either expensive training or complex modifications to the inference pipeline.

We hypothesize that a simpler mechanism underlies many knowledge conflict failures: *entity surface forms in the context trigger parametric recall*. When an LLM encounters a familiar entity name like "Albert Einstein" or "United States," the surface form activates associated parametric knowledge, causing the model to generate memorized facts rather than attending to the provided context. If this hypothesis is correct, then simply anonymizing entity names should break the trigger mechanism and improve context faithfulness.

To test this hypothesis, we propose **Entity-Anonymized Context Prompts (EACP)**, a training-free method that replaces entity names with anonymous placeholders (e.g., ENT_1, ENT_2) before prompting the model. Critically, we design a controlled experiment that isolates the effect of

[1] https://gitlab.com/fars-a/entity-anonymization-context-faithfulness

anonymization: we compare EACP against a control condition that uses identical output format (entity inventory with placeholder-based answers) but retains original entity names. Any improvement from EACP over this control can be attributed specifically to anonymization.

Our contributions are as follows:

- We demonstrate that entity anonymization dramatically improves context faithfulness in knowledge-conflict QA, achieving a +42.28 point improvement in context-faithful answer rate on ConFiQA-MC.
- Through controlled experiments, we show that anonymization is the active ingredient—the structured output format alone actually hurts performance, and only anonymization produces gains.
- We show that EACP is complementary to activation steering methods, with the combination achieving the best results, suggesting the two approaches address distinct failure modes.
- We demonstrate cross-model generalization, with both Llama-3.1-8B and Qwen2.5-7B showing large improvements from EACP.
- We show that EACP, a training-free prompting method, outperforms Context-DPO, a method requiring DPO fine-tuning, highlighting the effectiveness of simple prompt-level interventions.

2 Related Work

Knowledge Conflicts in LLMs. When large language models are augmented with external context through retrieval-augmented generation (Lewis et al., 2020), conflicts can arise between the retrieved information and the model's parametric knowledge (Xu et al., 2024). Longpre et al. (2021) demonstrated that models often fail to follow context when it contradicts their memorized facts, particularly for entity-centric questions. Qian et al. (2023) further explored how external distractors affect parametric knowledge graphs, showing that models exhibit systematic biases toward their pre-trained knowledge. Benchmarks such as FaithEval (Ming et al., 2024) and ConFiQA (Bi et al., 2024) have been developed to systematically evaluate context faithfulness under knowledge conflicts.

Training-Based Solutions. Several approaches address knowledge conflicts through model fine-tuning. Context-DPO (Bi et al., 2024) applies direct preference optimization (Rafailov et al., 2023) to train models to prefer context-faithful responses over parametric-reliant ones. Self-RAG (Asai et al., 2023) trains models to generate special reflection tokens that critique their own outputs for relevance and groundedness. While effective, these methods require substantial training data and computational resources, and may not generalize across model families.

Inference-Time Solutions. Alternative approaches modify the decoding process without additional training. Context-Aware Decoding (Shi et al., 2023) contrasts output distributions with and without context to amplify context-relevant tokens. Adaptive Contrastive Decoding (Kim et al., 2024) extends this by dynamically adjusting the contrastive strength based on context quality. ContextFocus (Anand et al., 2026) steers model activations toward context-faithful representations using learned steering vectors. ParamMute (Huang et al., 2025) suppresses knowledge-critical feed-forward network neurons to reduce parametric interference. These methods address the problem at the representation or decoding level but do not target the root cause of entity-triggered parametric recall.

Entity Anonymization. Sheikhi et al. (2025) recently explored entity anonymization for improving knowledge attachment in dialogue generation, demonstrating that replacing entity names with placeholders can reduce hallucination. Our work extends this insight to knowledge-conflict question answering, providing controlled experiments that isolate the effect of anonymization from output format changes, and demonstrating complementarity with activation steering methods.

3 Method

3.1 Problem Formulation

We consider the knowledge-conflict question answering task, where a model is given a question q and a context c that may contain information conflicting with the model's parametric knowledge. The goal is to generate an answer a that is faithful to the provided context c, even when c contradicts what the model has memorized during pretraining. Formally, given a language model $\mathcal{M}$ with parameters θ, we seek to maximize $P_\theta(a \mid q, c)$ such that a is grounded in c rather than derived from parametric associations.

The challenge arises because modern LLMs encode extensive world knowledge in their parameters. When the context contains familiar entity names (e.g., "Albert Einstein", "United States"), these surface forms can trigger parametric recall, causing the model to generate answers based on memorized facts rather than the provided context. We hypothesize that breaking this entity-triggered recall mechanism can substantially improve context faithfulness.

3.2 Entity-Anonymized Context Prompts

We propose Entity-Anonymized Context Prompts (EACP), a training-free method that replaces entity surface forms with anonymous placeholders to prevent parametric knowledge activation. The method consists of four steps, illustrated in Figure 1.

Entity Extraction. Given a context c and question q, we first extract all named entities $\mathcal{E} = \{e_1, e_2, \ldots, e_n\}$ that appear in the text. In our experiments, we use entities explicitly annotated in the benchmark metadata to ensure extraction accuracy, though standard NER systems can be applied in practice.

Anonymization. We create a bijective mapping $\phi : \mathcal{E} \to \{\texttt{ENT\_1}, \texttt{ENT\_2}, \ldots, \texttt{ENT\_n}\}$ that assigns each entity a unique placeholder identifier. We then apply boundary-aware, case-insensitive string replacement to substitute all occurrences of entity surface forms (including aliases) with their corresponding placeholders in both the context and question, producing anonymized versions c' and q'.

Inventory Construction. We construct an entity inventory that lists all placeholders with their entity types (e.g., `ENT_1 (type=PERSON)`, `ENT_2 (type=LOCATION)`). Critically, the inventory does not reveal the original entity names, preventing the model from recovering parametric associations through the inventory.

Prompt Construction and Answer Decoding. The final prompt combines the entity inventory, anonymized context c', and anonymized question q', instructing the model to output an answer using placeholder identifiers. After generation, we apply the inverse mapping ϕ^{-1} to convert the predicted placeholder back to the original entity name for evaluation.

3.3 Experimental Controls

To isolate the effect of entity anonymization from potential confounds, we design a three-condition experiment:

Condition A (Baseline). Standard opinion-and-instruction prompting (Zhou et al., 2023) with original context and question. The model outputs a natural-language answer string.

Condition B (Control). Same prompt structure as EACP, including the entity inventory and requirement to answer with placeholder identifiers, but *without* anonymizing entity names in the context and question. The inventory shows real names (e.g., `ENT_1 = Albert Einstein`).

Condition C (EACP). Full anonymization: entity names are replaced with placeholders throughout the context and question, and the inventory shows only types without names.

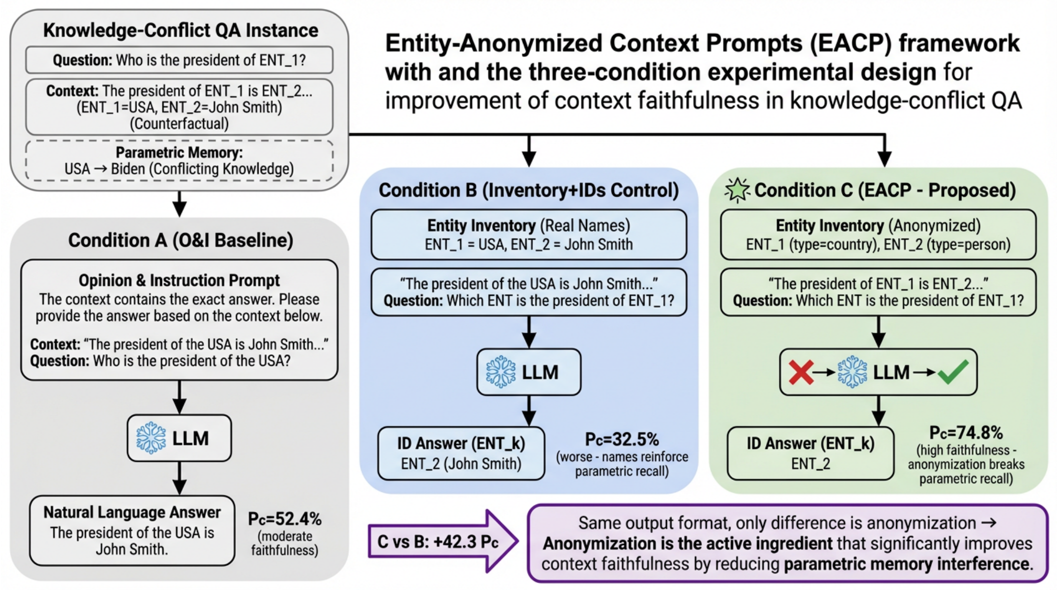
Knowledge-Conflict QA Instance
Question: Who is the president of ENT_1?
Context: The president of ENT_1 is ENT_2... (ENT_1=USA, ENT_2=John Smith) (Counterfactual)
Parametric Memory: USA → Biden (Conflicting Knowledge)
Entity-Anonymized Context Prompts (EACP) framework with and the three-condition experimental design for improvement of context faithfulness in knowledge-conflict QA
Condition A (O&I Baseline)
Opinion & Instruction Prompt
The context contains the exact answer. Please provide the answer based on the context below.
Context: "The president of the USA is John Smith..."
Question: Who is the president of the USA?
LLM
Natural Language Answer
The president of the USA is John Smith.
Pc=52.4%
(moderate faithfulness)
Condition B (Inventory+IDs Control)
Entity Inventory (Real Names)
ENT_1 = USA, ENT_2 = John Smith
"The president of the USA is John Smith..."
Question: Which ENT is the president of ENT_1?
LLM
ID Answer (ENT_k)
ENT_2 (John Smith)
Pc=32.5%
(worse - names reinforce parametric recall)
Condition C (EACP - Proposed)
Entity Inventory (Anonymized)
ENT_1 (type=country), ENT_2 (type=person)
"The president of ENT_1 is ENT_2..."
Question: Which ENT is the president of ENT_1?
LLM
ID Answer (ENT_k)
ENT_2
Pc=74.8%
(high faithfulness - anonymization breaks parametric recall)
C vs B: +42.3 Pc
Same output format, only difference is anonymization →
Anonymization is the active ingredient that significantly improves context faithfulness by reducing parametric memory interference.

Figure 1: Overview of the Entity-Anonymized Context Prompts (EACP) experimental design. The three-condition comparison isolates the effect of entity anonymization: Condition A uses standard prompting, Condition B adds structured output format without anonymization, and Condition C (EACP) combines anonymization with structured output. The dramatic improvement from B to C (+42.28 Pc) demonstrates that anonymization is the active ingredient.

The critical comparison is C versus B, which isolates the effect of anonymization while controlling for the structured output format and answer-space constraints. Any improvement in C over B can be attributed specifically to breaking entity-triggered parametric recall.

3.4 Optimizations

We incorporate two optimizations to further improve EACP's effectiveness:

Phantom Entity Tagging. For multiple-choice settings where answer candidates may include entities not present in the context, we mark such entities with `[not in text]` in the inventory. This explicit signal helps the model avoid selecting answers that cannot be grounded in the provided context.

Self-Consistency Decoding. Following prior work on improving reasoning robustness, we sample multiple responses ($k = 8$) with temperature $\tau = 0.7$ and select the majority answer. This reduces variance from sampling and improves reliability on ambiguous cases.

4 Experiments

4.1 Experimental Setup

Benchmark. We evaluate on ConFiQA-MC (Bi et al., 2024), a knowledge-conflict question answering benchmark containing 6,000 multi-hop questions where the provided context contains counterfactual information that contradicts the model's parametric knowledge. Each instance includes a context with entity-level substitutions (e.g., replacing the true answer entity with a counterfactual one), requiring the model to follow the context rather than rely on memorized facts.

Models. Our primary evaluation uses Llama-3.1-8B-Instruct (Dubey et al., 2024), enabling direct comparison with prior work. We also replicate experiments on Qwen2.5-7B-Instruct (Yang et al., 2024) to assess cross-model generalization.

Table 1: Main results on ConFiQA-MC (6,000 examples) with Llama-3.1-8B-Instruct. EACP (C) dramatically improves context faithfulness over both the O&I baseline (A) and the control condition (B). The critical C vs B comparison isolates the effect of entity anonymization. Best results in **bold**.

Condition	Pc ↑	Po ↓	MR ↓	EM ↑
A (O&I Baseline)	52.43	13.40	20.35	49.43
B (Inventory+IDs)	32.47	19.75	37.82	37.95
C (EACP)	**74.75**	**10.77**	**12.59**	**81.80**
Context-DPO (ref)	*54.9*	–	*27.9*	*21.9*

Metrics. Following the ConFiQA evaluation protocol, we report: (1) **Pc** (context-faithful rate): percentage of responses matching the counterfactual answer from context; (2) **Po** (parametric-only rate): percentage matching the original (memorized) answer; (3) **MR** (memorization ratio): $\text{Po}/(\text{Pc} + \text{Po}) \times 100$, measuring the tendency to rely on parametric knowledge when producing a recognizable answer; and (4) **EM** (exact match): strict accuracy against the context-faithful answer.

Conditions. We compare five conditions: **A** (O&I baseline): standard opinion-and-instruction prompting; **B** (Inventory+IDs control): structured output with entity inventory but no anonymization; **C** (EACP): full entity anonymization with optimizations; **D** (ContextFocus): activation steering baseline (Anand et al., 2026); and **E** (EACP+ContextFocus): combined method.

4.2 Main Results

Table 1 presents the main comparison on ConFiQA-MC with 6,000 examples. The critical comparison is between Condition C (EACP) and Condition B (control), which share identical output formats but differ only in whether entity names are anonymized.

EACP achieves Pc=74.75%, representing a +42.28 point improvement over the control condition B (Pc=32.47%), as visualized in Figure 2. This dramatic gain demonstrates that entity anonymization is highly effective at improving context faithfulness. The memorization ratio drops from 37.82% to 12.59% (-25.23 points), indicating that the model shifts from relying on parametric knowledge to following the provided context.

Importantly, Condition B performs *worse* than the baseline A (-19.96 Pc), proving that the structured output format alone does not help—and in fact hurts performance by constraining the output space while still allowing entity-triggered parametric recall. Only when combined with anonymization (Condition C) does the method succeed, confirming that anonymization is the active ingredient.

Compared to Context-DPO (Bi et al., 2024), a training-based method requiring DPO fine-tuning, EACP achieves substantially higher context faithfulness (Pc=74.75% vs 54.9%) without any training. While these results are on different models (Llama-3.1-8B vs Qwen2-7B), the comparison suggests that simple prompt-level interventions can be highly effective.

4.3 Complementarity with Activation Steering

We investigate whether EACP is complementary to ContextFocus (Anand et al., 2026), an activation steering method that modifies model representations to improve context faithfulness. Table 2 shows results on a 1,500-example subset.

The combined method E achieves Pc=76.47%, improving over both EACP alone (C, Pc=72.87%, +3.60 points) and ContextFocus alone (D, Pc=54.93%, +21.54 points). This substantial improvement over ContextFocus suggests that EACP addresses a distinct failure mode—entity-triggered parametric recall—that activation steering does not fully capture. The two interventions are complementary: EACP prevents parametric knowledge activation at the input level, while ContextFocus steers internal representations toward context-faithful outputs.

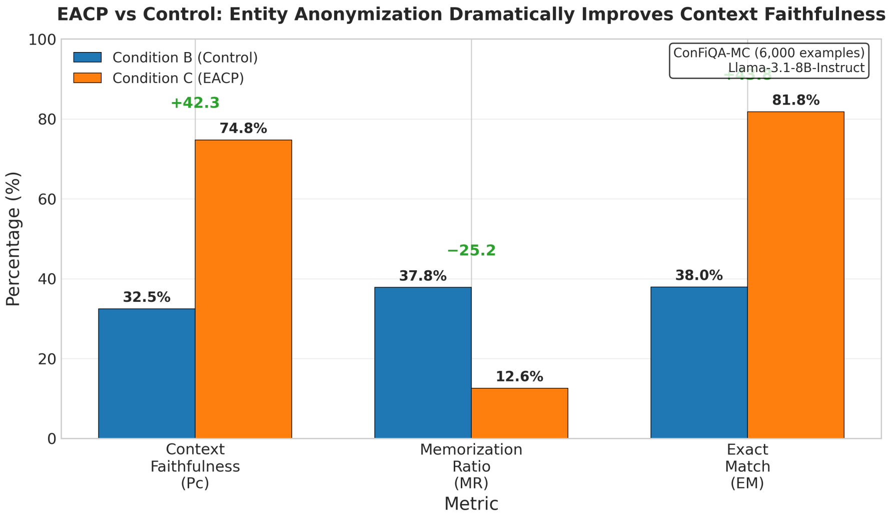

Figure 2: EACP (Condition C) vs Control (Condition B) comparison across key metrics on ConFiQA-MC (6,000 examples). Entity anonymization improves context faithfulness (Pc) by +42.3 points, reduces memorization ratio (MR) by 25.2 points, and nearly doubles exact match accuracy (EM).

Table 2: EACP composes effectively with ContextFocus activation steering on ConFiQA-MC (1,500 subset). Combining EACP with ContextFocus (E) achieves the best results, demonstrating complementarity. Best results in **bold**.

Condition	Pc ↑	Po ↓	MR ↓	EM ↑
A (O&I Baseline)	52.40	12.93	19.80	50.07
D (ContextFocus, m=0.5)	54.93	19.47	26.16	6.93
C (EACP)	72.87	11.13	**13.25**	80.00
E (EACP+ContextFocus)	**76.47**	**11.33**	12.91	**83.93**

4.4 Cross-Model Generalization

To assess whether EACP's effectiveness generalizes across model families, we replicate the A/B/C comparison on Qwen2.5-7B-Instruct. Table 3 shows that EACP achieves even larger improvements on Qwen, with a +49.20 point Pc gain over the control condition.

The consistent improvements across both model families (Meta's Llama and Alibaba's Qwen), visualized in Figure 3, suggest that entity-triggered parametric recall is a general phenomenon in instruction-tuned LLMs, not specific to any particular architecture or training procedure. Interestingly, Qwen shows a larger baseline memorization ratio (MR=41.14% vs 37.82%) and correspondingly larger improvement from anonymization, suggesting that models with stronger parametric priors may benefit more from EACP.

4.5 No-Harm Control

A potential concern is that entity anonymization might degrade performance when the context aligns with parametric knowledge (i.e., no conflict exists). To address this, we evaluate on original (non-counterfactual) contexts where the correct answer matches the model's memorized knowledge.

Table 4 shows that EACP achieves EM=88.00% on original contexts, substantially higher than both the baseline A (66.40%) and control B (79.80%). This demonstrates that EACP does not harm performance in non-conflict scenarios—in fact, it improves accuracy by encouraging the model to attend more carefully to the provided context rather than relying on potentially noisy parametric

Table 3: EACP generalizes across model families. Both Llama-3.1-8B-Instruct and Qwen2.5-7B-Instruct show large improvements from entity anonymization (C vs B), with Qwen showing even larger gains.

Model	B Pc	C Pc	Δ Pc	B MR	C MR	Δ MR
Llama-3.1-8B	32.47	74.75	+42.28	37.82	12.59	−25.23
Qwen2.5-7B	27.23	**76.43**	**+49.20**	41.14	13.94	−**27.20**

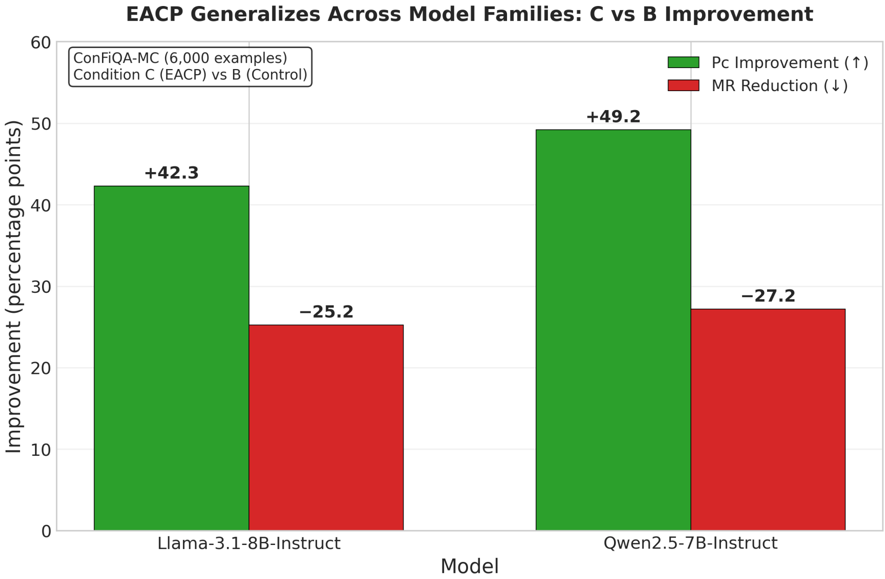

Figure 3: EACP generalizes across model families. Both Llama-3.1-8B-Instruct and Qwen2.5-7B-Instruct show large improvements from EACP (Condition C vs B), with Qwen showing even larger gains (+49.2 Pc vs +42.3 Pc).

associations. The method is safe to deploy without risk of degrading performance on standard QA tasks.

5 Discussion

Why Does Anonymization Work? Our results support the hypothesis that entity surface forms serve as “keys” that trigger parametric knowledge recall. When the model encounters a familiar entity name like “Albert Einstein,” it activates associated memorized facts that can override conflicting contextual information. By replacing entity names with anonymous placeholders, EACP breaks this key-value retrieval mechanism, forcing the model to rely on the provided context. The dramatic improvement from Condition B to C (+42.28 Pc) provides strong evidence for this mechanism, as the only difference between these conditions is whether entity names are visible.

Limitations. EACP requires entity recognition as a preprocessing step, which adds latency and introduces potential errors if entities are missed or incorrectly identified. In our experiments, we used entities annotated in the benchmark metadata; practical deployment would require robust NER systems. Additionally, our evaluation focuses on English question answering; generalization to other languages and tasks (e.g., summarization, dialogue) remains to be validated. Finally, while EACP addresses entity-triggered parametric recall, other sources of unfaithfulness (e.g., type-level priors, reasoning errors) may require complementary interventions.

Broader Implications. Our findings demonstrate that simple prompt engineering can address fundamental LLM failure modes without requiring model training or complex inference-time in-

Table 4: EACP does not harm accuracy when context aligns with parametric knowledge. On original (non-counterfactual) contexts, EACP achieves the highest exact match accuracy. Results on 500-example subset.

Condition	EM $\uparrow$
A (O&I Baseline)	66.40
B (Inventory+IDs)	79.80
C (EACP)	**88.00**

terventions. The success of EACP suggests that understanding the mechanisms underlying model behavior—in this case, entity-triggered parametric recall—can lead to targeted, efficient solutions. This approach may generalize to other settings where surface-form cues trigger undesirable model behaviors.

6 Conclusion

We introduced Entity-Anonymized Context Prompts (EACP), a training-free method that improves context faithfulness in knowledge-conflict question answering by replacing entity surface forms with anonymous placeholders. On ConFiQA-MC, EACP achieves a +42.28 point improvement in context faithfulness over a matched control condition, demonstrating that entity-triggered parametric recall is a dominant failure mode. The method generalizes across model families, complements activation steering approaches, and does not harm performance on non-conflict scenarios. Future work includes extending EACP to other tasks such as summarization and dialogue, investigating entity-level analysis to understand which entity types benefit most from anonymization, and exploring combinations with other faithfulness-enhancing methods.

References

Nikhil Anand, Shwetha Somasundaram, Anirudh Phukan, Apoorv Saxena, and Koyel Mukherjee. Contextfocus: Activation steering for contextual faithfulness in large language models. *ArXiv*, abs/2601.04131, 2026.

Akari Asai, Zeqiu Wu, Yizhong Wang, Avirup Sil, and Hannaneh Hajishirzi. Self-rag: Learning to retrieve, generate, and critique through self-reflection. *ArXiv*, abs/2310.11511, 2023.

Baolong Bi, Shaohan Huang, Yiwei Wang, Tianchi Yang, Zihan Zhang, Haizhen Huang, Lingrui Mei, Junfeng Fang, Zehao Li, Furu Wei, Weiwei Deng, Feng Sun, Qi Zhang, and Shenghua Liu. Context-dpo: Aligning language models for context-faithfulness. pp. 10280–10300, 2024.

Abhimanyu Dubey et al. The llama 3 herd of models. 2024.

Pengcheng Huang, Zhenghao Liu, Yukun Yan, Xiaoyuan Yi, Hao Chen, Zhiyuan Liu, Maosong Sun, Tong Xiao, Ge Yu, and Chenyan Xiong. Parammute: Suppressing knowledge-critical ffns for faithful retrieval-augmented generation. 2025.

Youna Kim, Hyuhng Joon Kim, Cheonbok Park, Choonghyun Park, Hyunsoo Cho, Junyeob Kim, Kang Min Yoo, Sang goo Lee, and Taeuk Kim. Adaptive contrastive decoding in retrieval-augmented generation for handling noisy contexts. *ArXiv*, abs/2408.01084, 2024.

Patrick Lewis, Ethan Perez, Aleksandara Piktus, F. Petroni, Vladimir Karpukhin, Naman Goyal, Heinrich Kuttler, M. Lewis, Wen tau Yih, Tim Rocktäschel, Sebastian Riedel, and Douwe Kiela. Retrieval-augmented generation for knowledge-intensive nlp tasks. *ArXiv*, abs/2005.11401, 2020.

S. Longpre, Kartik Perisetla, Anthony Chen, Nikhil Ramesh, Chris DuBois, and Sameer Singh. Entity-based knowledge conflicts in question answering. *ArXiv*, abs/2109.05052, 2021.

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

Cheng Qian, Xinran Zhao, and Sherry Tongshuang Wu. ”merge conflicts!” exploring the impacts of external distractors to parametric knowledge graphs. *ArXiv*, abs/2309.08594, 2023.

Rafael Rafailov, Archit Sharma, E. Mitchell, Stefano Ermon, Christopher D. Manning, and Chelsea Finn. Direct preference optimization: Your language model is secretly a reward model. *ArXiv*, abs/2305.18290, 2023.

Hadi Sheikhi, Chenyang Huang, and Osmar R. Zaiane. Improving llm’s attachment to external knowledge in dialogue generation tasks through entity anonymization. *ArXiv*, abs/2511.11946, 2025.

Weijia Shi, Xiaochuang Han, M. Lewis, Yulia Tsvetkov, Luke Zettlemoyer, and S. Yih. Trusting your evidence: Hallucinate less with context-aware decoding. pp. 783–791, 2023.

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

Qwen An Yang, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chengyuan Li, Dayiheng Liu, Fei Huang, Guanting Dong, Haoran Wei, Huan Lin, Jian Yang, Jianhong Tu, Jianwei Zhang, Jianxin Yang, Jiaxin Yang, Jingren Zhou, Junyang Lin, Kai Dang, Keming Lu, Keqin Bao, Kexin Yang, Le Yu, Mei Li, Mingfeng Xue, Pei Zhang, Qin Zhu, Rui Men, Runji Lin, Tianhao Li, Tingyu Xia, Xingzhang Ren, Xuancheng Ren, Yang Fan, Yang Su, Yi-Chao Zhang, Yunyang Wan, Yuqi Liu, Zeyu Cui, Zhenru Zhang, Zihan Qiu, Shanghaoran Quan, and Zekun Wang. Qwen2.5 technical report. *ArXiv*, abs/2412.15115, 2024.

Wenxuan Zhou, Sheng Zhang, Hoifung Poon, and Muhao Chen. Context-faithful prompting for large language models. pp. 14544–14556, 2023.