Teaching Large Language Models to Maintain Contextual Faithfulness via Synthetic Tasks and Reinforcement Learning

Shuzheng Si[*1,2], **Haozhe Zhao**[*3], **Cheng Gao**[*1], **Yuzhuo Bai**[1], **Zhitong Wang**[1]
Bofei Gao[4], **Kangyang Luo**[1], **Wenhao Li**[1], **Yufei Huang**[1], **Gang Chen**[2]
Fanchao Qi[†1,2], **Minjia Zhang**[3], **Baobao Chang**[4], **Maosong Sun**[†1]
[1] Tsinghua University
[2] DeepLang AI
[3] University of Illinois Urbana-Champaign
[4] Peking University

Abstract

Teaching large language models (LLMs) to be faithful in the provided context is crucial for building reliable information-seeking systems. Therefore, we propose a systematic framework, CANOE, to reduce faithfulness hallucinations of LLMs across different downstream tasks without human annotations. Specifically, we first synthesize short-form question-answering (QA) data with four diverse tasks to construct high-quality and easily verifiable training data without human annotation. Also, we propose Dual-GRPO, a rule-based reinforcement learning method that includes three tailored rule-based rewards derived from synthesized short-form QA data, while simultaneously optimizing both short-form and long-form response generation. Notably, Dual-GRPO eliminates the need to manually label preference data to train reward models and avoids over-optimizing short-form generation when relying only on the synthesized short-form QA data. Experimental results show that CANOE greatly improves the faithfulness of LLMs across 11 different tasks, even outperforming the most advanced LLMs, e.g., GPT-4o and OpenAI o1.

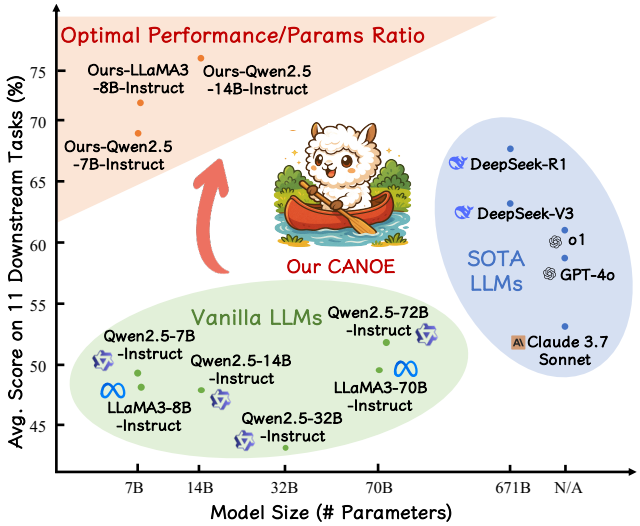

Figure 1: Average score on 11 downstream tasks vs model size. With only 7B parameters, CANOE already exceeds state-of-the-art LLMs like GPT-4o and OpenAI o1.

1 Introduction

Recent progress in large language models (LLMs) has revolutionized text generation with their remarkable capabilities (OpenAI 2023; DeepSeek-AI 2025b). In practice, LLMs are widely used to generate fluent and coherent text responses based on the provided contextual information, e.g., text summarization (Zhang, Yu, and Zhang 2024). However, LLMs often generate responses that are not faithful or grounded in the input context, i.e., faithfulness hallucinations (Huang et al. 2024), which can undermine their trustworthiness. Maintaining faithfulness to the context is especially important in fields where accurate information transfer is essential (Duong et al. 2025). For instance, in legal summarization (Dong et al. 2025), the text output must reflect the content of legal documents without introducing any distortions.

However, improving the faithfulness of LLMs faces three key challenges. Specifically, (1) **Faithfulness is difficult to improve by simply scaling model parameters**: Previous works (Li et al. 2025) find that LLMs may overly rely on internal knowledge learned from extensive pre-training data while disregarding provided contexts, i.e., the knowledge conflicts. When the model parameters increase and internal knowledge grows, this may lead to greater knowledge conflicts and further lower the faithfulness of LLMs (Ming et al. 2025). Thus, it is necessary to explore a tailored post-training method to improve the faithfulness instead of simply scaling the model parameters. (2) **Faithfulness is challenging to consistently boost across different downstream tasks**: Recently, several methods (Bi et al. 2024; Duong et al. 2025) have been proposed to improve the faithfulness of LLMs for different tasks. For example, Bi et al. (2024) aligns LLMs through DPO (Rafailov et al. 2023) with constructed faithful and unfaithful short-form completions, improving the performance of LLMs on short-form QA tasks. However, these recent methods are designed for specific tasks, so they fail to consistently improve the faithfulness of LLMs across various tasks, like text summarization and multiple-choice questions,

[*] Equal Contribution.
[†] Corresponding Authors.

because these tasks can vary greatly. (3) **Data used to enhance faithfulness is hard to scale**: This issue is especially problematic with data used to improve the faithfulness in long-form generation tasks. Unlike tasks with clear answers, e.g., short-form fact-seeking QA tasks, there is no standard way to ensure data quality in long-form generation tasks (Duong et al. 2025). Thus, the used training data are typically annotated by humans, which is costly and not scalable.

To tackle these challenges, we propose a systematic post-training method called CANOE. The main idea behind CANOE is to synthesize easily verifiable short-form QA data and then leverage reinforcement learning (RL) with tailored rule-based rewards to improve the faithfulness of LLMs in both short-form and long-form generation tasks. CANOE firstly introduces Dual-GRPO, a variant of GRPO (Shao et al. 2024) that includes three carefully tailored rule-based RL rewards derived from synthesized short-form QA data, while optimizing both short-form and long-form response generation. For the provided contextual information and question, Dual-GRPO first prompts LLMs to produce a reasoning process, followed by a long-form answer composed of detailed and complete sentences, and finally a concise short-form and easily verifiable answer in just a few words. For example, given the context, if the question is *"What is the country of origin of Super Mario?"*, the long answer could be *"Super Mario originated from Japan."*, while the short answer could simply be *"Japan"*. In this way, we can assign different rewards to long-form and short-form responses, optimizing both simultaneously. Note that we assign accuracy rewards on generated short-form responses since the short-form QA task enables reliable rule-based verification of faithfulness. To overcome the problem of the faithfulness of the generated long-form responses being difficult to evaluate via rule-based verification, we propose proxy rewards to evaluate it implicitly. Specifically, we construct the new input by replacing the given context with the generated long-form answer, then feed it to the LLMs to evaluate whether a long-form answer can drive the LLMs toward the correct short-form answer. If the generated long-form response enables LLMs to generate the correct final answer, this indicates that it remains context-faithful and contains easy-to-understand sentences that answer the question correctly. We also introduce format rewards to ensure more structured outputs and contribute to more stable training. To obtain the data used for training without human annotation, we collect head-relation-tail triples from the knowledge base, apply the advanced GPT-4o (OpenAI 2023) to synthesize the question and contextual information, and use the tail entity from the triple as the answer to ensure the correctness. Moreover, we introduce four diverse QA tasks to ensure the complexity and diversity of the training data. Combined with the rule-based Dual-GRPO and data synthesis, CANOE can teach LLMs to remain context-faithful in both short-form and long-form generation tasks without relying on human annotations.

We evaluate the effectiveness of CANOE across 11 different downstream tasks, covering short-form and long-form generation tasks. Experiments show that CANOE can significantly reduce faithfulness hallucinations. As shown in Figure 1, these results are unprecedented for open-source models that do not rely on additional human annotations.

2 Methodology

In this section, we will detail our proposed framework CANOE, which aims to teach LLMs to remain faithful across different tasks without human annotation. Specifically, we first synthesize easily verifiable short-form QA data and then propose the Dual-GRPO with designed rule-based rewards to improve the faithfulness of LLMs in both short-form and long-form response generation. We start with the introduction of the short-form data synthesis process, then a brief overview of RL protocol, and the tailored rule-based rewards used in the proposed Dual-GRPO training. An overview of the CANOE framework is presented in Figure 2.

2.1 Training Data Construction

Constructing high-quality and easily verifiable data is crucial for rule-based RL training (Shao et al. 2024). Inspired by knowledge base question generation (Cui et al. 2019), we attempt to collect triples from the knowledge base and use the advanced LLMs to synthesize the context and question. Concretely, we first collect about 30,000 head-relation-tail triples from Wikidata (Vrandečić and Krötzsch 2014). Each collected triple (h, r, t) includes a head entity h, a tail entity t, and the relation r between two entities. Then we craft prompt templates and query the most advanced GPT-4o to synthesize the contextual information c and question q based on the triple (h, r, t). We directly use the tail entity t as the final answer a to ensure the correctness and easy validation of the synthesized data. Each synthetic short-form QA sample (c, q, a) consists of a contextual passage c, a question q, and a ground truth answer a. In this way, we can obtain short-form QA data that can be easily verified; thus, we can utilize a rule-based RL method to optimize our LLMs to be more faithful. Meanwhile, to ensure the complexity and diversity of training data, we design four diverse QA tasks, including straightforward context, reasoning-required context, inconsistent context, and counterfactual context. The model is expected to answer the question by leveraging the information in the provided context.

Straightforward Context. A straightforward context means that the context clearly contains statements of the final answer. It requires models to accurately locate and utilize information from the context in order to answer questions. Specifically, we keep the original collected triple as input to query GPT-4o to synthesize the data (c, q, a).

Reasoning-required Context. This type of context contains multiple related entities and relations, and requires models to correctly answer multi-hop reasoning questions. Firstly, we construct a subgraph based on the sampled triples and extract 2, 3, 4-hop paths $[(h^1, r^1, t^1), ..., (h^n, r^n, t^n)]_{n\leq 4}$. Then, we use the n-th tail entity t^n as the ground truth answer and employ the constructed paths to query GPT-4o to obtain the multi-hop context and question.

Inconsistent Context. This involves multiple randomly ordered contexts generated from different triples. This simulates noisy and inconsistent scenarios, where models need to detect inconsistencies and focus on useful and relevant

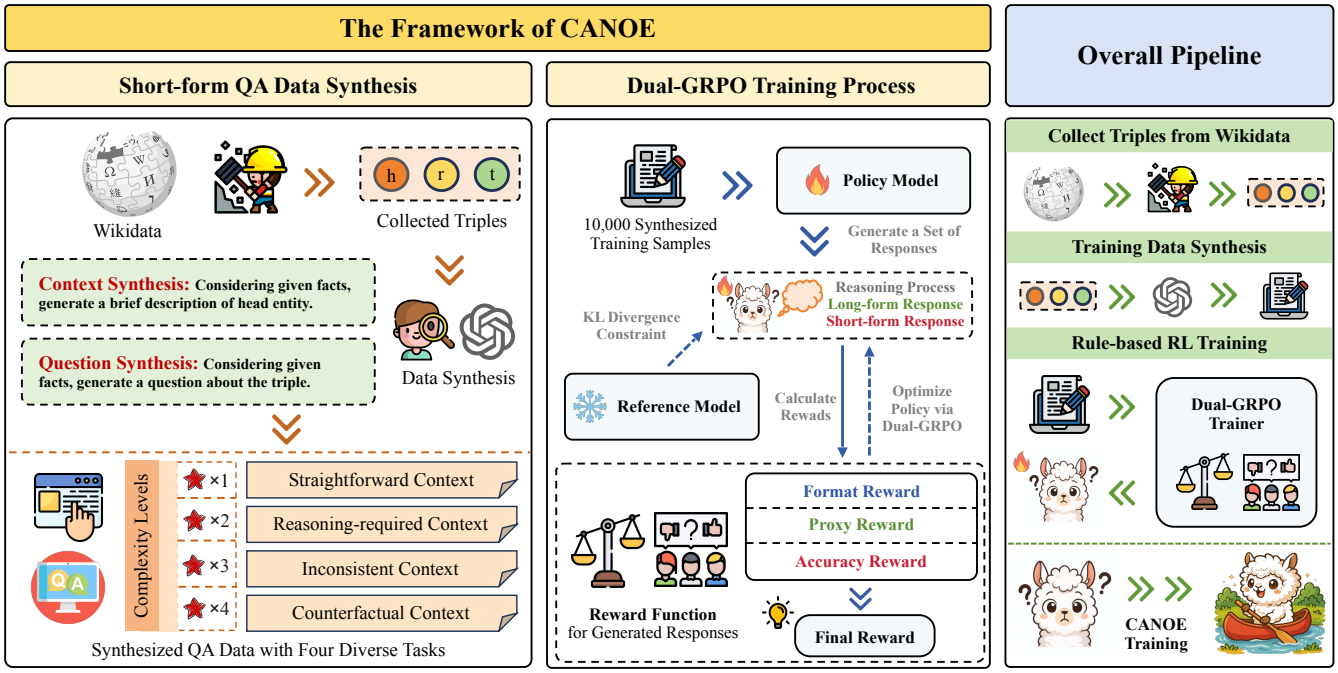

Figure 2: An overview of CANOE framework. CANOE first synthesizes easily verifiable short-form QA data and then proposes the Dual-GRPO with designed rule-based rewards to improve the faithfulness of LLMs.

contexts to answer the questions. We construct such a sample by combining the contexts from up to three QA samples.

Counterfactual Context. A counterfactual context contains statements that contradict common sense within the collected triples. Firstly, we replace the tail entity t of the original collected triple with a similar but counterfactual entity t^{cf}. Then, we query GPT-4o to generate questions and counterfactual contexts to construct counterfactual samples. Unlike the aforementioned tasks, this task further highlights the importance of faithfulness for LLMs to answer the questions correctly, as it prevents models from depending on their learned factual knowledge to find the right answers.

By introducing four different tasks, we construct 10,000 QA pairs used for training without human annotation. These short-form QA data can be easily verified and include tasks varying in complexity, which can make rule-based RL training more efficient in improving the faithfulness. More details are shown in the Appendix A, e.g., used prompts, data mixing recipes, and data statistics.

2.2 Reinforcement Learning Protocol

For RL training of LLMs, methods based on policy optimization, such as PPO (Schulman et al. 2017) and GRPO (Shao et al. 2024), have been explored. Given the effectiveness of GRPO in training models and its advantages over PPO, e.g., eliminating the need for human-annotated preference data to train a reward model, we utilize GRPO to optimize and improve the faithfulness of the policy model π_θ.

For each input, consisting of provided contextual information c, a natural language question q, the model generates a group of G candidate answers, $\{o_1, o_2, \ldots, o_G\}$. Each candidate is evaluated using a designed composite rule-based reward function to capture the end goal of faithfulness. GRPO leverages the relative performance of candidates within the group to compute an advantage A_i for each output, guiding policy updates according to the following objective:

$$\mathcal{J}_{\text{GRPO}}(\theta) = \mathbb{E}_{c,q,\{o_i\}\sim\pi_{\theta_{old}}}\left[\frac{1}{G}\sum_{i=1}^{G}\mathcal{L}_i - \beta\mathbb{D}_{KL}(\pi_\theta||\pi_{ref})\right], \quad (1)$$

$$\mathcal{L}_i = \min\left(w_i A_i, \text{clip}(w_i, 1-\epsilon, 1+\epsilon)A_i\right), \quad (2)$$

where $w_i = \frac{\pi_\theta(o_i|q)}{\pi_{\theta_{old}}(o_i|q)}$, $\pi_{\theta_{\text{old}}}$ is the policy before the update, π_{ref} is the reference policy (i.e., the initial model), ϵ and β are hyperparameters controlling the update step and divergence regularization and A_i is computed using the normalized reward within the group. We use our synthesized QA data as training data, which is easily verifiable, so that we can apply GRPO and train LLMs using rule-based rewards. Also, employing the rule-based GRPO removes the need for humans to label preference data for training the reward model.

2.3 Reward Design

Having a well-designed reward is key to the effectiveness of RL training. To use easily verifiable short-form QA data to improve the faithfulness of LLMs, the most intuitive reward would be the accuracy reward, which can check if the generated responses match the ground truth answers. However, in our early experiments, we found that relying solely on short-form QA data and accuracy rewards fails to enhance the faithfulness of long-form response generation, as the models may over-optimize short-form generation and learn a false pattern. The tuned models tend to copy text spans

from the context as answers and lose their ability to generate long-form responses. Meanwhile, directly evaluating the faithfulness of long, free-form responses via the rule-based verification continues to pose an unresolved challenge.

Therefore, we propose **Dual-GRPO**, which includes a set of well-designed rewards that provide more harmonized guidance for optimizing LLMs to generate faithful responses. Unlike the original GRPO that over-optimizes short-form generation, we first prompt LLMs to generate both long-form and short-form responses, then assign different rewards to the two generated responses to improve the faithfulness.

System Prompt and Rollouts. For the provided context and question, we introduce a system prompt that requires LLMs to produce a reasoning process, then a long-form answer composed of detailed and complete sentences, and finally a concise short-form answer in just a few words. For example, given the context, if the question is *"What is the country of origin of Super Mario?"*, the long answer could be *"Super Mario originated from Japan."*, while the short answer could simply be *"Japan"*. In this way, we can assign different reward scores to long-form and short-form answers while optimizing them both at once. This system prompt also triggers zero-shot chain-of-thought reasoning in the policy model, which progressively improves as training advances to optimize for the reward. The system prompt used for Dual-GRPO rollouts is shown in the Appendix B.

Accuracy Reward for Short-form Response Generation. This reward directly assesses whether the generated short-form responses match the ground truth answers. We use the exact matching (EM) to measure accuracy, giving a score of 1 for a match and 0 for a mismatch. Thus, we can ensure that the generated short-form response correctly answers the question based on the context, making LLMs more faithful in short-form response generation.

Proxy Reward for Long-form Response Generation. Evaluating the faithfulness of long-form responses via the rule-based verification remains challenging. This is because these long-form answers are often free-form, making rule-based verification ineffective. Therefore, instead of directly evaluating the faithfulness of the long-form response, we propose a proxy reward to evaluate it implicitly, as the faithfulness of a long-form answer can be measured by its ability to drive the LLMs toward a correct short-form answer. Specifically, for each generated long-form answer y_{lf}, we replace the given context c with it as new input and feed it to the LLM to check whether the LLM can produce the correct short-form answer based on y_{lf}. If the generated long-form response can enable the LLM to generate the correct answer, it indicates that the long-form response stays faithful to the context, contains complete and easy-to-understand sentences, and correctly addresses the question. Thus, we assign a reward score of 1 for the positive long-form response that helps the LLM to produce the correct final answer, and a reward score of 0 for those that lead to an incorrect answer.

Format Reward. We also include a format reward that encourages adherence to a predefined output structure (e.g., using <think>, <long_answer>, and <short_answer>tags). Outputs that conform to this pattern receive a reward boost, thereby enhancing consistency. We use string matching to evaluate whether the generated responses adhere to the format, giving a score of 1 for a match and 0 for a mismatch.

Finally, we use the sum of these three rewards as the final composite reward. It enhances the efficacy of the RL training framework, guiding the model toward generating more faithful responses in both short-form and long-form tasks. More details are shown in the Appendix B.

3 Experiments

In this section, we conduct experiments and provide analyses to justify the effectiveness of CANOE.

3.1 Tasks and Datasets

To evaluate our method CANOE comprehensively, we select a range of downstream datasets, including short-form and long-form generation tasks.

Short-form Generation Tasks. For short-form generation tasks, we use two counterfactual QA datasets, including ConFiQA (Bi et al. 2024) and CNQ (Longpre et al. 2021), a multiple-choice questions dataset FaithEval (Ming et al. 2025), and a factual QA dataset FiQA (Bi et al. 2024) that is the factual version of ConFiQA. These datasets ensure the answers appear in the contexts to evaluate the faithfulness. We also evaluate our method on four open-domain QA datasets within the FollowRAG benchmark (Dong et al. 2024) to evaluate the abilities of LLMs in real-world retrieval-augmented generation (RAG) scenarios, including NaturalQA (Kwiatkowski et al. 2019b), TriviaQA (Joshi et al. 2017), HotpotQA (Yang et al. 2018), and WebQSP (Yih et al. 2016). In real-world RAG scenarios, the answer may not appear in the retrieved passages, and these passages tend to be noisy. We evaluate models based on whether gold answers are included in the generated responses (i.e., Acc) and exact matching (EM) for QA tasks. For multiple-choice questions, we follow Ming et al. (2025) and use keyword matching to verify the accuracy.

Long-form Generation Tasks. We include a text summarization task XSum (Narayan, Cohen, and Lapata 2018), a text simplification task WikiLarge (Zhang and Lapata 2017), and a long-form QA task CLAPNQ (Rosenthal et al. 2025). To evaluate the faithfulness of generated long-form answers, called FaithScore (FS), we use MiniCheck (Tang, Laban, and Durrett 2024) to check whether the model response is grounded in the provided context. MiniCheck is a state-of-the-art method to recognize if LLM output can be grounded in given contexts. If the model response contains at least one statement that cannot be inferred from the context, we consider it as a negative response; otherwise, it is a positive response. We also query GPT-4o to evaluate the quality of generated responses, namely QualityScore.

More details can be found in the Appendix C.

3.2 Baselines and Implementation Details

Baselines. We compare several baselines, including (1) **Vanilla LLMs**: including LLaMA-3-Instruct (Meta 2024) and Qwen-2.5-Instruct (Yang et al. 2024) of different sizes. We also conduct supervised fine-tuning on synthesized 10,000 short-form data as SFT baselines. (2) **SOTA LLMs**:

Model	Short-form Generation Tasks									Long-form Generation Tasks			Avg. Score	
	ConFiQA		FiQA		CNQ		FaithEval	FollowRAG		XSum	WikiLarge	CLAPNQ		
	EM	Acc	EM	Acc	EM	Acc	Acc	EM	Acc	FS	FS	FS	Avg EM	Avg Acc
The state-of-the-art LLMs														
GPT-4o	31.5	42.7	66.8	79.6	43.4	55.9	47.5	42.2	57.8	80.7	88.1	70.3	58.8	65.3
GPT-4o mini	49.5	63.7	67.1	78.8	47.8	54.3	50.9	38.5	51.3	75.4	91.0	66.0	60.8	66.4
DeepSeek V3	49.5	58.6	67.0	76.5	54.6	67.3	51.0	37.7	55.2	82.8	85.6	71.0	62.4	68.5
Claude 3.7 Sonnet	26.0	36.0	56.4	72.2	41.4	65.0	45.6	36.3	53.7	78.3	81.7	68.3	54.3	62.6
OpenAI o1	49.0	57.9	78.0	89.7	29.5	39.1	52.0	40.5	57.0	81.0	88.1	68.0	60.8	66.6
DeepSeek R1	68.4	74.3	68.4	80.7	60.3	70.2	60.1	42.9	56.6	80.3	83.0	73.5	67.1	72.3
Claude 3.7 Sonnet-Thinking	27.1	38.7	59.5	76.7	42.1	67.0	57.0	38.8	55.3	79.0	81.4	72.2	57.1	65.9
LLaMA-3-Instruct Series														
LLaMA-3-Instruct-8B	49.2	58.2	11.4	59.3	37.8	45.2	52.0	31.1	44.8	64.2	77.1	58.5	47.7	57.4
LLaMA-3-Instruct-70B	38.1	54.5	9.1	66.8	54.2	65.0	50.9	38.7	45.7	72.0	77.4	47.2	48.5	59.9
SFT-8B	65.1	70.3	35.9	59.9	52.6	65.7	43.0	19.2	21.0	62.2	74.2	55.3	50.9	56.4
Context-DPO-8B	66.3	72.9	40.9	59.5	54.6	62.3	37.5	29.9	43.8	65.2	78.2	59.1	54.0	59.8
SCOPE$_{sum}$-8B	35.7	64.6	7.1	68.7	33.8	60.6	55.7	30.1	46.2	70.3	80.3	59.8	46.6	63.3
CANOE-LLaMA-8B	**73.5**	**80.9**	**82.7**	**84.9**	**66.7**	**73.4**	**74.6**	**40.9**	**51.7**	**74.4**	**84.4**	**64.9**	**70.3**	**73.6**
Δ Compared to Vanilla.	+24.3	+22.6	+71.3	+25.6	+28.9	+28.2	+22.6	+9.8	+6.9	+10.2	+7.3	+6.4	+22.6	+16.2
Qwen-2.5-Instruct Series														
Qwen-2.5-Instruct-7B	52.5	61.0	13.2	68.4	55.3	68.2	56.1	32.6	45.3	63.4	57.8	61.2	49.0	60.2
Qwen-2.5-Instruct-14B	34.1	47.3	0.8	61.4	43.1	64.3	51.6	34.8	51.2	68.2	82.3	63.4	47.3	61.2
Qwen-2.5-Instruct-32B	44.5	66.4	39.2	81.1	37.7	66.4	47.0	33.9	53.1	20.2	57.7	31.7	39.0	52.9
Qwen-2.5-Instruct-72B	43.7	52.3	4.8	67.3	51.8	62.2	45.2	38.5	55.7	71.2	90.4	64.8	51.3	63.6
SFT-7B	62.8	69.8	48.8	76.6	60.1	65.3	50.3	29.0	41.7	55.2	51.3	57.2	51.8	58.4
Context-DPO-7B	64.5	70.6	57.1	78.2	62.3	70.1	45.7	31.0	43.7	60.2	53.4	62.8	54.6	60.6
SCOPE$_{sum}$-7B	39.3	47.9	12.9	60.9	50.2	55.3	52.3	30.6	46.0	68.3	72.0	63.2	48.6	58.2
CANOE-Qwen-7B	**67.6**	**75.2**	**78.1**	**83.5**	**67.2**	**76.4**	**70.5**	**37.0**	**50.2**	**72.4**	**86.1**	**65.2**	**68.0**	**72.4**
Δ Compared to Vanilla.	+15.1	+14.2	+64.9	+15.0	+11.9	+8.2	+14.4	+4.4	+4.9	+9.0	+28.3	+4.0	+19.0	+12.3
CANOE-Qwen-14B	**85.7**	**87.4**	**87.8**	**88.5**	**81.8**	**84.2**	**67.4**	**46.1**	**54.6**	**75.7**	**91.1**	**68.4**	**75.5**	**77.2**
Δ Compared to Vanilla.	+51.6	+40.1	+87.0	+27.1	+38.7	+19.9	+15.8	+11.3	+3.4	+7.5	+8.8	+5.0	+28.2	+16.0

Table 1: Experimental results (%) on eleven datasets. The FollowRAG results represent the results averaged over these four open-domain QA datasets, as shown in the Appendix C, including NaturalQA, TriviaQA, HotpotQA, and WebQSP. **Bold** numbers indicate the best performance of models with the same model size. Avg EM/Acc represents the average score between short-form task metrics (EM/Acc) and long-form task metric FaithScore (FS).

We further evaluate the most advanced LLMs, including GPT-4o, GPT-4o-mini, OpenAI o1 (Jaech et al. 2024), Claude 3.7 Sonnet (Anthropic 2025), Claude 3.7 Sonnet-Thinking, DeepSeek R1, and DeepSeek V3 (DeepSeek-AI 2025a,b). (3) **The Methods to Improve Faithfulness**: Context-DPO (Bi et al. 2024) aligns LLMs through DPO with constructed faithful and unfaithful short-form answers, thus improving the faithfulness in short-form generation. SCOPE (Duong et al. 2025) proposes a pipeline to generate self-supervised task-specific data and applies preference training to enhance the faithfulness in a special task. We train it on the sampled training set of the summarization task XSum as SCOPE$_{sum}$, regarding it as the method designed to improve the faithfulness of long-form response generation.

Implementation Details. Our experiments are conducted on LLaMA-3-Instruct and Qwen-2.5-Instruct. Details are shown in Appendix D, e.g., hyperparameters.

3.3 Main Results

CANOE Improves the Faithfulness of LLMs in Both Short-form and Long-form Response Generation. As shown in Table 1, CANOE shows significant improvements on 11 faithfulness-related benchmarks. CANOE achieves substantial improvements in the overall score compared to original LLMs, e.g., **22.6%** for *Llama3-8B* and **19.0%** for *Qwen2.5-7B* in Avg EM score. CANOE also surpasses the most advanced LLMs (e.g., GPT-4o) in the overall score (both Avg EM and Avg Acc scores). This shows that CANOE effectively aligns LLMs to be context-faithful. Also, for real-world RAG scenarios, our proposed CANOE can also improve the performance even though the answer may not appear in the retrieved passages, and these passages are often noisy.

Model	XSum	WikiLarge	CLAPNQ	Avg
GPT-4o	98.5	97.5	81.2	92.4
LLaMA-3-Instruct-8B	70.9	82.9	39.2	64.3
LLaMA-3- Instruct-70B	86.2	83.0	30.1	66.4
CANOE-LLaMA-8B	**85.8**	**87.8**	**65.5**	**79.7**
Qwen-2.5-Instruct-7B	79.4	79.0	64.6	74.3
Qwen-2.5-Instruct-14B	90.5	83.1	63.6	79.1
Qwen-2.5-Instruct-32B	90.3	83.9	58.6	77.6
Qwen-2.5-Instruct-72B	95.7	94.1	75.4	88.4
CANOE-Qwen-7B	**91.5**	**87.3**	**68.2**	**82.3**
CANOE-Qwen-14B	**91.9**	**89.7**	**73.5**	**85.0**

Table 2: QualityScore (%) on long-form generation tasks.

CANOE Maintains the Factuality of LLMs. We further evaluate whether CANOE will reduce the factuality of LLMs. Following Ming et al. (2025), we modify the original FaithEval and make it a closed-book QA setting, where no context is provided and LLMs need to give factual answers. In this case, the models rely entirely on their parametric knowledge of common facts, and we find that our proposed CANOE maintains the factuality compared to the untuned LLM as shown in Figure 3. However, when a new context with counterfactual

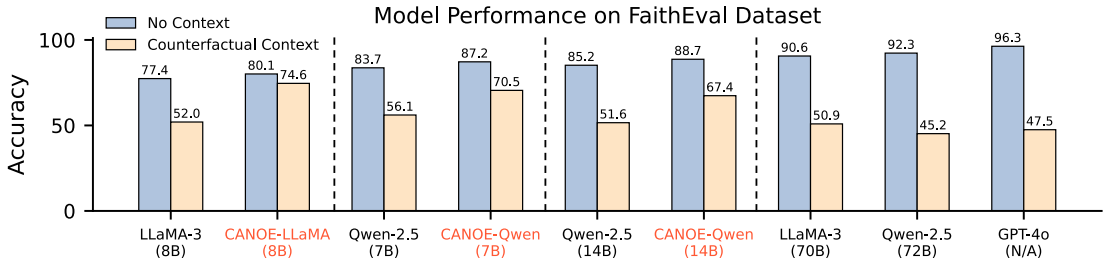

Figure 3: Model performance comparison on FaithEval in a closed-book QA setting and counterfactual context setting. Our models are colored in orange. We report the results from the chat version of LLaMA-3 and Qwen-2.5.

Model	Acc			EM		
	QA	MR	MC	QA	MR	MC
GPT-4o	52.2	45.6	30.3	43.3	32.4	18.7
LLaMA-3-Instruct-8B	69.7	55.9	49.1	60.0	47.9	39.6
CANOE-LLaMA-8B	**82.7**	**80.1**	**79.8**	**76.4**	**73.5**	**70.5**
Qwen-2.5-Instruct-7B	72.8	59.1	51.1	64.9	50.2	42.5
Qwen-2.5-Instruct-14B	62.4	44.9	34.7	44.7	34.3	23.3
Qwen-2.5-Instruct-32B	74.1	65.9	59.3	55.9	42.8	34.8
Qwen-2.5-Instruct-72B	63.3	50.3	43.3	54.3	42.2	34.7
CANOE-Qwen-7B	**79.5**	**76.1**	**70.1**	**73.3**	**67.9**	**61.7**
CANOE-Qwen-14B	**91.8**	**86.4**	**84.1**	**89.7**	**85.2**	**82.1**

Table 3: Results (%) on three tasks in ConFiQA.

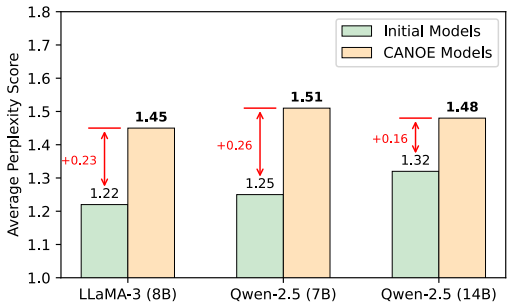

Figure 4: The average perplexity score of 110 negative samples for each model from eleven datasets.

Model	Short-form Tasks		Long-form Tasks	
	EM	Acc	FaithScore	QualityScore
CANOE-LLaMA-8B	**67.7**	**73.1**	**74.6**	**79.7**
-w/o. Our Method	36.3	51.9	66.6	64.3
-w/o. Dual-GRPO	60.5	66.6	N/A	23.5
-w/o. Reasoning-required Data.	63.7	69.4	71.7	75.3
-w/o. Inconsistent Data.	64.4	70.2	70.2	72.5
-w/o. Counterfactual Data.	62.6	67.8	69.7	73.7

Table 4: Results (%) of ablation study. EM/Acc in short-form tasks represents the average score between QA metrics (EM/Acc) and the accuracy of FaithEval. N/A means a false generation pattern hacks this metric.

evidence that contradicts the model's parametric knowledge is introduced, performance declines sharply. For example, GPT-4o achieves 96.3% accuracy on factual closed-book QA task but only 47.5% on counterfactual QA task that evaluates the faithfulness of LLMs. This highlights that, unlike factuality, the faithfulness of LLMs is difficult to improve by simply scaling model parameters, which further indicates the necessity of a post-training method to improve faithfulness.

CANOE Improves the Quality of Long-form Response Generation. As shown in Table 2, we can find that our proposed CANOE consistently improves the generation quality in the three long-form tasks. This is because the proxy reward implicitly requires LLMs to generate easy-to-understand responses, which further optimizes the response quality.

CANOE Enhances LLMs' Reasoning in Short-form Response Generation. ConFiQA consists of three different tasks: question answering (QA), multi-hop reasoning (MR), and multi-conflicts reasoning (MC). QA focuses on the single-hop task with context containing one corresponding answer, while MR and MC involve multi-hop reasoning tasks with context containing one and multiple related counterfactual contexts, respectively. As shown in Table 3, CANOE not only improves the faithfulness in the single-hop QA task but also enhances the reasoning ability in reasoning tasks.

CANOE Mitigates Overconfidence Bias. For each model, we select a total of 110 unfaithful samples with the highest perplexity from the 11 datasets, 10 samples per dataset. Then we report the average perplexity score on these negative samples shown in Figure 4. We can find that CANOE produces the high perplexity scores, indicating low confidence scores, for these bad cases. This shows that CANOE mitigates overconfidence in these false statements.

3.4 Analysis

Ablation Study. We conduct an ablation study in Table 4. The result reveals that our proposed CANOE (including Dual-GRPO and the designed short-form data synthesis) significantly improves the faithfulness of LLMs in both short-form and long-form generation tasks. For Dual-GRPO, we observe that directly applying GRPO to synthesized short-form data leads to over-optimizing short-form generation and a false response generation pattern. We find that tuned models tend to directly copy text spans from the given context as the final answer instead of following instructions in long-form generation tasks (we show the case study in Table 5 to visually show this phenomenon). Thus, the generated responses do not contain syntactically and semantically complete sentences for

Text Summarization Task

Inputs: {Documents}

Instruction: Refer to the passage above and provide a summary as the response.

GRPO-LLaMA-8B: Comedy clubs. (✗)

CANOE-LLaMA-8B: Kareem Badr and his friends were able to turn around the struggling comedy club, the Hideout, by applying their skills in problem-solving and efficiency. They broke down the business into small chunks and figured out how to make each aspect better and more efficient. This approach helped them to improve the club's performance and eventually expand it. The passage also mentions that the growing popularity of stand-up comedy is fueled in part by younger audiences, who view humorous TV hosts as their primary source for news. Many comedy clubs now require talent, and there is a career and a living to be made by good comedians who are not superstars. The clubs also make money from selling drinks and dinners, as well as from cover charges and talent fees. In addition, digital media has given live comedy a boost, with one terrific joke being able to get a million hits online. (✓)

Figure 5: Case study between the models trained via GRPO and our model for the summarization task.

Question: What do Greeks say when they break plates?

Qwen-2.5-Instruct-7B: Based on the information provided in the passages, there is no specific mention of what exactly Greeks say when they break plates during the plate-breaking ceremony. (✗)

CANOE-Qwen-7B: According to the information given, in the Three's Company episode "Opening Night," when Jack breaks the plates at the end of the celebration, he exclaims "Opa!".(✓) This suggests that "Opa!" might be a phrase commonly used in Greek culture during such festivities. (✓) "Opa!" is often used to express joy, excitement, or approval, which aligns with the celebratory nature of plate smashing in Greek traditions. (✓)

Figure 6: Case study from long-form QA task CLAPNQ. For different useful statements, we use different colors.

long-form generation tasks, which leads to low QualityScore performance and also invalidates the metric used for evaluating faithfulness. For the designed QA tasks used to ensure the complexity and diversity of training data, we replace the designed Reasoning-required Context and Inconsistent Context samples with the same number of samples that contain Straightforward Context. We find that involving these more challenging instances can improve the effectiveness of RL training. We also replace the data points that contain Counterfactual Context with the same number of factual samples. The designed Counterfactual Context improves the final performance as it prevents models from depending on their learned factual knowledge to find the right answers.

Case Study. We find that directly applying GRPO instead of our Dual-GRPO to synthesized short-form data leads to over-optimizing short-form generation and a false response generation pattern. As shown in Figure 5, we find that the tuned model GRPO-LLaMA-8B tends to copy text spans from the given context as the final answer instead of following instructions in the summarization task. We also observe the same phenomenon in the tasks of text simplification and long-form QA; more details can be found in the Appendix F. However, when we apply Dual-GPRO, we find that trained models can generate fluent and complete sentences. Thus, Dual-GRPO not only improves the faithfulness of LLMs in two types of response generation but also ensures the utility. We also conduct a case study in Figure 6 to show the advantages of CANOE. CANOE ensures the statements are faithful and comprehensive, and the text flows naturally.

4 Related Work

Recently, the demand for utilizing LLMs to generate coherent text responses based on the provided contexts has continued to grow (Si et al. 2022, 2023; Zhao et al. 2025), e.g., retrieval-augmented generation and text summarization. However, LLMs are often criticized for generating outputs that deviate from the provided contents, namely *faithfulness hallucination* (Huang et al. 2024). Many approaches have been proposed to improve the faithfulness of language models (Si et al. 2025a; Wang et al. 2025). The first line of work focuses on the inference stage, such as designing prompts to encourage context integration (Zhou et al. 2023) and improving context quality via explicit denoising (Xu, Shi, and Choi 2024). Although effective, these approaches primarily serve as a compensatory way rather than enabling the model to inherently learn to prevent generating unfaithful responses. Thus, many studies attempt to apply post-training methods to improve the faithfulness. Bi et al. (2024) utilizes constructed faithful and unfaithful short-form completions and applies DPO to align LLMs to be context-faithful in short-form QA tasks. Huang et al. (2025) trains LLMs to discriminate between faithful and unfaithful responses in long-form QA tasks by unfaithful response synthesis and contrastive tuning. Li et al. (2025) introduces a self-evolving framework to maintain the faithfulness of LLMs in long-form QA tasks through iterative self-improvement. Duong et al. (2025) proposes a pipeline to generate a self-supervised task-specific dataset and applies preference training to enhance the faithfulness for a special task. However, these methods struggle to consistently improve the faithfulness of LLMs across various tasks, as these methods are designed for specific tasks.

5 Conclusion

In this paper, we propose CANOE, a systematic post-training method for teaching LLMs to remain faithful in both short-form and long-form generation tasks without human annotations. By synthesizing diverse short-form QA data and introducing Dual-GRPO, a tailored RL method with three well-designed rule-based rewards, CANOE effectively improves the faithfulness of LLMs. We first synthesize short-form QA data with four diverse tasks to construct high-quality and easily verifiable training data without human annotation. We then propose Dual-GRPO, a rule-based RL method that includes three tailored rule-based rewards derived from synthesized short-form QA data, while optimizing both short-form and long-form response generation simultaneously. Experimental results show that CANOE consistently improves the faithfulness of LLMs across diverse downstream tasks.

Acknowledgment

This work is supported by Beijing Municipal Science and Technology Plan Project (Z241100001324025).

References

Anthropic. 2025. Claude 3.7 Sonnet System Card.

Bai, Y.; Lv, X.; Zhang, J.; Lyu, H.; Tang, J.; Huang, Z.; Du, Z.; Liu, X.; Zeng, A.; Hou, L.; Dong, Y.; Tang, J.; and Li, J. 2023. LongBench: A Bilingual, Multitask Benchmark for Long Context Understanding. *arXiv preprint arXiv:2308.14508*.

Bi, B.; Huang, S.; Wang, Y.; Yang, T.; Zhang, Z.; Huang, H.; Mei, L.; Fang, J.; Li, Z.; Wei, F.; Deng, W.; Sun, F.; Zhang, Q.; and Liu, S. 2024. Context-DPO: Aligning Language Models for Context-Faithfulness. arXiv:2412.15280.

Clark, P.; Cowhey, I.; Etzioni, O.; Khot, T.; Sabharwal, A.; Schoenick, C.; and Tafjord, O. 2018. Think you have Solved Question Answering? Try ARC, the AI2 Reasoning Challenge. arXiv:1803.05457.

Cui, W.; Zhou, M.; Zhao, R.; and Norouzi, N. 2019. KB-NLG: From Knowledge Base to Natural Language Generation. In Axelrod, A.; Yang, D.; Cunha, R.; Shaikh, S.; and Waseem, Z., eds., *Proceedings of the 2019 Workshop on Widening NLP*, 80–82. Florence, Italy: Association for Computational Linguistics.

DeepSeek-AI. 2025a. DeepSeek-R1: Incentivizing Reasoning Capability in LLMs via Reinforcement Learning. arXiv:2501.12948.

DeepSeek-AI. 2025b. DeepSeek-V3 Technical Report. arXiv:2412.19437.

Dong, G.; Song, X.; Zhu, Y.; Qiao, R.; Dou, Z.; and Wen, J.-R. 2024. Toward General Instruction-Following Alignment for Retrieval-Augmented Generation. arXiv:2410.09584.

Dong, X.; Li, W.; Le, Y.; Jiang, Z.; Zhong, J.; and Wang, Z. 2025. TermDiffuSum: A Term-guided Diffusion Model for Extractive Summarization of Legal Documents. In Rambow, O.; Wanner, L.; Apidianaki, M.; Al-Khalifa, H.; Eugenio, B. D.; and Schockaert, S., eds., *Proceedings of the 31st International Conference on Computational Linguistics*, 3222–3235. Abu Dhabi, UAE: Association for Computational Linguistics.

Duong, S.; Bronnec, F. L.; Allauzen, A.; Guigue, V.; Lumbreras, A.; Soulier, L.; and Gallinari, P. 2025. SCOPE: A Self-supervised Framework for Improving Faithfulness in Conditional Text Generation. In *The Thirteenth International Conference on Learning Representations*.

He, W.; Liu, K.; Liu, J.; Lyu, Y.; Zhao, S.; Xiao, X.; Liu, Y.; Wang, Y.; Wu, H.; She, Q.; Liu, X.; Wu, T.; and Wang, H. 2018. DuReader: a Chinese Machine Reading Comprehension Dataset from Real-world Applications. In Choi, E.; Seo, M.; Chen, D.; Jia, R.; and Berant, J., eds., *Proceedings of the Workshop on Machine Reading for Question Answering*, 37–46. Melbourne, Australia: Association for Computational Linguistics.

Huang, L.; Feng, X.; Ma, W.; Fan, Y.; Feng, X.; Ye, Y.; Zhong, W.; Gu, Y.; Wang, B.; Wu, D.; Hu, G.; and Qin, B. 2025. Improving Contextual Faithfulness of Large Language Models via Retrieval Heads-Induced Optimization. arXiv:2501.13573.

Huang, L.; Yu, W.; Ma, W.; Zhong, W.; Feng, Z.; Wang, H.; Chen, Q.; Peng, W.; Feng, X.; Qin, B.; and Liu, T. 2024. A Survey on Hallucination in Large Language Models: Principles, Taxonomy, Challenges, and Open Questions. *ACM Trans. Inf. Syst.* Just Accepted.

Jaech, A.; Kalai, A.; Lerer, A.; Richardson, A.; El-Kishky, A.; Low, A.; Helyar, A.; Madry, A.; Beutel, A.; Carney, A.; et al. 2024. Openai o1 system card. *arXiv preprint arXiv:2412.16720*.

Joshi, M.; Choi, E.; Weld, D.; and Zettlemoyer, L. 2017. TriviaQA: A Large Scale Distantly Supervised Challenge Dataset for Reading Comprehension. In Barzilay, R.; and Kan, M.-Y., eds., *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, 1601–1611. Vancouver, Canada: Association for Computational Linguistics.

Kwiatkowski, T.; Palomaki, J.; Redfield, O.; Collins, M.; Parikh, A.; Alberti, C.; Epstein, D.; Polosukhin, I.; Devlin, J.; Lee, K.; Toutanova, K.; Jones, L.; Kelcey, M.; Chang, M.-W.; Dai, A. M.; Uszkoreit, J.; Le, Q.; and Petrov, S. 2019a. Natural Questions: A Benchmark for Question Answering Research. *Transactions of the Association for Computational Linguistics*, 7: 452–466.

Kwiatkowski, T.; Palomaki, J.; Redfield, O.; Collins, M.; Parikh, A.; Alberti, C.; Epstein, D.; Polosukhin, I.; Kelcey, M.; Devlin, J.; Lee, K.; Toutanova, K. N.; Jones, L.; Chang, M.-W.; Dai, A.; Uszkoreit, J.; Le, Q.; and Petrov, S. 2019b. Natural Questions: a Benchmark for Question Answering Research. *Transactions of the Association of Computational Linguistics*.

Li, K.; Zhang, T.; Li, Y.; Luo, H.; Moustafa, A.; Wu, X.; Glass, J.; and Meng, H. 2025. Generate, Discriminate, Evolve: Enhancing Context Faithfulness via Fine-Grained Sentence-Level Self-Evolution. arXiv:2503.01695.

Longpre, S.; Perisetla, K.; Chen, A.; Ramesh, N.; DuBois, C.; and Singh, S. 2021. Entity-Based Knowledge Conflicts in Question Answering. In Moens, M.-F.; Huang, X.; Specia, L.; and Yih, S. W.-t., eds., *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, 7052–7063. Online and Punta Cana, Dominican Republic: Association for Computational Linguistics.

Meta. 2024. The Llama 3 Herd of Models. arXiv:2407.21783.

Ming, Y.; Purushwalkam, S.; Pandit, S.; Ke, Z.; Nguyen, X.-P.; Xiong, C.; and Joty, S. 2025. FaithEval: Can Your Language Model Stay Faithful to Context, Even If "The Moon is Made of Marshmallows". In *The Thirteenth International Conference on Learning Representations*.

Narayan, S.; Cohen, S. B.; and Lapata, M. 2018. Don't Give Me the Details, Just the Summary! Topic-Aware Convolutional Neural Networks for Extreme Summarization. In Riloff, E.; Chiang, D.; Hockenmaier, J.; and Tsujii, J., eds., *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, 1797–1807. Brussels, Belgium: Association for Computational Linguistics.

OpenAI. 2023. GPT-4 Technical Report. *arXiv preprint arXiv:2303.08774*.

Rafailov, R.; Sharma, A.; Mitchell, E.; Manning, C. D.; Ermon, S.; and Finn, C. 2023. Direct Preference Optimization: Your Language Model is Secretly a Reward Model. In

Thirty-seventh Conference on Neural Information Processing Systems.

Ravichander, A.; Ghela, S.; Wadden, D.; and Choi, Y. 2025. HALoGEN: Fantastic LLM Hallucinations and Where to Find Them. arXiv:2501.08292.

Rosenthal, S.; Sil, A.; Florian, R.; and Roukos, S. 2025. CLAPnq: Cohesive Long-form Answers from Passages in Natural Questions for RAG systems. *Transactions of the Association for Computational Linguistics*, 13: 53–72.

Schulman, J.; Wolski, F.; Dhariwal, P.; Radford, A.; and Klimov, O. 2017. Proximal Policy Optimization Algorithms. arXiv:1707.06347.

Shao, Z.; Wang, P.; Zhu, Q.; Xu, R.; Song, J.; Bi, X.; Zhang, H.; Zhang, M.; Li, Y. K.; Wu, Y.; and Guo, D. 2024. DeepSeekMath: Pushing the Limits of Mathematical Reasoning in Open Language Models. arXiv:2402.03300.

Si, S.; Ma, W.; Gao, H.; Wu, Y.; Lin, T.-E.; Dai, Y.; Li, H.; Yan, R.; Huang, F.; and Li, Y. 2023. SpokenWOZ: A Large-Scale Speech-Text Benchmark for Spoken Task-Oriented Dialogue Agents. In Oh, A.; Neumann, T.; Globerson, A.; Saenko, K.; Hardt, M.; and Levine, S., eds., *Advances in Neural Information Processing Systems*, volume 36, 39088–39118. Curran Associates, Inc.

Si, S.; Zeng, S.; Lin, J.; and Chang, B. 2022. SCL-RAI: Span-based Contrastive Learning with Retrieval Augmented Inference for Unlabeled Entity Problem in NER. In Calzolari, N.; Huang, C.-R.; Kim, H.; Pustejovsky, J.; Wanner, L.; Choi, K.-S.; Ryu, P.-M.; Chen, H.-H.; Donatelli, L.; Ji, H.; Kurohashi, S.; Paggio, P.; Xue, N.; Kim, S.; Hahm, Y.; He, Z.; Lee, T. K.; Santus, E.; Bond, F.; and Na, S.-H., eds., *Proceedings of the 29th International Conference on Computational Linguistics*, 2313–2318. Gyeongju, Republic of Korea: International Committee on Computational Linguistics.

Si, S.; Zhao, H.; Chen, G.; Gao, C.; Bai, Y.; Wang, Z.; An, K.; Luo, K.; Qian, C.; Qi, F.; Chang, B.; and Sun, M. 2025a. Aligning Large Language Models to Follow Instructions and Hallucinate Less via Effective Data Filtering. In Che, W.; Nabende, J.; Shutova, E.; and Pilehvar, M. T., eds., *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, 16469–16488. Vienna, Austria: Association for Computational Linguistics. ISBN 979-8-89176-251-0.

Si, S.; Zhao, H.; Chen, G.; Li, Y.; Luo, K.; Lv, C.; An, K.; Qi, F.; Chang, B.; and Sun, M. 2025b. GATEAU: Selecting Influential Samples for Long Context Alignment. In Christodoulopoulos, C.; Chakraborty, T.; Rose, C.; and Peng, V., eds., *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, 7391–7422. Suzhou, China: Association for Computational Linguistics. ISBN 979-8-89176-332-6.

Tang, L.; Laban, P.; and Durrett, G. 2024. MiniCheck: Efficient Fact-Checking of LLMs on Grounding Documents. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics.

Vrandečić, D.; and Krötzsch, M. 2014. Wikidata: a free collaborative knowledgebase. *Commun. ACM*, 57(10): 78–85.

Wang, H.; Prasad, A.; Stengel-Eskin, E.; and Bansal, M. 2025. Retrieval-Augmented Generation with Conflicting Evidence. arXiv:2504.13079.

Wu, H.; Zhan, M.; Tan, H.; Hou, Z.; Liang, D.; and Song, L. 2023. VCSUM: A Versatile Chinese Meeting Summarization Dataset. In Rogers, A.; Boyd-Graber, J.; and Okazaki, N., eds., *Findings of the Association for Computational Linguistics: ACL 2023*, 6065–6079. Toronto, Canada: Association for Computational Linguistics.

Xu, F.; Shi, W.; and Choi, E. 2024. RECOMP: Improving Retrieval-Augmented LMs with Context Compression and Selective Augmentation. In *The Twelfth International Conference on Learning Representations*.

Yang, A.; Yang, B.; Zhang, B.; Hui, B.; Zheng, B.; Yu, B.; Li, C.; Liu, D.; Huang, F.; Wei, H.; Lin, H.; Yang, J.; Tu, J.; Zhang, J.; Yang, J.; Yang, J.; Zhou, J.; Lin, J.; Dang, K.; Lu, K.; Bao, K.; Yang, K.; Yu, L.; Li, M.; Xue, M.; Zhang, P.; Zhu, Q.; Men, R.; Lin, R.; Li, T.; Xia, T.; Ren, X.; Ren, X.; Fan, Y.; Su, Y.; Zhang, Y.; Wan, Y.; Liu, Y.; Cui, Z.; Zhang, Z.; and Qiu, Z. 2024. Qwen2.5 Technical Report. *arXiv preprint arXiv:2412.15115*.

Yang, Z.; Qi, P.; Zhang, S.; Bengio, Y.; Cohen, W.; Salakhutdinov, R.; and Manning, C. D. 2018. HotpotQA: A Dataset for Diverse, Explainable Multi-hop Question Answering. In Riloff, E.; Chiang, D.; Hockenmaier, J.; and Tsujii, J., eds., *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, 2369–2380. Brussels, Belgium: Association for Computational Linguistics.

Yih, W.; Richardson, M.; Meek, C.; Chang, M.-W.; and Suh, J. 2016. The Value of Semantic Parse Labeling for Knowledge Base Question Answering. In *Annual Meeting of the Association for Computational Linguistics*.

Zhang, H.; Yu, P. S.; and Zhang, J. 2024. A Systematic Survey of Text Summarization: From Statistical Methods to Large Language Models. arXiv:2406.11289.

Zhang, X.; and Lapata, M. 2017. Sentence Simplification with Deep Reinforcement Learning. In Palmer, M.; Hwa, R.; and Riedel, S., eds., *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, 584–594. Copenhagen, Denmark: Association for Computational Linguistics.

Zhao, H.; Si, S.; Chen, L.; Zhang, Y.; Sun, M.; Chang, B.; and Zhang, M. 2025. Looking Beyond Text: Reducing Language Bias in Large Vision-Language Models via Multimodal Dual-Attention and Soft-Image Guidance. In Christodoulopoulos, C.; Chakraborty, T.; Rose, C.; and Peng, V., eds., *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, 19677–19701. Suzhou, China: Association for Computational Linguistics. ISBN 979-8-89176-332-6.

Zhou, W.; Zhang, S.; Poon, H.; and Chen, M. 2023. Context-faithful Prompting for Large Language Models. In Bouamor, H.; Pino, J.; and Bali, K., eds., *Findings of the Association for Computational Linguistics: EMNLP 2023*, 14544–14556. Singapore: Association for Computational Linguistics.