REFO: Reinforced Evolutionary Faithfulness Optimization for Large Language Models

Yi Wang*, Xiaqiang Tang*, Keyu Hu, Haojie Lu, Sihong Xie†

The Hong Kong University of Science and Technology (Guangzhou)
sihongxie@hkust-gz.edu.cn

Abstract

Despite its success in enriching LLMs with external knowledge, RAG remains plagued by faithfulness hallucinations, where generated text contradicts the retrieved source information. Previous research on faithfulness hallucination in LLMs is frequently hindered by prohibitive manual annotation costs and a dependency on static datasets, which caps their performance and adaptability. Furthermore, these models lack a clear training mechanism to explicitly promote contextual focus. In this work, we propose a novel iterative self-evolution framework to enhance model faithfulness. This framework autonomously generates high-quality data and leverages it for the continuous self-optimization of the model, leading to significant improvements in faithfulness. Our experimental analysis reveals that improving model faithfulness encourages a closer alignment of the attention distribution with the given context. Based on this finding, we design an attention-based loss function to further promote this process. Experimental results show that our model achieves state-of-the-art faithfulness on a range of context-based question-answering datasets, marking a significant advancement over previous approaches.

Code & Datasets — https://github.com/chkwy/REFO

Introduction

The widespread adoption of Retrieval-Augmented Generation (RAG) (Lewis et al. 2020) has made it crucial for Large Language Models (LLMs) to be able to generate precise and consistent answers based on the provided information (Song et al. 2025, 2024; Niu et al. 2024), especially in scenarios where their internal parametric knowledge is insufficient or outdated. As a result, faithfulness hallucination (Zhou et al. 2023) has emerged as a critical challenge, especially within the nascent field of Context Engineering (Mei et al. 2025), where such failures undermine verifiability, pose a risk of cascading errors in agentic systems (Zhang et al. 2025), and violate the core tenets of traceability and auditability in high-stakes domains.

While post-training methods like Supervised Fine-Tuning (SFT)(Liu et al. 2025) and Direct Preference Optimization (DPO)(Bi et al. 2025) have been explored to enhance faithfulness, they face two fundamental limitations. First, they typically rely on static, offline datasets, which inherently limits their performance ceiling and adaptability. Second, they often lack an explicit mechanism to guide the model's focus, optimizing for preference alignment rather than directly for faithfulness to the context. Although a recent study has proposed a self-supervised approach(Duong et al. 2025), it is constrained by the inherent limitations of the model itself, leading to the generation of low-quality data, as the method lacks an effective way to improve its quality.

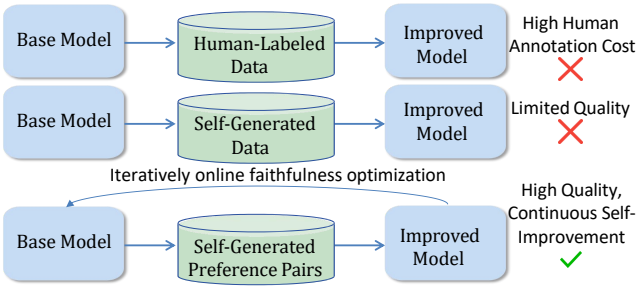

Figure 1: Comparison of three training paradigms. Human annotation is costly, while self-generation is limited by model capability. Our iterative method, however, continuously improves data quality.

Given these challenges, we seek a fully automated method that continuously generates progressively higher-quality data to improve model faithfulness. We draw inspiration from iterative evolution(Gulcehre et al. 2023; Singh et al. 2024) to propose a framework capable of iteratively optimizing the model and improving data generation quality.

Specifically, we introduce **REFO**, an iterative architecture that achieves self-improvement by learning directly from its own generated responses. This approach operates independently of pre-existing, human-annotated datasets, enabling two key functions: i) automatically curating high-quality data, and ii) leveraging that data to continuously enhance its own faithfulness via self-training.

This paper is the first to demonstrate the effectiveness of iterative evolution in the domain of hallucination mitigation. We introduce a novel nested iterative framework.

*These authors contributed equally.
†Corresponding author.

In this framework, the first step leverages the discrepancy between external and parametric knowledge as a reward signal, while the second step employs the output differences between successive model generations. To instantiate this reward mechanism, we utilize a dedicated faithfulness evaluation model as a scorer and train the model using Direct Preference Optimization (DPO)(Rafailov et al. 2023). This approach significantly enhances the model's faithfulness.

Furthermore, building on the observation that an improvement in faithfulness correlates with the model's enhanced attention to context(Chuang et al. 2024), we theoretically incorporate an attention mechanism into the DPO loss function. To this end, we design a novel loss objective hypothesized to promote this process. Our experiments subsequently validate that this approach significantly accelerates its faithfulness training efficiency.

Our contributions are threefold:

- We propose REFO, the first self-evolving optimization framework that improves LLM faithfulness reducing reliance on offline annotated datasets.
- We introduce a novel attention-guided loss that effectively directs the model's generation process towards the provided context to enhance faithfulness, based on the theoretical analysis that the improvement in model faithfulness is related to allocating more attention over the context.
- We conduct comprehensive evaluations across diverse LLMs and benchmarks, demonstrating that REFO achieves state-of-the-art performance in faithfulness.

Related Work

Faithfulness Hallucination of LLM

RAG enhances LLM by retrieving external documents, demonstrating significant success in applications like open-domain question answering(Han et al. 2024). Nevertheless, "unfaithful generation" outputs that are unsupported by or even contradict the retrieved evidence—persists as a key challenge.(Chen, Zhang, and Choi 2022; Holtzman et al. 2021) This problem often stems from the model's tendency to prioritize its internal parametric knowledge over the provided external evidence, which undermines system reliability by producing factually inconsistent outputs. Within the emerging paradigm of "context engineering"(Pajo 2025) which builds upon RAG to optimize information payloads for LLMs, ensuring output faithfulness has become a central concern.

Existing approaches to mitigate faithfulness hallucination face notable drawbacks. Post-training adjustment methods, while capable of optimizing models via supervised fine-tuning(Hu et al. 2022) or preference alignment(Rafailov et al. 2023) (e.g., constructing preference pairs), demand costly supervision(Liu et al. 2025; Song et al. 2025, 2024; Bi et al. 2025) and are susceptible to catastrophic forgetting(Kirkpatrick et al. 2017), which undermines generalization. In contrast, decoding strategy-based methods employ techniques like contrastive decoding at inference time to bolster faithfulness(Shi et al. 2024). Although these methods are training-free, they incur a substantial increase in computational overhead. More recently, self-supervised training methods have been proposed(Duong et al. 2025), however, ensuring the quality and fidelity of the preference data remains a critical challenge.

Self-Training of LLM

Self-training is a well-established semi-supervised learning methodology that leverages unlabeled data to improve model performance(Scudder 1965). In the context of LLM, this capability enables models to self-generate labels, thereby creating self-learning approaches that eliminate the reliance on computationally expensive teacher models with high API costs.

Iterative self-improvement methods fine-tune LLMs on generated data, such as ReST (Gulcehre et al. 2023) using an external reward model and ReST^{EM} (Singh et al. 2024) using a verifier. Other approaches rely on the LLM's own solutions, like STaR (Zelikman et al. 2022) and rejection fine-tuning (Yuan et al. 2023). More advanced methods, including ReST-MCTS* (Zhang et al. 2024) and rStar_Math (Guan et al. 2025), integrate MCTS for process-reward guidance. Attempts at self-evolution, such as (Li et al. 2025), also exist but often rely heavily on human-labeled data, preventing them from constituting a complete and autonomous self-evolutionary cycle. A common limitation of these methods is their reliance on fine-tuning only on correct answers, which is inefficient and discards valuable information from the more abundant incorrect responses.

Another family of methods evolves iteratively by generating preference pairs. For example, approaches like Self-Rewarding Language Models (Yuan et al. 2024) and Iterative Reasoning Preference Optimization (Pang et al. 2024) apply the DPO algorithm (Rafailov et al. 2023) in each iteration on newly generated data. (Wang et al. 2023) introduced a contrastive loss function to increase the likelihood of correct solutions. However, these methods typically use only data generated by the current model, effectively operating in an on-policy manner. This wastes valuable data from previous iterations and forfeits the off-policy advantages inherent to algorithms like DPO.

In this paper, we propose a novel framework that fundamentally differs by leveraging the distributional discrepancy between historical and current model generations as a dense reward signal. Unlike standard approaches that discard past experiences, our method incorporates an off-policy replay mechanism to fully exploit the exploration trajectories from previous iterations. By dynamically contrasting self-generated positive and negative reasoning paths, we establish a robust, self-sustaining evolutionary loop. This approach not only significantly maximizes sample efficiency by recycling "waste" data but also autonomously steers the model towards higher faithfulness, effectively circumventing the inefficiencies and external dependencies characteristic of current methods.

Method

This section details our proposed REFO framework, a novel approach for iteratively boosting model faithfulness, as depicted in Figure 2. We also analyze the principles of faithful-

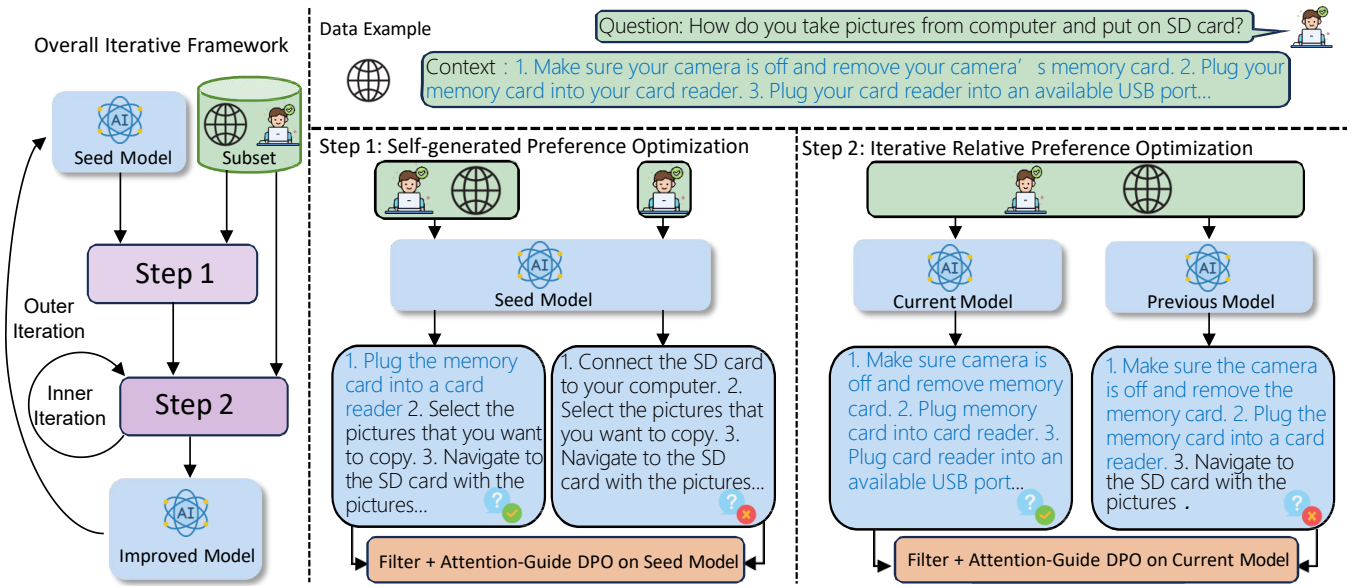

Figure 2: Overview of the REFO framework. Our framework employs a nested-loop architecture for iterative model improvement. The outer loop begins with a seed model and context-question data. Within each outer iteration, the model is first optimized using preference pairs generated from contrastive prompting (Step 1). Subsequently, it enters an inner loop (Step 2) for further refinement, where it iteratively learns from preference pairs formed by comparing responses from its current and previous generations. The final improved model is then fed back as the new seed for the next outer loop.

ness optimization and introduce an attention-guided training algorithm.

Reinforced Evolutionary Faithfulness Optimization Framework

Training Data Construction Our training process is based on the MS MARCO dataset, a pivotal large-scale benchmark for deep learning in search, Information Retrieval (IR), and Natural Language Processing (NLP) (Bajaj et al. 2018). For each training iteration, we sample a subset D_n of 1,000 instances from the MS MARCO passage-ranking corpus. Each instance provides an authentic query (question) and a document (context) to be used as input for our model.

Overview Iteration Our framework operates as a nested-loop iterative process that evolves over N main iterations. Each iteration n is composed of two sequential phases: an Initialization Optimization step, followed by up to T sub-iterations of a Relative Preference Optimization step. The resulting model from iteration n serves as the initialization for the subsequent iteration $n+1$, thereby creating a continuous, self-evolving training pipeline. Formally, we denote the seed model for iteration n as $M_{n,0}$, the model after the first phase as $M_{n,1}$, and the model after the t-th sub-iteration as $M_{n,t+1}$.

Bootstrap Preference Warm-up Our approach begins with a strong baseline model, $M_{0,0}$, which is a robust pre-trained LLM that has undergone initial fine-tuning for faithfulness. To generate training data, we create two distinct prompt variations for each context-question pair in the training subset D_n. The first is a **context-present** prompt, which includes both the context and the question, while the second is a **context-absent** prompt, containing only the question.

Using the model $M_{n,0}$, we generate a response for each prompt variant. These two responses form a preference pair, (y_w, y_l), where the response generated with the context (y_w) is designated as the winning response and the one generated without it (y_l) as the losing response. This pair is subsequently used for DPO training. This training paradigm is designed to explicitly teach the model to prioritize contextual information.

Iterative Relative Preference Optimization Traditional iterative frameworks (e.g., ReST(Gulcehre et al. 2023), ReST$^{\text{EM}}$(Singh et al. 2024), and ReST-MCTS*(Zhang et al. 2024)) involve repeated training on a fixed dataset until performance converges. However, this approach introduces significant risks of unnecessary overfitting and catastrophic model collapse(Dohmatob et al. 2025). To mitigate these issues, we propose a novel framework inspired by the principles of REBEL (Gao et al. 2024).

Our methodology shifts the focus from repeatedly optimizing for absolute rewards of "faithful" versus "unfaithful" responses. Instead, we reframe the objective to optimize the relative difference in faithfulness between distinct responses. Specifically, We propose a "relative preference" framework wherein contrastive pairs consist of two responses to the same context-present prompt: the response from the **current model**, designated as the winning response (y_w), and the response from the **previous model**, designated as the losing response (y_l). This relative preference framework forms the core of our novel REFO architecture.

Algorithm 1: REFO Framework

Initialize: Seed model $M_{0,0}$; Outer iterations N_{iter}; Inner iteration T_{step}; Thresholds $\tau_{\text{initial}}, \tau_{\text{relative}}$.
Training:
for $i = 0 \rightarrow N_{\text{iter}}$ **do**
  ▷ Step 1: Initial Data Generation and Training
  $D_{i,1} \leftarrow \text{getInitData}(M_{i,0}, \tau_{\text{initial}}[i])$
  $M_{i,1} \leftarrow \text{trainModel}(M_{i,0}, D_{i,1})$
  $S_{i,1} \leftarrow \text{evalModel}(M_{i,1})$
  $M_{\text{improved}} \leftarrow M_{i,1}$
  ▷ Step 2: Inner Loop for Relative Optimization
  for $t = 1 \rightarrow T_{\text{step}}$ **do**
    $D_{i,t+1} \leftarrow \text{getRelData}(M_{i,t}, M_{i,t-1}, \tau_{\text{relative}}[i])$
    $M_{i,t+1} \leftarrow \text{trainModel}(M_{i,t}, D_{i,t+1})$
    $S_{i,t+1} \leftarrow \text{evalModel}(M_{i,t+1})$
    if $S_{i,t+1} > S_{i,t}$ **then**
      $M_{\text{improved}} \leftarrow M_{i,t+1}$
    else
      break
    end if
  end for
  ▷ Update the model for the next outer iteration
  $M_{i+1,0} \leftarrow M_{\text{improved}}$
end for
return Final model $M_{improved}$

Faithfulness Scoring and Filtering We employ a textual faithfulness detection model (Bao et al. 2024) to quantify the contextual support for each response in a preference pair, yielding scores whose difference serves as the reward signal: reward $= \text{score}_{y_w} - \text{score}_{y_l}$. We then filter out sample pairs with reward values below a positive threshold and, inspired by curriculum learning, sort the resulting training set in descending order by reward value. Empirical results show this sorting strategy significantly enhances the smoothness and stability of the training process (Appendix A). The complete training procedure is detailed in Algorithm 1.

Attention Guided Training Algorithm

Recent studies (Fang et al. 2025; Chuang et al. 2024) suggest that a gradual loss of attention to the initial context, as response length grows, is a key cause of declining faithfulness and hallucination. In Appendix B, we provide a theoretical analysis demonstrating how the **DPO** process can implicitly enhance attention scores on relevant tokens. To further investigate this, we adopt the "Lookback Ratio" (Chuang et al. 2024). Specifically, for a transformer model with L layers and H heads, generating the t-th token y_t requires an input consisting of the context sequence $X = \{x_1, \ldots, x_N\}$ and the previously generated sequence $Y = \{y_1, \ldots, y_{t-1}\}$. The "Lookback Ratio" then measures the attention a single head (the h-th head in the l-th layer) pays to the context versus the generated tokens. It is formally defined as follows:

$$LR_t^{l,h} = \frac{A_t^{l,h}(\text{context})}{A_t^{l,h}(\text{context}) + A_t^{l,h}(\text{generation})} \quad (1)$$

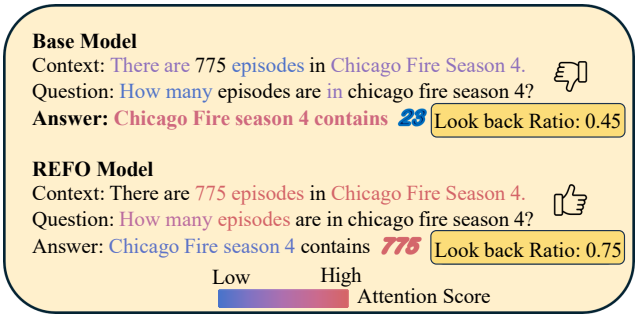

Figure 3: Token-level attention distribution of Llama-3-8B-Instruct before and after training.

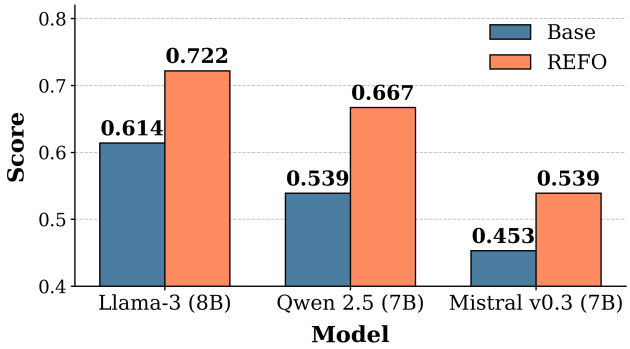

Figure 4: Average Lookback Ratio for different models on NQ-Swap

where

$$A_t^{l,h}(\text{context}) = \frac{1}{N_c} \sum_{i=1}^{N_c} a_i^{l,h} \quad (2)$$

$$A_t^{l,h}(\text{generation}) = \frac{1}{N_g} \sum_{j=N_c+1}^{N} a_j^{l,h} \quad (3)$$

We partition the model's attention distribution during token generation into two components: attention allocated to the **context** and attention allocated to the **previously generated tokens**. Our core hypothesis is that more faithful responses concentrate their attention on the context, as they leverage external knowledge. In contrast, responses that focus on preceding tokens are more likely to rely on the model's internal parametric knowledge. To validate this hypothesis, we measured the model's attention distribution on the NQ-Swap dataset before and after our proposed training. A visualization of a specific instance is presented in Figure 3. The statistical results, shown in Figure 4, demonstrate that the model allocates significantly more attention to the context post-training.

$$\mathcal{L}_{\text{DPO}}(\pi_\theta; \pi_{\text{ref}}) = - \mathbb{E}_{(\mathbf{x}, y_w, y_l)\sim\mathcal{D}} \left[\log \sigma \left(\beta \log \frac{\pi_\theta(y_w \mid \mathbf{x})}{\pi_{\text{ref}}(y_w \mid \mathbf{x})} - \beta \log \frac{\pi_\theta(y_l \mid \mathbf{x})}{\pi_{\text{ref}}(y_l \mid \mathbf{x})} \right) \right] \quad (4)$$

Based on these findings, we propose to implicitly incorporate a reward signal related to the lookback ratio during training. Inspired by research on β-DPO(Wu et al. 2024), which highlights the effectiveness of assigning distinct β values to different examples, we utilize the lookback ratio to compute a unique scaling factor for each sample.

First, we define the sentence-level Lookback Ratio (LR_s) as the average of token-level ratios across all layers (L), attention heads (H), and decoding steps (T):

$$LR_s = \frac{1}{LHT} \sum_{l=1}^{L} \sum_{h=1}^{H} \sum_{t=1}^{T} LR_t^{l,h} \quad (5)$$

We then define the scaling factor for β as the Attention Ratio. For a given sample s within a dataset of S sentences, its Attention Ratio is defined as its own LR_s normalized by the average LR_s across the entire dataset:

$$\text{Attention\_Ratio}_s = \frac{LR_s}{\frac{1}{S}\sum_{i=1}^{S} LR_s} \quad (6)$$

This factor modulates the β parameter, thereby introducing an implicit reward into the training process. This leads to our modified DPO loss, which we term Attention-DPO (Att-DPO), where k is a model-specific hyperparameter that controls the degree of scaling:

$$\mathcal{L}_{\text{Att-DPO}}(\pi_\theta; \pi_{\text{ref}}) = - \mathbb{E}_{(\mathbf{x}, y_w, y_l)\sim\mathcal{D}} \left[\log \sigma \left(\beta \left(\log \frac{\pi_\theta(y_w|\mathbf{x})}{\pi_{\text{ref}}(y_w|\mathbf{x})} - \log \frac{\pi_\theta(y_l|\mathbf{x})}{\pi_{\text{ref}}(y_l|\mathbf{x})} \right) \cdot \text{Attention\_Ratio}_{y_w}^{k} \right) \right] \quad (7)$$

We substitute the standard DPO with our proposed Att-DPO and integrate it into the REFO framework, thereby creating the Att-REFO framework.

Experiments

Baselines and Datasets

Datasets To evaluate our model, we first utilized the MemoTrap (Liu and Liu 2023) and NQ-Swap (Longpre et al. 2021) datasets to assess robustness, following the methodology of prior work. These datasets are specifically designed to evaluate the model's ability to handle adversarial conflicts within its parametric knowledge. Second, for response quality, we selected widely-used datasets including NQ-Open (Lee, Chang, and Toutanova 2019) and SQuAD (Rajpurkar et al. 2016) for contextual short-form question answering, as well as ELI5 (Fan et al. 2019) for long-form text generation. Additionally, we employed the recently-released ConFiQA (Bi et al. 2025) dataset. This dataset is specifically designed to address faithfulness and hallucination in large language models, including three subsets: QA (Question-Answering), MR (Multi-hop Reasoning), and MC (Multi-Conflicts). It evaluates faithfulness by simulating various "knowledge conflict" scenarios. We also used FollowBench (Jiang et al. 2024) to assess the model's instruction-following capabilities.

Metrics For short-form QA datasets such as NQ-Open, SQuAD, MemoTrap, and ConFiQA, we adopt a zero-shot setting that simulates a Retrieval-Augmented Generation (RAG) scenario. Performance is evaluated using the span Extraction Matching (span EM) score, where a prediction is deemed correct if a generated span precisely matches any of the reference answers. For the long-form generation dataset ELI5, we use the ROUGE-1, ROUGE-2, and ROUGE-L scores to comprehensively assess content quality. For the instruction-following task, FollowBench, we report the Consistent Satisfaction Levels (CSL) (Jiang et al. 2024), which measures the number of consecutive instruction hardness levels the model can satisfy.

Models and Baselines We subjected our proposed method to a rigorous evaluation to validate its generality and effectiveness. Our framework utilizes a diverse set of foundation models, including the LLaMA 3, Qwen 2.5, and Mistral families. Furthermore, we performed a comparative analysis against a suite of powerful baselines focused on faithfulness: the decoding-based method DECORE (Gema et al. 2024); post-training methods such as ChatQA (Liu et al. 2025), TrustAlign (Song et al. 2025), and Context-DPO (Bi et al. 2025); and the self-supervised training method SCOPE (Duong et al. 2025). Since SCOPE's approach involves selecting specific training datasets for different tasks, we used the same dataset as our own method for a fair comparison in our experiments.

Implementation Details

Following the procedure outlined in section 3.1, we sample 1,000 instances from the MS MARCO dataset for each training iteration. All models are fine-tuned on a cluster of 8 NVIDIA A6000 GPUs with BF16 precision using the REFO method. For efficient parameter tuning, we employ Low-Rank Adaptation (LoRA) with a rank (r) of 8 and an alpha (α) of 16. The optimization is performed using the AdamW optimizer (Loshchilov and Hutter 2019) (with $\beta_1 = 0.9, \beta_2 = 0.95$), combined with a cosine learning rate schedule with a peak learning rate of 2×10^{-5}. In the REFO configuration, limited by computational resources, we set the number of outer- and inner-loop iterations to $N = 5$ and $T = 2$, the initial threshold list was set to [0.2,0.3,0.4,0.4,0.4] and the relative threshold list to [0.1,0.2,0.2,0.3,0.3], respectively. For the Att-REFO configuration, we set the value of k to 3 for LLaMA-3 and Mistral, and to 7 for Qwen2.5. A consistent decoding temperature of 0.1 is maintained across all experiments.

Evaluation of Model Faithfulness

Our empirical results illustrated in Table 1 and 2 yield three key findings. **First**, our core iterative evolutionary method, embodied by REFO and its variant Att-REFO, demonstrates significant effectiveness in enhancing contextual faithfulness.

Model	Method	Robustness		Response Quality				
						Eli5		
		NQ-Swap Span EM ↑	Memo-Trap Span EM ↑	NQ-Open Span EM ↑	SQuAD Span EM ↑	R-1 F1 ↑	R-2 F1 ↑	R-L F1 ↑
Llama-3-8B	Vanilla	73.54%	73.60%	80.15%	88.20%	26.88%	9.12%	23.88%
	DECORE	80.53%	74.40%	82.03%	84.90%	28.45%	10.96%	25.65%
	ChatQA	67.70%	30.60%	76.80%	88.50%	28.61%	12.67%	26.26%
	Trust-Align	75.56%	70.95%	77.38%	50.90%	11.36%	10.55%	20.15%
	Context-DPO	82.76%	72.90%	82.86%	89.90%	28.15%	10.43%	24.98%
	SCOPE	76.72%	10.04%	80.38%	68.80%	22.05%	8.89%	14.02%
	REFO	**90.08%**	**80.19%**	**92.02%**	**93.30%**	30.97%	12.83%	27.77%
	Att-REFO	88.22%	79.28%	91.64%	92.90%	**34.03%**	**14.93%**	**30.48%**
Qwen2.5-7B	Vanilla	79.35%	54.19%	82.29%	90.30%	23.08%	6.06%	20.55%
	DECORE	81.93%	54.56%	83.76%	82.80%	24.31%	7.01%	21.46%
	Trust-Align	79.69%	53.71%	77.93%	80.30%	23.15%	6.45%	20.24%
	Context-DPO	82.13%	55.34%	83.13%	91.80%	23.68%	6.53%	21.13%
	SCOPE	79.75%	44.55%	87.98%	78.50%	**36.60%**	**17.33%**	24.76%
	REFO	84.77%	**58.68%**	86.67%	93.60%	27.18%	9.11%	24.16%
	Att-REFO	**87.61%**	57.78%	**91.94%**	**93.90%**	28.83%	10.10%	**25.29%**
Mistral-7B	Vanilla	67.76%	34.34%	79.13%	84.80%	23.20%	5.74%	20.38%
	DECORE	78.17%	30.68%	86.52%	85.30%	23.90%	6.48%	21.19%
	Context-DPO	79.62%	33.20%	80.68%	86.50%	24.97%	7.02%	22.00%
	SCOPE	49.58%	5.99%	64.71%	54.00%	26.71%	12.23%	17.42%
	REFO	87.84%	**54.75%**	**91.90%**	**95.30%**	**36.20%**	**17.95%**	**33.47%**
	Att-REFO	**88.50%**	42.73%	89.53%	93.90%	31.86%	13.67%	28.53%

Table 1: Experimental results evaluating different methods across three models on robustness and response quality tasks.

Model	Method	Faithfulness			Instruction Following
		ConFiQA-MC Span EM ↑	ConFiQA-MR Span EM ↑	ConFiQA-QA Span EM ↑	FollowBench CSL ↑
Llama-3-8B	Vanilla	33.68%	54.77%	81.93%	2.75
	DECORE	51.20%	55.92%	63.93%	2.65
	ChatQA	56.82%	57.17%	70.43%	1.09
	Trust-Align	0.15%	1.73%	50.87%	0.06
	Context-DPO	71.70%	76.67%	89.55%	2.65
	SCOPE	63.80%	65.33%	82.47%	0.16
	REFO	**74.03%**	**78.55%**	**93.40%**	**2.81**
	Att-REFO	74.13%	77.78%	92.02%	2.64
Qwen2.5-7B	Vanilla	40.18%	61.42%	82.72%	2.90
	DECORE	58.65%	62.97%	68.77%	2.73
	Trust-Align	16.67%	25.83%	69.72%	0.60
	Context-DPO	47.00%	66.80%	85.62%	2.89
	SCOPE	31.03%	33.48%	85.47%	0.42
	REFO	43.18%	65.08%	89.05%	**2.92**
	Att-REFO	**50.38%**	**69.60%**	**91.02%**	2.87
Mistral-7B	Vanilla	27.77%	41.63%	73.97%	2.32
	DECORE	41.32%	52.38%	77.32%	2.29
	Context-DPO	51.52%	60.80%	86.42%	2.52
	SCOPE	26.65%	28.03%	74.10%	0.02
	REFO	69.80%	69.45%	**91.30%**	**2.58**
	Att-REFO	**73.28%**	**75.80%**	91.12%	2.31

Table 2: Experimental results evaluating different methods across three models on faithfulness-focused and instruction following tasks.

We observe substantial improvements over base models: a 4–21% increase in robustness against parametric knowledge; a 5–13% gain in response quality; and a 9–33% enhancement on the ConFiQA dataset. **Second**, our methods consis-

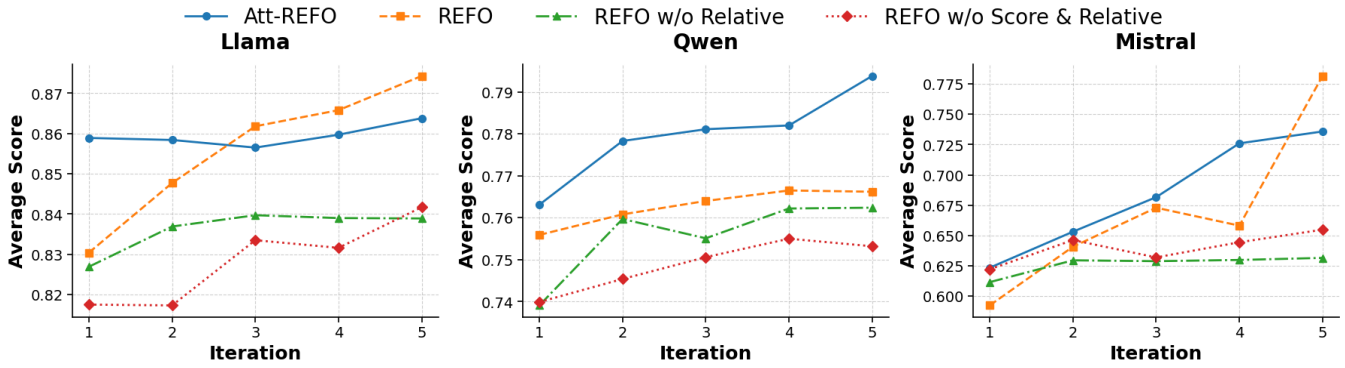

Figure 5: Ablation study of the REFO framework's components across three base models. The x-axis denotes the improved model after each training iteration, while the y-axis represents the average score on the NQ, NQ-Swap, and MemoTrap datasets.

tently achieve **state-of-the-art performance**, outperforming a range of existing approaches, including post-training methods (ChatQA, Trust-Align, Context-DPO), decoding-based strategies (DECORE), and self-supervised training (SCOPE). **Finally**, REFO-tuned models exhibit a positive externality on instruction-following. They show a modest but measurable improvement in general instruction-following, as indicated by CSL scores on FollowBench. **This suggests that our approach enhances faithfulness without sacrificing the model's ability to generalize.** We provide more detailed statistical analysis in Appendix C.

Effectiveness of Attention-Guide DPO Loss

The attention-guided DPO loss accelerates the training process while maintaining stability. As shown in Figure 5, we compare the average span-EM scores of LlaMA, Qwen, and Mistral on the NQ-Swap, NQ, and MemoTrap datasets over five iterations to demonstrate the loss's impact. Att-REFO exhibited significantly superior early-stage performance and faster convergence compared to the REFO baseline, suggesting a more effective utilization of training data. While such accelerated training might raise concerns about the risk of overfitting and a subsequent performance decline, our experiments confirmed that the model remained stable. It showed no signs of performance degradation or collapse throughout the observed iterations, thereby validating both the effectiveness of our approach and the robustness of the training process itself.

Effectiveness of Relative Preference Optimization

We investigate the impact of removing Step 2 from our iterative optimization process, thereby relying solely on Step 1 shown in Figure 5 The empirical results indicate that this variant suffers from significantly reduced optimization efficiency of Llama and mistral. Furthermore, their final performance ceiling is substantially constrained, highlighting the critical role of Step 2 in refining the optimization trajectory and achieving superior performance. Our experimental results demonstrate that our method of relative preference optimization effectively raises the performance convergence ceiling. **This key insight suggests that prior research on iterative evolution, which focused exclusively on on-policy data, overlooked the valuable learning signals embedded in the evolution process between model generations.**

Effectiveness of Score Model

In the absence of a score model, we cannot reliably determine the preference between answers in relative pairs. Therefore, in our ablation study, we simultaneously ablated both the score model and relative pairs. In the absence of a score model, this is equivalent to training on all generated data. As shown in Figure 5, for the Llama and Qwen models, incorporating the score model ensures that training is conducted on high-quality data, leading to faster and more stable improvements with each iteration. However, for the Mistral model, the performance is suboptimal. This could be because the data generated by the base Mistral model is of lower quality, and the filtering threshold we set may have been too high, resulting in an insufficient amount of training data and thus poor performance. In summary, the score model proves effective in filtering for high-quality data.

Conclusion

In this paper, we propose REFO, a self-training alignment framework designed to mitigate faithfulness hallucinations in RAG systems. By iteratively optimizing the model responses and introducing a theoretically grounded attention-guided DPO loss, we significantly accelerate training convergence without compromising stability. Crucially, our exploration of relative preference optimization underscores the value of leveraging evolutionary signals between model generations, a dimension often overlooked in prior on-policy research. Extensive experiments across various benchmarks confirm that REFO achieves state-of-the-art faithfulness and robustness while strictly preserving general instruction-following capabilities. Ultimately, this work provides a scalable paradigm for enhancing the reliability of LLMs in retrieval-augmented contexts, offering a promising direction for future research into self-evolving, trustworthy AI systems.

Acknowledgements

Sihong Xie was supported by the Department of Science and Technology of Guangdong Province (2023CX10X079), National Key R&D Program of China (Grant No.2023YFF0725001), the Guangzhou-HKUST(GZ) Joint Funding Program (Grant No.2023A03J0008), and Education Bureau Guangzhou Municipality.

References

Bajaj, P.; Campos, D.; Craswell, N.; Deng, L.; Gao, J.; Liu, X.; Majumder, R.; McNamara, A.; Mitra, B.; Nguyen, T.; Rosenberg, M.; Song, X.; Stoica, A.; Tiwary, S.; and Wang, T. 2018. MS MARCO: A Human Generated MAchine Reading COmprehension Dataset. arXiv:1611.09268.

Bao, F.; Li, M.; Luo, R.; and Mendelevitch, O. 2024. HHEM-2.1-Open.

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. 2025. Context-DPO: Aligning Language Models for Context-Faithfulness. In Che, W.; Nabende, J.; Shutova, E.; and Pilehvar, M. T., eds., *Findings of the Association for Computational Linguistics, ACL 2025, Vienna, Austria, July 27 - August 1, 2025*, 10280–10300. Association for Computational Linguistics.

Chen, H.-T.; Zhang, M.; and Choi, E. 2022. Rich Knowledge Sources Bring Complex Knowledge Conflicts: Recalibrating Models to Reflect Conflicting Evidence. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, 2292–2307.

Chuang, Y.-S.; Qiu, L.; Hsieh, C.-Y.; Krishna, R.; Kim, Y.; and Glass, J. 2024. Lookback Lens: Detecting and Mitigating Contextual Hallucinations in Large Language Models Using Only Attention Maps. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, 1419–1436.

Dohmatob, E.; Feng, Y.; Subramonian, A.; and Kempe, J. 2025. Strong Model Collapse. In *The Thirteenth International Conference on Learning Representations*.

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*.

Fan, A.; Jernite, Y.; Perez, E.; Grangier, D.; Weston, J.; and Auli, M. 2019. ELI5: Long Form Question Answering. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, 3558–3567.

Fang, Y.; Sun, T.; Shi, Y.; and Gu, X. 2025. AttentionRAG: Attention-Guided Context Pruning in Retrieval-Augmented Generation. arXiv:2503.10720.

Gao, Z.; Chang, J. D.; Zhan, W.; Oertell, O.; Swamy, G.; Brantley, K.; Joachims, T.; Bagnell, J. A.; Lee, J. D.; and Sun, W. 2024. REBEL: reinforcement learning via regressing relative rewards. In *Proceedings of the 38th International Conference on Neural Information Processing Systems*, 52354–52400.

Gema, A. P.; Jin, C.; Abdulaal, A.; Diethe, T.; Teare, P.; Alex, B.; Minervini, P.; and Saseendran, A. 2024. DeCoRe: decoding by contrasting retrieval heads to mitigate hallucinations. *arXiv preprint arXiv:2410.18860*.

Guan, X.; Zhang, L. L.; Liu, Y.; Shang, N.; Sun, Y.; Zhu, Y.; Yang, F.; and Yang, M. 2025. rStar-Math: Small LLMs Can Master Math Reasoning with Self-Evolved Deep Thinking. In *Forty-second International Conference on Machine Learning*.

Gulcehre, C.; Paine, T. L.; Srinivasan, S.; Konyushkova, K.; Weerts, L.; Sharma, A.; Siddhant, A.; Ahern, A.; Wang, M.; Gu, C.; Macherey, W.; Doucet, A.; Firat, O.; and de Freitas, N. 2023. Reinforced Self-Training (ReST) for Language Modeling. arXiv:2308.08998.

Han, R.; Zhang, Y.; Qi, P.; Xu, Y.; Wang, J.; Liu, L.; Wang, W. Y.; Min, B.; and Castelli, V. 2024. RAG-QA Arena: Evaluating Domain Robustness for Long-form Retrieval Augmented Question Answering. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, 4354–4374.

Holtzman, A.; West, P.; Shwartz, V.; Choi, Y.; and Zettlemoyer, L. 2021. Surface Form Competition: Why the Highest Probability Answer Isn't Always Right. 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*, 7038–7051. Online and Punta Cana, Dominican Republic: Association for Computational Linguistics.

Hu, E. J.; Shen, Y.; Wallis, P.; Allen-Zhu, Z.; Li, Y.; Wang, S.; Wang, L.; Chen, W.; et al. 2022. Lora: Low-rank adaptation of large language models. *ICLR*, 1(2): 3.

Jiang, Y.; Wang, Y.; Zeng, X.; Zhong, W.; Li, L.; Mi, F.; Shang, L.; Jiang, X.; Liu, Q.; and Wang, W. 2024. FollowBench: A Multi-level Fine-grained Constraints Following Benchmark for Large Language Models. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, 4667–4688.

Kirkpatrick, J.; Pascanu, R.; Rabinowitz, N.; Veness, J.; Desjardins, G.; Rusu, A. A.; Milan, K.; Quan, J.; Ramalho, T.; Grabska-Barwinska, A.; et al. 2017. Overcoming catastrophic forgetting in neural networks. *Proceedings of the national academy of sciences*, 114(13): 3521–3526.

Lee, K.; Chang, M.-W.; and Toutanova, K. 2019. Latent Retrieval for Weakly Supervised Open Domain Question Answering. In Korhonen, A.; Traum, D.; and Màrquez, L., eds., *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, 6086–6096. Florence, Italy: Association for Computational Linguistics.

Lewis, P.; Perez, E.; Piktus, A.; Petroni, F.; Karpukhin, V.; Goyal, N.; Küttler, H.; Lewis, M.; Yih, W.-t.; Rocktäschel, T.; et al. 2020. Retrieval-augmented generation for knowledge-intensive nlp tasks. *Advances in Neural Information Processing Systems*, 33: 9459–9474.

Li, K.; Zhang, T.; Li, Y.; Luo, H.; Moustafa, A. M. S. S.; Wu, X.; Glass, J. R.; and Meng, H. M. 2025. Generate, Discriminate, Evolve: Enhancing Context Faithfulness via

Fine-Grained Sentence-Level Self-Evolution. In Che, W.; Nabende, J.; Shutova, E.; and Pilehvar, M. T., eds., *Findings of the Association for Computational Linguistics: ACL 2025*, 17091–17105. Vienna, Austria: Association for Computational Linguistics. ISBN 979-8-89176-256-5.

Liu, A.; and Liu, J. 2023. The MemoTrap Dataset. https://github.com/liujch1998/memo-trap.

Liu, Z.; Ping, W.; Roy, R.; Xu, P.; Lee, C.; Shoeybi, M.; and Catanzaro, B. 2025. ChatQA: surpassing GPT-4 on conversational QA and RAG. In *Proceedings of the 38th International Conference on Neural Information Processing Systems*, NIPS '24. Red Hook, NY, USA: Curran Associates Inc. ISBN 9798331314385.

Longpre, S.; Perisetla, K.; Chen, A.; Ramesh, N.; DuBois, C.; and Singh, S. 2021. Entity-Based Knowledge Conflicts in Question Answering. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*.

Loshchilov, I.; and Hutter, F. 2019. Decoupled Weight Decay Regularization. In *International Conference on Learning Representations*.

Mei, L.; Yao, J.; Ge, Y.; Wang, Y.; Bi, B.; Cai, Y.; Liu, J.; Li, M.; Li, Z.-Z.; Zhang, D.; Zhou, C.; Mao, J.; Xia, T.; Guo, J.; and Liu, S. 2025. A Survey of Context Engineering for Large Language Models. arXiv:2507.13334.

Niu, C.; Wu, Y.; Zhu, J.; Xu, S.; Shum, K.; Zhong, R.; Song, J.; and Zhang, T. 2024. RAGTruth: A Hallucination Corpus for Developing Trustworthy Retrieval-Augmented Language Models. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, 10862–10878.

Pajo, P. 2025. Context Engineering: Enhancing Large Language Model Performance Through Comprehensive Contextual Management.

Pang, R. Y.; Yuan, W.; He, H.; Cho, K.; Sukhbaatar, S.; and Weston, J. 2024. Iterative reasoning preference optimization. *Advances in Neural Information Processing Systems*, 37: 116617–116637.

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. *Advances in Neural Information Processing Systems*, 36: 53728–53741.

Rajpurkar, P.; Zhang, J.; Lopyrev, K.; and Liang, P. 2016. SQuAD: 100,000+ Questions for Machine Comprehension of Text. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, 2383–2392.

Scudder, H. 1965. Probability of error of some adaptive pattern-recognition machines. *IEEE Transactions on Information Theory*, 11(3): 363–371.

Shi, W.; Han, X.; Lewis, M.; Tsvetkov, Y.; Zettlemoyer, L.; and Yih, W.-t. 2024. Trusting your evidence: Hallucinate less with context-aware decoding. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 2: Short Papers)*, 783–791.

Singh, A.; Co-Reyes, J. D.; Agarwal, R.; Anand, A.; Patil, P.; Garcia, X.; Liu, P. J.; Harrison, J.; Lee, J.; Xu, K.; Parisi, A. T.; Kumar, A.; Alemi, A. A.; Rizkowsky, A.; Nova, A.; Adlam, B.; Bohnet, B.; Elsayed, G. F.; Sedghi, H.; Mordatch, I.; Simpson, I.; Gur, I.; Snoek, J.; Pennington, J.; Hron, J.; Kenealy, K.; Swersky, K.; Mahajan, K.; Culp, L. A.; Xiao, L.; Bileschi, M.; Constant, N.; Novak, R.; Liu, R.; Warkentin, T.; Bansal, Y.; Dyer, E.; Neyshabur, B.; Sohl-Dickstein, J.; and Fiedel, N. 2024. Beyond Human Data: Scaling Self-Training for Problem-Solving with Language Models. *Transactions on Machine Learning Research*. Expert Certification.

Song, J.; Wang, X.; Zhu, J.; Wu, Y.; Cheng, X.; Zhong, R.; and Niu, C. 2024. RAG-HAT: A Hallucination-Aware Tuning Pipeline for LLM in Retrieval-Augmented Generation. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing: Industry Track*, 1548–1558.

Song, M.; Sim, S. H.; Bhardwaj, R.; Chieu, H. L.; Majumder, N.; and Poria, S. 2025. Measuring and Enhancing Trustworthiness of LLMs in RAG through Grounded Attributions and Learning to Refuse. In *The Thirteenth International Conference on Learning Representations*.

Wang, P.; Li, L.; Chen, L.; Song, F.; Lin, B.; Cao, Y.; Liu, T.; and Sui, Z. 2023. Making Large Language Models Better Reasoners with Alignment. arXiv:2309.02144.

Wu, J.; Xie, Y.; Yang, Z.; Wu, J.; Gao, J.; Ding, B.; Wang, X.; and He, X. 2024. β-DPO: direct preference optimization with dynamic β. In *Proceedings of the 38th International Conference on Neural Information Processing Systems*, 129944–129966.

Yuan, W.; Pang, R. Y.; Cho, K.; Li, X.; Sukhbaatar, S.; Xu, J.; and Weston, J. 2024. Self-rewarding language models. In *Proceedings of the 41st International Conference on Machine Learning*, 57905–57923.

Yuan, Z.; Yuan, H.; Li, C.; Dong, G.; Lu, K.; Tan, C.; Zhou, C.; and Zhou, J. 2023. Scaling Relationship on Learning Mathematical Reasoning with Large Language Models. arXiv:2308.01825.

Zelikman, E.; Wu, Y.; Mu, J.; and Goodman, N. 2022. Star: Bootstrapping reasoning with reasoning. *Advances in Neural Information Processing Systems*, 35: 15476–15488.

Zhang, D.; Zhoubian, S.; Hu, Z.; Yue, Y.; Dong, Y.; and Tang, J. 2024. Rest-mcts*: Llm self-training via process reward guided tree search. *Advances in Neural Information Processing Systems*, 37: 64735–64772.

Zhang, Z.; Dai, Q.; Bo, X.; Ma, C.; Li, R.; Chen, X.; Zhu, J.; Dong, Z.; and Wen, J.-R. 2025. A Survey on the Memory Mechanism of Large Language Model based Agents. *ACM Trans. Inf. Syst.* Just Accepted.

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