Generate, Discriminate, Evolve: Enhancing Context Faithfulness via Fine-Grained Sentence-Level Self-Evolution

Kun Li[♡*], **Tianhua Zhang**[♡*], **Yunxiang Li**[♡], **Hongyin Luo**[♢], **Abdalla Moustafa**[♡], **Xixin Wu**[♡], **James Glass**[♢], **Helen Meng**[♡]
[♡]The Chinese University of Hong Kong, Hong Kong SAR, China
[♢]Massachusetts Institute of Technology, Cambridge MA, USA
kunli@se.cuhk.edu.hk, thzhang@link.cuhk.edu.hk

Abstract

Improving context faithfulness in large language models is essential for developing trustworthy retrieval augmented generation systems and mitigating hallucinations, especially in long-form question answering (LFQA) tasks or scenarios involving knowledge conflicts. Existing methods either intervene LLMs only at inference without addressing their inherent limitations or overlook the potential for self-improvement. In this paper, we introduce GenDiE (**Gen**erate, **Di**scriminate, **E**volve), a novel self-evolving framework that enhances context faithfulness through fine-grained sentence-level optimization. GenDiE combines both generative and discriminative training, equipping LLMs with self-generation and self-scoring capabilities to facilitate iterative self-evolution. This supports both data construction for model alignment and score-guided search during inference. Furthermore, by treating each sentence in a response as an independent optimization unit, GenDiE effectively addresses the limitations of previous approaches that optimize at the holistic answer level, which may miss unfaithful details. Experiments on ASQA (in-domain LFQA) and ConFiQA (out-of-domain counterfactual QA) datasets demonstrate that GenDiE surpasses various baselines in both faithfulness and correctness, and exhibits robust performance for domain adaptation[1].

1 Introduction

Large language models have achieved remarkable success across various natural language processing tasks (OpenAI, 2024; Anthropic, 2025; DeepSeek-AI et al., 2025). Despite their impressive performance, LLMs are prone to *hallucinations*—generating plausible yet nonfactual information (Zhang et al., 2023; Li et al., 2024). This limitation poses critical risks in domains where accuracy and reliability are paramount. Retrieval-Augmented Generation (RAG) has emerged as a promising framework to mitigate hallucinations and enhance context-faithfulness (Nguyen et al., 2024) by grounding LLM outputs in provided knowledge (Gao et al., 2024; Fan et al., 2024; Luo et al., 2023). Despite its advantages, *knowledge conflicts* (Xie et al., 2024; Xu et al., 2024) between LLMs' parametric memory and external context can undermine its effectiveness. LLMs may overly rely on their internal priors while disregard provided contexts, failing to meet user requirements or incorporate the latest updates (Jin et al., 2024; Bi et al., 2024). Additionally, when tasked with *long-form* generation, such as long-form question answering (LFQA) (Stelmakh et al., 2022; Fan et al., 2019) that aims to provide in-depth and paragraph-length responses, maintaining context-faithfulness throughout the text is still challenging (Stolfo, 2024). Consequently, it is crucial to develop robust mechanism to alleviate faithfulness hallucination and ensure trustworthiness.

Many recent efforts enhance context faithfulness of LLMs through inference-time interventions, such as improving prompting strategies (Zhou et al., 2023) and context-aware decoding (Shi et al., 2024) to increase the output probability on contextual information. While effective, these method do not fundamentally address the models' inherent limitations (Bi et al., 2024). Training-based approaches, including adaptive retrieval and self-critiquing generations with reflection tokens (Asai et al., 2024) and aligning LLMs via Direct Preference Optimization (DPO) towards faithful responses (Bi et al., 2024), improve faithfulness by updating model parameters. However, these approaches typically rely on one-round optimization and do not fully explore the potential of continuous refinement. Furthermore, existing methods often train models in complete answer level holistically, which can overlook

[*] Equal contribution.
[1] The source code is available here.

Findings of the Association for Computational Linguistics: ACL 2025, pages 17091–17105

unfaithful details, particularly in the cases of long-form generation (Lai et al., 2024).

To address these challenges, we introduce a novel self-evolving framework GenDiE (**Gen**erate, **Di**scriminate, **E**volve) for enhancing LLMs' context faithfulness through fine-grained, sentence-level optimization. Our framework addresses the limitations of conventional answer-level training paradigms by operating at the sentence level, treating each constituent sentence of a response as an independent optimization unit. This paradigm enables fine-grained control over the quality of a complete response. Central to our approach is a unified training strategy that integrates sentence-level generation and evaluation capabilities. Specifically, 1) during training, the model learns to produce context-grounded sentences while developing discriminative capabilities to distinguish between faithful and unfaithful sentences. Between every two iterations, GenDiE generates candidate sentences through tree-structured sampling, self-scores their faithfulness, and constructs contrastive sentence pairs serving as training data in the next iteration. This design enables continuous model improvement through successive self-evolve cycles. 2) For inference, we design hierarchical decoding that first generates candidate sentences through standard methods, then selects optimal outputs using the model's learned scoring capacity to fully utilize both generative and discriminative capabilities that conventional single-stage decoding neglects. Overall, the sentence-level paradigm enables both fine-grained supervision for training and a comprehensive search for inference through optimization in sentence space.

We evaluate GenDiE on two benchmarks for context-faithful generation and the results demonstrate its effectiveness. GenDiE surpasses various baselines in two dimensions, faithfulness and correctness, and exhibits robust performance even in out-of-domain settings. Remarkably, our approach enables the model to consistently improve with each successive training iteration. We further conduct comprehensive experiments to verify the effectiveness of the self-scoring function, as well as the superiority of sentence-level optimization.

In summary, we make the following contributions:

- We propose GenDiE, a novel self-evolving framework that addresses the critical challenge of maintaining faithfulness in LLM responses through iterative self-improvement.
- GenDiE operates at a fine-grained sentence level, offering more precise control over faithfulness compared to previous methods that typically operate on entire response sequences.
- GenDiE integrates both generative and discriminative capabilities through multi-task training, enabling models to not only generate faithful responses but also effectively discriminate between faithful and unfaithful content.

2 Methodology

Task Formulation We focus on long-form question answering (LFQA) task which requires models to generate long and detailed answers by leveraging the evidence documents provided in the input (Xu et al., 2023). Our goal is to train the model M to generate faithful long-form answer $A = \{a_1, a_2, ...a_{|A|}\}$, where a_i denotes the i-th sentence, in response to a given input question q and its corresponding evidence passages $P = \{p_j\}_{j=1}^{k}$. For training, we assume the availability of answer labels a^* from a seed (in-domain) dataset to construct the initial training data. To comprehensively assess context faithfulness, we also employ a counterfactual dataset for out-of-domain evaluation.

Overview We present GenDiE, a *self-evolving* framework (§2.1.2) that enhances LLMs context-faithfulness in fine-grained sentence-level optimization (§2.1.1), integrating both generation and discrimination capabilities. This enables LLMs to distinguish between faithful and unfaithful responses, facilitating self-scoring for both training data construction (§2.2.2) and score-guided search during inference (§2.3). Our approach employs iterative self-training, where the models can self-generate, self-score, and therefore self-improve through multiple training iterations.

2.1 Training

2.1.1 Sentence-Level Optimization

The majority of existing approaches train models at answer level, treating an entire answer as the training target. We believe this paradigm provides limited supervision signal for learning faithful generation, especially in LFQA task where the sentences in a lengthy answer often exhibit varying levels of faithfulness. By treating the answer as a

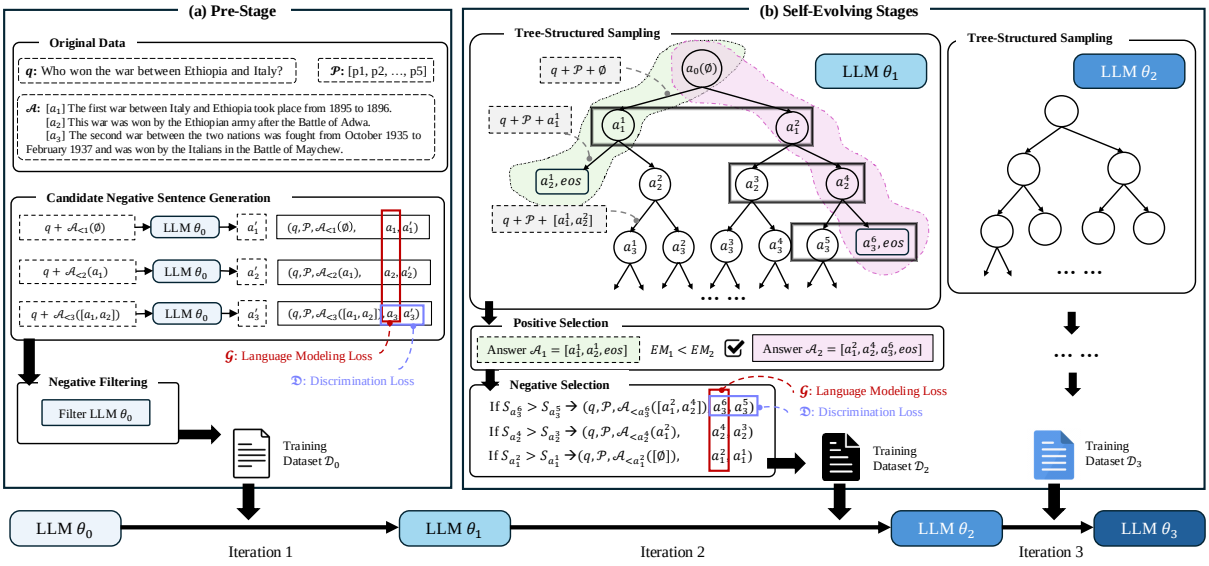

Figure 1: An overview of GenDiE: (a) Pre-stage (§2.2.1) uses gold answer sentences from a seed dataset as target faithful instances, while filtered self-generated sentences—produced without access to supporting passages—serve as negative samples. (b) Self-evolving stages (§2.1.2) leverage models from previous iteration for both self-generation and self-scoring, constructing training datasets via tree-structured sampling. Throughout all stages of the self-evolving framework, both language modeling loss (optimizing towards the target instances a) and discrimination loss (assigning higher faithfulness scores to a over a') are incorporated (§2.1).

monolithic unit, these approaches fail to capture the nuanced differences in faithfulness across individual sentences. To train a model with finer-grained supervision, we propose a sentence-level optimization. We split a target answer A into a set of sentences $\{a_i\}_{i=1}^{|A|}$, and each sentence is used as a separate training instance[2].

In addition to introducing finer-grained supervision, the sentence-level optimization can also enable sentence-level search during inference (§2.3).

2.1.2 Self-Evolving

Iterative Training for Self-evolving We propose training the model iteratively to enable self-evolving. At t-iteration, the answer sentences used for training are generated and evaluated (in terms of their faithfulness) by the model θ_{t-1} obtained at $t-1$-iteration. In this way, the model gets self-evolved with progressively improved training data.

Training Objective To achieve the goal of self-evolving within our single-model framework, the model should grasp the abilities of both generation and scoring in faithfulness. With this motivation, we train the model with multi-task training paradigm. Furthermore, inspired by Odds Ratio Preference Optimization (Hong et al., 2024), an algorithm that optimizes language generation and preference alignment simultaneously, we train our model by maximizing

$$\mathcal{L} = \sum_{a \in A} \Bigg(\underbrace{\log \mathcal{P}_\theta(a|q,P,A_{\prec a})}_{\text{Language Modeling Objective}} + \underbrace{\lambda \log \sigma \left(\log \frac{\mathcal{P}_\theta(a|q,P,A_{\prec a})(1-\mathcal{P}_\theta(a'|q,P,A_{\prec a}))}{\mathcal{P}_\theta(a'|q,P,A_{\prec a})(1-\mathcal{P}_\theta(a|q,P,A_{\prec a}))} \right)}_{\text{Discrimination Objective}} \Bigg); \tag{1}$$

where

$$\mathcal{P}_\theta(a|q,P,A_{\prec a}) = \prod_{t=1}^{|a|} \mathcal{P}_\theta(a[t]|q,P,A_{\prec a},a[1:t-1])^{1/|a|}. \tag{2}$$

a and a' in Eq.1 denote the target and negative sentence respectively, and they share a common prefix $A_{\prec a}$. The prefix $A_{\prec a}$ includes all the sentences in the answer A preceding a (a'). $a[j]$ in Eq.2 denotes the j-th token of a. In Eq.1, the first term in the summation is the vanilla language modeling objective, aimed to maximize the probability of generating a by the model. Following Zhang et al. (2024a), we define the faithfulness score of sentence a to passages P as $S_a = \log \mathcal{P}_\theta(a|q,P,A_{\prec a})$. Therefore, the second term in the summation in Eq.1, the discrimination objective, optimizes the model to assign a higher faithfulness score to a over a'.

Note that the above training objective is applied

[2]Answers are split based on sentence-ending punctuation marks, including periods, exclamation and question marks.

throughout the whole training process, but the training sets $\{(q, A_{\prec a}, a, a^{'})\}$, which are constructed using the latest trained model, vary across different training iterations. §2.2 will elaborate the process of data construction.

2.2 Iterative Data Construction

To enable sentence-level optimization with Eq.1, the training instances must include 1) a contrastive sentence pair $(a, a^{'})$ that share a common prefix $A_{\prec a}$; and 2) the relative relation in the faithfulness degree of the pair, i.e., $S_a > S_{a'}$. Based on this, we design following data construction methods for pre-stage (first iteration) and self-evolving stages (remaining iterations), respectively.

2.2.1 Pre-Stage

Initially, the pre-trained model θ_0 lacks sufficient self-scoring capability to directly differentiate answer candidates in varying faithfulness among self-generations. To address this, we construct the pre-stage training dataset $\mathcal{D}_0$, in which sentences from ground-truth answers a^* serve as target instances, and model-generated responses, produced without access to evidence passages (i.e., conditioned only on the question and answer prefix), serve as negative instances as shown in Fig.1(a). Specifically, for each target sentence a in a ground-truth answer, we obtain corresponding negative sentence $a^{'}$ as $a^{'} \sim \mathcal{P}_{\theta_0}(*|q, A_{\prec a})$. Although $a^{'}$ are likely to lack sufficient faithfulness due to the absence of supporting passages, it remains possible for θ_0 to answer the question correctly given its extensive pre-training on massive corpora. Consequently, we implement a heuristic negative sample filtering process for quality control with model θ_0 (see details App. A), trying to ensure that the faithfulness score of the positive instance exceeds that of its negative counterpart, i.e., $S_a > S_{a'}$.

2.2.2 Self-evolving Stages

Tree-structured Sampling For self-evolving stages, we no longer use the ground-truth answers on the training set. Instead, given the question and passages on the training set, we sample and score answer sentence pairs with the latest model– the answer sentences used for t-iteration training are generated by the model θ_{t-1} obtained at $t-1$-iteration. Furthermore, to efficiently obtain diverse and high-quality sentence pairs, we devise a tree-structured sampling method, the operation process of which can be illustrated with an n-ary tree as shown in Fig.1(b).

The root node of the tree represents an empty string. Each non-root node indicates a sentence a, which is sampled based on its all prefix sentences along the path from the root to its parent node. Therefore, each path from the root to a leaf node constitutes a complete answer and the node in i-th layer a_i is thus the $i-$th sentence in the answer. Formally, $a_i \sim \mathcal{P}_\theta(*|q, P, a_1, ..., a_{i-1})$. We sample n distinct sentences for each generation, finally leading to an n-ary tree (using $n = 2$ as an example in Fig.1(b)). During the expansion of the tree, a path will be terminated once $[eos]$ is output. The tree expands layer by layer until it reaches the specified size or all paths are terminated.

Sentence Pairs Selection We then construct contrastive sentence pairs from the generated tree. Specifically, we evaluate all the terminated paths on the tree, and select the one that achieves the highest accuracy, measured by Exact Match (EM), against the gold answer. Each node a along the path, together with its sibling node a', will compose a contrastive sentence pair (a, a'), if $S_a > S_{a'}$. Note that the sentences of sibling nodes must have a common prefix.

2.3 Hierarchical Inference

When making inference with the trained model, one can use some standard decoding methods like greedy search. However, this will make the model's scoring ability idle. To make full use the ability of sentence-level scoring, we propose hierarchical inference for answer generation. This method utilizes the model's scoring ability to carry out sentence-level search during generation.

The hierarchical inference is a two-level inference method. The inner one is token-level inference, used to determine the tokens within a sentence. It is similar to the standard decoding methods which decode text token by token, so conventional methods like beam search or top-k/p sampling can be applied to the token-level inference.

More importantly, we also want to optimize answers in sentence level during inference. We view a complete sentence as a step and generate multiple candidate sentence iteratively. Based on their faithfulness scores, we then utilize beam search algorithm to determine the sentence at each step. In details, a fixed number N of beams are maintained throughout the generation process. When generating the i-th sentence, for each beam, based on q, P and $A_{<i}$, M (beam width) candidate sentences are

sampled through token-level inference. Among the NM continuations, the top N beams with the highest length-normalized faithfulness scores, defined as $\frac{\sum_{j=1}^{i} S_{a_j}}{i}$, are selected for the next step of generation. Again, each of these selected beams are expanded by sampling M next sentences. The process will stop once all the selected beams end with $[eos]$ or the maximum search depth is exceeded. Finally, the beam with the highest length-normalized faithfulness score will be returned as the answer. An illustrative example is provided in §5.4.

3 Experiments

3.1 Datasets & Metrics

We conduct experiments on two benchmarks: 1) **ASQA** (Stelmakh et al., 2022), a long-form factoid question answering dataset derived from AmbigQA (Min et al., 2020) , which includes crowd-sourced, paragraph-long answers to ambiguous questions; and 2) **ConFiQA** (Bi et al., 2024), a dataset designed for retrieval-augmented generation (RAG) scenarios, incorporating knowledge conflicts through counterfactual passages to assess context-faithfulness. For all reported results, we use the same retrieval results with GTR (Ni et al., 2021) as dense retriever for ASQA dataset following Gao et al. (2023); Aly et al. (2024), and the counterfactual contexts provided by Bi et al. (2024) for ConFiQA. We train our model exclusively on ASQA (in-domain) and evaluate it on ConFiQA (out-of-domain) to assess its generalization and adaptation ability. See data statistics in App.B.1.

We follow previous literature (Aly et al., 2024; Gao et al., 2023) to measure correctness (via EM Recall and hit rate), and faithfulness (via an NLI-trained T5-11B model (Honovich et al., 2022) and AlignScore (Zha et al., 2023)). In the case of ConFiQA, the ground-truth answer is a single entity name, along with its aliases. Consequently, EM Recall is equivalent to the hit rate.

3.2 Implementation Details

Model Training: λ in Eq.1 is set to 0.5. We train the model for 3 iterations, and each iteration takes one epoch. `Llama-3.1-8b` (Grattafiori et al., 2024) is used as the base model. Refer to App.B.2 for more training details.

Data construction: During tree-structured sampling at self-evolving stages, for each continuation, $n = 3$ sentences are sampled with random sampling (p=0.9, temperature=1).

Hierarchical Decoding: For sentence-level inference, both N and M are 3. We thus use beam search with beam width as 3 for token-level inference and produce 3 sentences at each step. The maximum number of steps for sentence-level inference is set to 6.

3.3 Baselines

To provide a comprehensive evaluation, we compare GenDiE with two types of baselines.

Training-free Approaches 1) **In-context Prompting** generates answers with two demonstration examples by `gpt-4o` (OpenAI, 2024) and `Llama-3.1-8b-Instruct` (Grattafiori et al., 2024). The complete prompt is shown in App.D. 2) **CAD** (Shi et al., 2024) uses contrastive decoding to amplify the difference in output probabilities with and without context, reinforcing the model's attention to input context during inference. The backbone model is `Llama-3.1-8b` with standard fine-tuning described in baseline-4. 3) **Extractive Sentence Selection** uses embedding models `stella_en_1.5B_v5`[3] and `instructor-large`[4] to select top-K relevant passage sentences according to the question as final answer, where K depends on GenDiE for fair comparison.

Training-based Approaches 4) **Standard SFT** directly fine-tune LLMs to replicate ground-truth answers. 5) **GenDiE$_{\text{answer-level}}$** is a variant of our method that using the same training paradigm but optimize in *answer* level instead of *sentence* level. In this setting, complete answers instead of separate sentences, are sampled and scored by the model to construct contrastive answer pairs. 6) **GenDiE$_{\text{gold-answer}}$** always uses sentences from ground-truth answers as target instead of self-generations with highest self-scored values, which is the setting used in the first iteration of GenDiE.

4 Main Results

GenDiE with greedy search outperforms most of training-free and training-based baselines in various metrics. Furthermore, even with the same checkpoint, our approach earns significant boost by using hierarchical inference over greedy search or vanilla beam search, demonstrating the benefit of sentence-level search during inference.

[3] https://huggingface.co/NovaSearch/stella_en_1.5B_v5

[4] https://huggingface.co/hkunlp/instructor-large

		ASQA				ConFiQA		
		Faithfulness		Correctness		Faithfulness		Correctness
Type	Method (Decoding)	AlignScore	T5NLI	EM Rec.	Hit	AlignScore	T5NLI	Hit
Training-Free	In-context Promping: `gpt-4o`	75.11	71.52	**47.21**	18.57	44.56	34.11	35.96
	In-context Promping: `llama3.1-8b`	76.90	70.39	40.30	14.35	50.67	30.78	55.14
	CAD: `llama3.1-8b` Standard SFT, greedy	72.07	66.18	42.72	17.61	77.46	52.79	49.38
	Extractive Sentence Selection: `stella-1.5b`	–	–	41.00	17.19	–	–	42.31
	Extractive Sentence Selection: `instructor-large`	–	–	36.58	14.45	–	–	39.17
Training-based	Standard SFT: `llama3.1-8b`, greedy	65.27	57.54	41.93	16.35	44.94	40.74	34.58
	Standard SFT: `llama3.1-8b`, beam3	73.88	66.47	44.64	19.51	66.03	62.90	59.68
	$\text{GenDiE}_{\text{answer-level}}$, greedy	64.48	56.51	43.13	19.20	69.81	66.17	77.91
	$\text{GenDiE}_{\text{gold-answer}}$, greedy	71.66	64.76	42.88	17.62	64.72	59.89	65.27
Ours	Greedy search	73.87	69.16	43.61	18.57	72.32	<u>70.17</u>	73.72
	Vanilla beam search (beam3)	<u>82.30</u>	<u>79.42</u>	43.71	<u>18.88</u>	<u>78.54</u>	69.33	<u>80.95</u>
	Hierarchical inference (beam3-beam3)	**84.90**	**82.03**	<u>45.75</u>	**21.52**	**80.73**	**80.69**	**84.63**

Table 1: Performance results of different methods on ASQA (in-domain) and ConFiQA (out-of-domain) benchmarks. **Bold** and <u>underline</u> numbers denote the best and second-best performance. Note that `Extractive Sentence Selection` directly uses input passage sentences as answers, making its faithfulness inherently 100%. Consequently, we denote its faithfulness as "–" to indicate that faithfulness evaluation is not applicable when compared to other generative approaches. See additional experimental results in App. C.

For training-free methods, prompting with `gpt-4o` performs well on ASQA dataset, achieving the best EM Recall, mainly due to its extensive world knowledge obtained during pretraining. However, `gpt-4o` faces challenges in knowledge conflict scenarios on ConFiQA dataset, as it heavily relies on its own knowledge rather than the provided passages. This observation aligns with the findings reported by Bi et al. (2024). A similar pattern is observed in prompting with `Llama`, implying that simply with prompting-based method, the models often disregard those external knowledge that conflicts with their parametric knowledge. Extractive Sentence Selection selects the most relevant sentences as answers (and thereby enjoys high faithfulness), but the concatenation of separated sentences, rather than free-form answer generation, often results in low correctness and readability due to its inflexibility.

Focusing on training-based methods, with the same training data, $\text{GenDiE}_{\text{gold-answer}}$ outperform Standard SFT, especially in terms of faithfulness, demonstrating the efficacy of our multi-task training objective. This suggests that the additional training with discrimination objective can also contribute to faithful generation. Furthermore, GenDiE with greedy search is superior to $\text{GenDiE}_{\text{gold-answer}}$ across most metrics, which can be attributed to the iterative update of training data (more discussion in §5.1). However, $\text{GenDiE}_{\text{answer-level}}$, which also underwent data update, shows lower faithfulness than Standard SFT on ASQA. This discrepancy between above two comparisons highlights the necessity of sentence-level operation for the data update, as evaluating faithfulness at sentence level is more feasible than at answer level (§5.3).

5 Analysis

In this section, we conduct ablation studies on the key components of our approach to evaluate their impact. Unless otherwise specified, vanilla greedy search is employed for decoding in all experiments within this section. Details of some variants in experiments are in §3.3.

5.1 Effectiveness of Self-Evolving

Our approach trains the model with continuously updated data. To investigate the effect of this paradigm, we take a closer look into the performance variation of GenDiE and $\text{GenDiE}_{\text{gold-answer}}$ across different iterations.

As shown in Fig.2, with the training progressing, for all metrics, GenDiE shows continuous improvement on both benchmarks and outperforms $\text{GenDiE}_{\text{gold-answer}}$ consistently across iterations. Notably, the gap in faithfulness (measured by both AlignScore and T5NLI) between two methods becomes wider from the second to the third iteration. This comparative result underscores the advantage of self-evolving through iterative data update. On ConFiQA, $\text{GenDiE}_{\text{gold-answer}}$ even experiences declines across various metrics but GenDiE still gets improved, implying that self-evolving could help alleviate negative effects of out-of-domain settings.

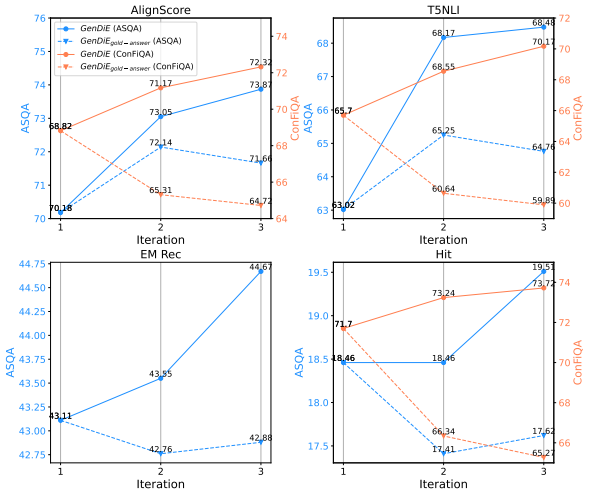

Figure 2: Performance comparisons between GenDiE and $\text{GenDiE}_{\text{gold-answer}}$ across iterations. Dual y-axes is used, with the left y-axis representing the performances on ASQA and the right one representing the performances on ConFiQA.

Dataset	Checkpoint	AlignScore	T5NLI	EM Rec.	Hit
ASQA	$\text{GenDiE}_{\text{iter1}}$	70.18	63.02	43.11	18.46
	$\text{GenDiE}_{\text{iter2}}$	73.05	68.17	43.55	18.46
	$\text{GenDiE}_{\text{T5}}$	73.80	70.31	42.30	17.09
ConFiQA	$\text{GenDiE}_{\text{iter1}}$	68.82	65.70	-	71.70
	$\text{GenDiE}_{\text{iter2}}$	71.17	68.55	-	73.24
	$\text{GenDiE}_{\text{T5}}$	71.73	69.16	-	72.69

Table 2: Performance comparison between $\text{GenDiE}_{\text{iter2}}$ and $\text{GenDiE}_{\text{T5}}$, both of which are trained from $\text{GenDiE}_{\text{iter1}}$.

5.2 Effectiveness of Self-Scoring

For data construction in self-evolving stage (§2.2.2), GenDiE relies the built-in scoring component to assess the faithfulness degrees of sentences, which is necessary for enabling self-evolving. To study the effectiveness of the self-scoring component, we replace it with the NLI-trained T5-11B model for evaluating sentence faithfulness during data construction. The NLI-trained T5 is also used as the evaluation tool to measure faithfulness in previous experiments. With these self-generated but T5-scored data, we train $\text{GenDiE}_{\text{T5}}$ from $\text{GenDiE}_{\text{iter1}}$ checkpoint for one iteration, employing the same training objective (Eq.1). We then compare it with $\text{GenDiE}_{\text{iter2}}$. Note that two methods score and select contrastive pairs from the same collection of sentence pairs sampled from $\text{GenDiE}_{\text{iter1}}$. We assess their performances after just one iteration of training, in order to exclude the effect brought by the different self-generated training data at following iterations.

$\text{GenDiE}_{\text{iter2}}$ exhibits slightly lower T5NLI scores than $\text{GenDiE}_{\text{T5}}$. This is expected since $\text{GenDiE}_{\text{T5}}$ is trained with the direct supervision of T5NLI scores, while GenDiE relies on the trained self-scoring component. Nevertheless, $\text{GenDiE}_{\text{iter2}}$ still shows a comparable level of faithfulness on both benchmarks, highlighting the reliable role of the self-scoring component in assessing the faithfulness of sentences when constructing sentence pairs. Also note that $\text{GenDiE}_{\text{T5}}$ incurs a degradation in correctness, compared with $\text{GenDiE}_{\text{iter1}}$.

5.3 Effectiveness of Sentence-Level Optimization

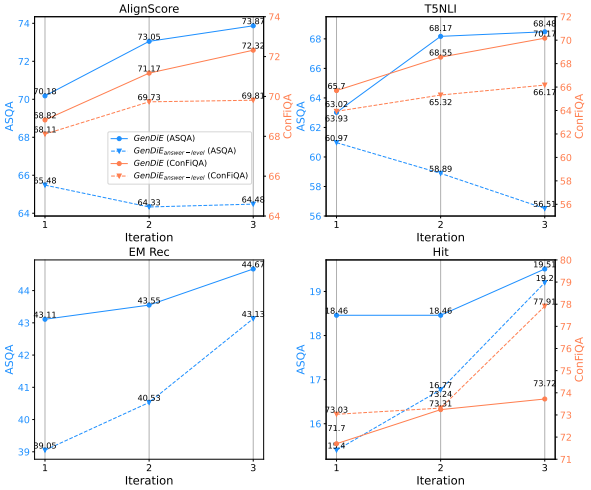

Figure 3: The comparisons between GenDiE and $\text{GenDiE}_{\text{answer-level}}$ across iterations. Dual y-axes is used, with the left y-axis representing the performances on ASQA and the right one representing the performances on ConFiQA.

The experiment in §4 demonstrates the significant improvement achieved by hierarchical inference, which is enabled by sentence-level optimization. Besides this, we would also like to explore whether sentence-level optimization would contribute to model training. With this aim, we compare the performances between GenDiE and $\text{GenDiE}_{\text{answer-level}}$, both with greedy decoding.

Fig. 3 plots the changes in the models' performances across different iterations. $\text{GenDiE}_{\text{answer-level}}$ displays opposite patterns in faithfulness on two benchmarks: it shows consistent improvement on ConFiQA but a decline on ASQA. This is primarily because the questions on ASQA typically demand longer and more detailed answers than on ConFiQA. On ConFiQA, the answer to a question can be an entity, while ASQA often requires paragraph-long answers.

GenDiE$_{\text{answer-level}}$, which lacks fine-grained supervision during training, underperforms in yielding faithful long-answers. Moreover, as evaluating faithfulness at answer level is more challenging, it is more difficult to ensure the reliability of self-scored contrastive answer pairs, leading to performance degradation during the self-evolving process. In comparison, GenDiE demonstrates increasing faithfulness and consistently outperforms GenDiE$_{\text{answer-level}}$, substantiating the effectiveness of sentence-level optimization.

Tab. 3 shows a comparative case. The generation by GenDiE$_{\text{answer-level}}$ is only partially faithful to the passages, with some sentences not being supported by the passages. While GenDiE provides a fully faithful and coherent answer.

[Question] When was the first Apple iPhone made?
[Passages] The iPhone (colloquially known as the iPhone 2G after 2008) is the first smartphone model designed and marketed by Apple Inc. After years of rumors and speculation, it was officially announced on January 9, 2007, and was later released in the United States on June 29, 2007. ... The original iPhone was introduced by Steve Jobs on January 9, 2007 in a keynote address at the Macworld Conference & Expo held in Moscone West in San Francisco, California. ...
[GenDiE generation] The first iPhone was released on June 29, 2007. It was announced by Steve Jobs on January 9, 2007 at the Macworld Conference & Expo held in Moscone West in San Francisco, California.
[GenDiE$_{\text{answer-level}}$ generation] The first iPhone was released on June 29, 2007. It was a revolutionary device that combined the features of a smartphone, a music player, and a camera into one sleek and stylish device. The iPhone was a game-changer in the mobile phone industry and paved the way for the development of smartphones as we know them today.

Table 3: Generation by GenDiE and GenDiE$_{\text{answer}}$ respectively. The questions are from ASQA. The text in red denotes the content that can not be supported by the provided passages.

5.4 Case Study

As illustrated in §4, the proposed hierarchical inference can significantly advance GenDiE's overall performance. This approach leverages the model's dual capabilities in sentence-level generation and scoring, thereby enabling efficient sentence-level search at each iteration. Tab. 4 provides an illustrative example of its execution.

6 Related Work

Faithfulness Although large language models (LLMs) have showcased impressive capabilities in various tasks (DeepSeek-AI et al., 2025; Zhang et al., 2024b; Minaee et al., 2024), they are often criticized for generating outputs that deviate from the provided contents, a phenomenon often termed *faithfulness hallucination* (Huang et al., 2024b; Zhang et al., 2023). This issue is particularly pronounced when *knowledge conflict* exists between model's parametric memory and the external evidence, as LLMs may overly rely on their internal priors (Jin et al., 2024). Many approaches have been proposed to improve contextual faithfulness of LLMs. CAD (Shi et al., 2024) leverages contrastive decoding to amplify the difference in output probabilities with and without context, reinforcing the model's attention to input context during inference. Self-RAG (Asai et al., 2024) trains models to selectively retrieve knowledge and reflect on retrieved information. Luo et al. (2023) introduces search-augmented instruction learning to ground LLM's generation on search results. Bi et al. (2024) aligns LLMs through DPO (Rafailov et al., 2024) with constructed faithful and stubborn responses. While effective, these approaches either intervene

[Question] Where is the Rogers Cup held in montreal?
[Generated sequences and scores after the 1st step]
| The Rogers Cup is held in Montreal at the Uniprix Stadium. [-0.176✓]
| The Rogers Cup is a tennis tournament held in Montreal, Quebec, Canada. [-0.242✓]
| The Rogers Cup is a tennis tournament that is held in both Montreal and Toronto. [-0.257×]
[Generated sequences and scores after the 2nd step]
| The Rogers Cup is held in Montreal at the Uniprix Stadium. The Uniprix Stadium is an outdoor hard court tennis stadium located in Montreal, Quebec, Canada. [-0.155✓]
| The Rogers Cup is held in Montreal at the Uniprix Stadium. The Uniprix Stadium is a tennis stadium located in Montreal, Quebec, Canada. [-0.174×]
| The Rogers Cup is held in Montreal at the Uniprix Stadium. The Uniprix Stadium is an outdoor hard court tennis stadium. [-0.186×]
| The Rogers Cup is a tennis tournament held in Montreal, Quebec, Canada. The men's event is held at the Uniprix Stadium, while the women's event is held at the IGA Stadium. [-0.167✓]
| The Rogers Cup is a tennis tournament held in Montreal, Quebec, Canada. The men's event is held at the Uniprix Stadium, while the women's event is held at the Aviva Centre. [-0.171×]
| The Rogers Cup is a tennis tournament held in Montreal, Quebec, Canada. The women's event is held at the Aviva Centre. [-0.190×]

Table 4: A running example of hierarchical inference with $M = 3$ and $N = 2$, where the passages are omitted for brevity. After the first step, the top-N sequences with the highest length-normalized faithfulness scores are selected by the sentence-level inference (marked with ✓). The second step begins with the selected sequences. After the second step, we have $N * M = 6$ candidate sequences, as the token-level inference produces M sequences for each prefix (a sequence at the second step and its corresponding prefix sequence from the first step share the same color). Again, the sentence-level inference selects the top-N sequences, which will serve as prefixes for the third step.

LLMs only at inference without addressing inherent limitations (Zhou et al., 2023), or fail to equip LLMs with self-judging abilities to explore the potential of self-improvement. Additionally, those answer-level optimization may overlook unfaithful details (Lai et al., 2024), particularly in long-form generation. In contrast, we propose a novel self-evolving framework with sentence-level optimization, enabling LLMs to enhance context faithfulness through self-generated and self-scored data, fostering iterative self-improvement. Another line of research aims to enhance the accuracy of passage citation in model-generated texts (Gao et al., 2023; Huang et al., 2024b; Ye et al., 2024; Aly et al., 2024), thereby increasing the trustworthiness of LLMs. However, these studies primarily focus on citation quality rather than improving the overall answer faithfulness to the input contexts.

Self-Evolving The field of self-evolving mechanisms for large language models (LLMs) is gaining traction as researchers seek to enhance model capabilities beyond current limitations. Self-evolving allows LLMs to autonomously improve and adapt to complex tasks without heavy reliance on human supervision. Huang et al. (2022) illustrates how LLMs can refine reasoning through self-generated rationale-augmented answers, thereby deepening their explanatory capabilities. Self-Align (Sun et al., 2023) proposes a principle-driven self-alignment model, trained from scratch and requiring little human annotation through self-generated data. Moreover, Self-Rewarding Language Models (Yuan et al., 2024) and MathShepherd (Wang et al., 2024) present mechanisms where models self-assign high-quality rewards, facilitating their own learning processes. Self-Evolved Reward Learning (Huang et al., 2024a) trains the reward model itself using the selected self-labeled data. Similarly, our approach employs a iterative self-training framework, allowing models to self-generate, self-score, and self-evolve through multiple training iterations.

7 Conclusion

We introduce GenDiE, a self-evolved approach for enhancing context-faithful generation. A distinctive feature of GenDiE is its capability for generating-then-self-scoring, which facilitates the model's self-evolution via iterative updates to the training data. Additionally, GenDiE functions at the sentence level, enabling fine-grained control over faithfulness. Experiments on benchmarks of long-form QA and counterfactual QA show that GenDiE achieves superior performances in both faithfulness and correctness over various baselines. We also verify the effectiveness of the key designs of GenDiE. GenDiE demonstrates promise in building self-evolved and trustworthy RAG system.

Acknowledgement

This study was supported by the Centre for Perceptual and Interactive Intelligence (CPII) Ltd., a CUHK-led InnoCentre under the InnoHK initiative of the Innovation and Technology Commission of the Hong Kong Special Administrative Region Government.

Limitations

GenDiE demonstrates notable performances on the benchmarks of long-form QA and counterfactual QA, which are two question-answering tasks with most intensive demand of faithful generation. Nevertheless, we need to show GenDiE can generalize to other QA tasks that require fast-changing world knowledge. Second, although we study the efficacy of multi-task training in §4, due to computational constraints, we did not conduct investigation of how other training objective for multi-task training, other than the ORPO-based one used in the paper, might affect the performance outcomes. our primary focus lies in achieving self-evolution through models endowed with generative and discriminative capabilities. Delving deeper into the effects of various multi-task training objectives presents a promising opportunity for advancing self-evolution Lastly, while hierarchical inference significantly enhances performance, it also introduces additional computational overhead compared to standard inference methods, stemming from the multiple generations and explorations required for each sentence. Further efforts to reduce this overhead can be pursued through algorithm optimization. Nevertheless, as observed in other research on LLM reasoning, we should recognize that this test-time scaling yields benefits that substantially outweigh the associated overhead.

Ethics Statement

While our framework significantly enhances LLMs' ability to generate contextually faithful responses through self-evolving sentence-level optimization,

we emphasize critical ethical considerations. Improved faithfulness scoring does not inherently guarantee factual correctness, as even self-scored "faithful" propositions may inherit biases or contextual omissions from training data. Our work is dedicated to maintaining ethical integrity, ensuring openness in methodology, and advancing the ethical application of AI innovations for societal good.

References

Rami Aly, Zhiqiang Tang, Samson Tan, and George Karypis. 2024. Learning to generate answers with citations via factual consistency models. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 11876–11896, Bangkok, Thailand. Association for Computational Linguistics.

Anthropic. 2025. Claude 3.5 sonnet.

Akari Asai, Zeqiu Wu, Yizhong Wang, Avirup Sil, and Hannaneh Hajishirzi. 2024. Self-RAG: Learning to retrieve, generate, and critique through self-reflection. In *The Twelfth International Conference on Learning Representations*.

Baolong Bi, Shaohan Huang, Yiwei Wang, Tianchi Yang, Zihan Zhang, Haizhen Huang, Lingrui Mei, Junfeng Fang, Zehao Li, Furu Wei, Weiwei Deng, Feng Sun, Qi Zhang, and Shenghua Liu. 2024. Context-dpo: Aligning language models for context-faithfulness. *Preprint*, arXiv:2412.15280.

DeepSeek-AI, Daya Guo, Dejian Yang, Haowei Zhang, Junxiao Song, Ruoyu Zhang, Runxin Xu, Qihao Zhu, Shirong Ma, Peiyi Wang, Xiao Bi, Xiaokang Zhang, Xingkai Yu, Yu Wu, Z. F. Wu, Zhibin Gou, Zhihong Shao, Zhuoshu Li, Ziyi Gao, Aixin Liu, Bing Xue, Bingxuan Wang, Bochao Wu, Bei Feng, Chengda Lu, Chenggang Zhao, Chengqi Deng, Chenyu Zhang, Chong Ruan, Damai Dai, Deli Chen, Dongjie Ji, et al. 2025. Deepseek-r1: Incentivizing reasoning capability in llms via reinforcement learning. *Preprint*, arXiv:2501.12948.

Tim Dettmers, Artidoro Pagnoni, Ari Holtzman, and Luke Zettlemoyer. 2023. Qlora: Efficient finetuning of quantized llms. *ArXiv*, abs/2305.14314.

Angela Fan, Yacine Jernite, Ethan Perez, David Grangier, Jason Weston, and Michael Auli. 2019. ELI5: Long form question answering. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 3558–3567, Florence, Italy. Association for Computational Linguistics.

Wenqi Fan, Yujuan Ding, Liangbo Ning, Shijie Wang, Hengyun Li, Dawei Yin, Tat-Seng Chua, and Qing Li. 2024. A survey on rag meeting llms: Towards retrieval-augmented large language models. In *Proceedings of the 30th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, KDD '24, page 6491–6501, New York, NY, USA. Association for Computing Machinery.

Tianyu Gao, Howard Yen, Jiatong Yu, and Danqi Chen. 2023. Enabling large language models to generate text with citations. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 6465–6488, Singapore. Association for Computational Linguistics.

Yunfan Gao, Yun Xiong, Xinyu Gao, Kangxiang Jia, Jinliu Pan, Yuxi Bi, Yi Dai, Jiawei Sun, Meng Wang, and Haofen Wang. 2024. Retrieval-augmented generation for large language models: A survey. *Preprint*, arXiv:2312.10997.

Aaron Grattafiori, Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Alex Vaughan, et al. 2024. The llama 3 herd of models. *Preprint*, arXiv:2407.21783.

Jiwoo Hong, Noah Lee, and James Thorne. 2024. ORPO: Monolithic preference optimization without reference model. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 11170–11189, Miami, Florida, USA. Association for Computational Linguistics.

Or Honovich, Roee Aharoni, Jonathan Herzig, Hagai Taitelbaum, Doron Kukliansy, Vered Cohen, Thomas Scialom, Idan Szpektor, Avinatan Hassidim, and Yossi Matias. 2022. TRUE: Re-evaluating factual consistency evaluation. In *Proceedings of the Second DialDoc Workshop on Document-grounded Dialogue and Conversational Question Answering*, pages 161–175, Dublin, Ireland. Association for Computational Linguistics.

J. Edward Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, and Weizhu Chen. 2021. Lora: Low-rank adaptation of large language models. *ArXiv*, abs/2106.09685.

Chenghua Huang, Zhizhen Fan, Lu Wang, Fangkai Yang, Pu Zhao, Zeqi Lin, Qingwei Lin, Dongmei Zhang, S. Rajmohan, and Qi Zhang. 2024a. Self-evolved reward learning for llms. *ArXiv*, abs/2411.00418.

Chengyu Huang, Zeqiu Wu, Yushi Hu, and Wenya Wang. 2024b. Training language models to generate text with citations via fine-grained rewards. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2926–2949, Bangkok, Thailand. Association for Computational Linguistics.

Jiaxin Huang, Shixiang Shane Gu, Le Hou, Yuexin Wu, Xuezhi Wang, Hongkun Yu, and Jiawei Han. 2022. Large language models can self-improve. *arXiv preprint arXiv:2210.11610*.

Zhuoran Jin, Pengfei Cao, Yubo Chen, Kang Liu, Xiaojian Jiang, Jiexin Xu, Qiuxia Li, and Jun Zhao. 2024.

Tug-of-war between knowledge: Exploring and resolving knowledge conflicts in retrieval-augmented language models. *Preprint*, arXiv:2402.14409.

Xin Lai, Zhuotao Tian, Yukang Chen, Senqiao Yang, Xiangru Peng, and Jiaya Jia. 2024. Step-dpo: Step-wise preference optimization for long-chain reasoning of llms. *Preprint*, arXiv:2406.18629.

Kun Li, Tianhua Zhang, Xixin Wu, Hongyin Luo, James Glass, and Helen Meng. 2024. Decoding on graphs: Faithful and sound reasoning on knowledge graphs through generation of well-formed chains. *Preprint*, arXiv:2410.18415.

Ilya Loshchilov and Frank Hutter. 2017. Decoupled weight decay regularization. In *International Conference on Learning Representations*.

Hongyin Luo, Yung-Sung Chuang, Yuan Gong, Tianhua Zhang, Yoon Kim, Xixin Wu, Danny Fox, Helen Meng, and James Glass. 2023. Sail: Search-augmented instruction learning. *Preprint*, arXiv:2305.15225.

Sewon Min, Julian Michael, Hannaneh Hajishirzi, and Luke Zettlemoyer. 2020. AmbigQA: Answering ambiguous open-domain questions. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 5783–5797, Online. Association for Computational Linguistics.

Shervin Minaee, Tomas Mikolov, Narjes Nikzad, Meysam Chenaghlu, Richard Socher, Xavier Amatriain, and Jianfeng Gao. 2024. Large language models: A survey. *Preprint*, arXiv:2402.06196.

Xuan-Phi Nguyen, Shrey Pandit, Senthil Purushwalkam, Austin Xu, Hailin Chen, Yifei Ming, Zixuan Ke, Silvio Savarese, Caiming Xong, and Shafiq Joty. 2024. Sfr-rag: Towards contextually faithful llms. *Preprint*, arXiv:2409.09916.

Jianmo Ni, Chen Qu, Jing Lu, Zhuyun Dai, Gustavo Hernández Ábrego, Ji Ma, Vincent Y. Zhao, Yi Luan, Keith B. Hall, Ming-Wei Chang, and Yinfei Yang. 2021. Large dual encoders are generalizable retrievers. *Preprint*, arXiv:2112.07899.

OpenAI. 2024. Hello gpt-4o.

Rafael Rafailov, Archit Sharma, Eric Mitchell, Stefano Ermon, Christopher D. Manning, and Chelsea Finn. 2024. Direct preference optimization: Your language model is secretly a reward model. *Preprint*, arXiv:2305.18290.

Weijia Shi, Xiaochuang Han, Mike Lewis, Yulia Tsvetkov, Luke Zettlemoyer, and Wen-tau Yih. 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)*, pages 783–791, Mexico City, Mexico. Association for Computational Linguistics.

Ivan Stelmakh, Yi Luan, Bhuwan Dhingra, and Ming-Wei Chang. 2022. ASQA: Factoid questions meet long-form answers. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 8273–8288, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Alessandro Stolfo. 2024. Groundedness in retrieval-augmented long-form generation: An empirical study. In *Findings of the Association for Computational Linguistics: NAACL 2024*, pages 1537–1552, Mexico City, Mexico. Association for Computational Linguistics.

Zhiqing Sun, Yikang Shen, Qinhong Zhou, Hongxin Zhang, Zhenfang Chen, David D. Cox, Yiming Yang, and Chuang Gan. 2023. Principle-driven self-alignment of language models from scratch with minimal human supervision. *ArXiv*, abs/2305.03047.

Cunxiang Wang, Xiaoze Liu, Yuanhao Yue, Xiangru Tang, Tianhang Zhang, Cheng Jiayang, Yunzhi Yao, Wenyang Gao, Xuming Hu, Zehan Qi, Yidong Wang, Linyi Yang, Jindong Wang, Xing Xie, Zheng Zhang, and Yue Zhang. 2023. Survey on factuality in large language models: Knowledge, retrieval and domain-specificity. *Preprint*, arXiv:2310.07521.

Peiyi Wang, Lei Li, Zhihong Shao, Runxin Xu, Damai Dai, Yifei Li, Deli Chen, Yu Wu, and Zhifang Sui. 2024. Math-shepherd: Verify and reinforce llms step-by-step without human annotations. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 9426–9439.

Jian Xie, Kai Zhang, Jiangjie Chen, Renze Lou, and Yu Su. 2024. Adaptive chameleon or stubborn sloth: Revealing the behavior of large language models in knowledge conflicts. In *The Twelfth International Conference on Learning Representations*.

Fangyuan Xu, Yixiao Song, Mohit Iyyer, and Eunsol Choi. 2023. A critical evaluation of evaluations for long-form question answering. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 3225–3245, Toronto, Canada. Association for Computational Linguistics.

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

Xi Ye, Ruoxi Sun, Sercan Arik, and Tomas Pfister. 2024. Effective large language model adaptation for improved grounding and citation generation. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 6237–6251, Mexico City, Mexico. Association for Computational Linguistics.

Wenhao Yu, Dan Iter, Shuohang Wang, Yichong Xu, Mingxuan Ju, Soumya Sanyal, Chenguang Zhu, Michael Zeng, and Meng Jiang. 2023. Generate rather than retrieve: Large language models are strong context generators. *Preprint*, arXiv:2209.10063.

Weizhe Yuan, Richard Yuanzhe Pang, Kyunghyun Cho, Sainbayar Sukhbaatar, Jing Xu, and Jason Weston. 2024. Self-rewarding language models. *arXiv preprint arXiv:2401.10020*.

Yuheng Zha, Yichi Yang, Ruichen Li, and Zhiting Hu. 2023. AlignScore: Evaluating factual consistency with a unified alignment function. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 11328–11348, Toronto, Canada. Association for Computational Linguistics.

Longhui Zhang, Yanzhao Zhang, Dingkun Long, Pengjun Xie, Meishan Zhang, and Min Zhang. 2024a. A two-stage adaptation of large language models for text ranking. In *Findings of the Association for Computational Linguistics: ACL 2024*, pages 11880–11891, Bangkok, Thailand. Association for Computational Linguistics.

Tianhua Zhang, Jiaxin Ge, Hongyin Luo, Yung-Sung Chuang, Mingye Gao, Yuan Gong, Yoon Kim, Xixin Wu, Helen Meng, and James Glass. 2024b. Natural language embedded programs for hybrid language symbolic reasoning. In *Findings of the Association for Computational Linguistics: NAACL 2024*, pages 4131–4155, Mexico City, Mexico. Association for Computational Linguistics.

Yue Zhang, Yafu Li, Leyang Cui, Deng Cai, Lemao Liu, Tingchen Fu, Xinting Huang, Enbo Zhao, Yu Zhang, Yulong Chen, Longyue Wang, Anh Tuan Luu, Wei Bi, Freda Shi, and Shuming Shi. 2023. Siren's song in the ai ocean: A survey on hallucination in large language models. *Preprint*, arXiv:2309.01219.

Wenxuan Zhou, Sheng Zhang, Hoifung Poon, and Muhao Chen. 2023. Context-faithful prompting for large language models. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 14544–14556, Singapore. Association for Computational Linguistics.

A Method Details

A.1 Quality Control

Due to extensive pre-training on massive corpora, LLMs encapsulate substantial world knowledge within their parameters (Yu et al., 2023; Wang et al., 2023). As a result, it is possible that θ_0 can answer the question correctly and faithfully even without access to the evidence passages, provided it has encountered similar information during pre-training. To mitigate this, we involve a heuristic negative sample filtering process for quality control. Given an input sequence x, we denote the (length-normalized) negative log-likelihood (NLL) loss in generating the output sequence y with m tokens as follows:

$$\mathcal{L}(y|x) = -\frac{1}{m}\sum_{t=1}^{m}\log P(y_t|x, y_{<t}) \quad (3)$$

For each candidate training instance $\{(q, A_{\prec a}, a, a^{'})\}$, we compare the NLL loss reduction ratio in generating a candidate with and without input passages to select valid negative sentences:

$$\frac{\mathcal{L}([A_{\prec a}, a]|q, P) - \mathcal{L}([A_{\prec a}, a]|q)}{\mathcal{L}([A_{\prec a}, a]|q)} < \frac{\mathcal{L}([A_{\prec a}, a^{'}]|q, P) - \mathcal{L}([A_{\prec a}), a^{'}]|q)}{\mathcal{L}([A_{\prec a}, a^{'}]|q)} \quad (4)$$

where $[*]$ denotes the concatenation. The motivation is to ensure that the selected negative samples a' remain less faithful to the input question than the positive answer a, even when evidence passages are provided. This filtering process tries to preserve the integrity of our faithfulness relation between positive and negative sentences.

	AlignScore	T5NLI	EM Rec.	Hit
non-filtered	68.55	62.02	42.81	17.93
filtered	**70.46**	**63.02**	**43.11**	**18.46**

Table 5: Performance comparison for pre-stage with and without filtering on ASQA dataset.

The effectiveness of pre-stage filtering is reported in Tab. 5. For each training item, we use model θ_0 to sample six negative sentence candidates, as described in §2.2.1. After filtering via Eq. 4, we retain at most two of the most negative sentences, ranked by the NLL loss reduction ratio, for `filtered` setting (some items may contain only one valid negative sentence after filtering). In contrast, the `non-filtered` setting uses two randomly selected sentences from the six candidates. As demonstrated in the Tab. 5, heuristic negative sample filtering improves both faithfulness and correctness, with `filtered` outperforming `non-filterd` across all metrics.

B Implementation Details

B.1 Data Statistics

Tab. 6 shows the statistics of the two benchmarks. For ASQA, we refine the original 4353 instances

by retaining only 3414 where the top-5 passages contain at least one candidate answer from the given reference list. This ensures that the question is at least partially answerable by the input passages. Truncating the answer sentences yields 14841 items, from which we retain 11531 valid items after filtering for model training.

Dataset	Train (orig. ans./kept ans./trunc. sent./final sent.)	Test
ASQA	4353/3414/14841/11531	948
ConFiQA	-	18000

Table 6: Data statistics for two benchmarks.

B.2 Training Details

We use AdamW (Loshchilov and Hutter, 2017) optimizer with learning rates of 1e-5 . The learning rates undergo a warmup of 10% of overall training steps, followed by a linear decrease until 0. We utilize quantized LoRA (QLoRA) (Hu et al., 2021; Dettmers et al., 2023) as the parameter-efficient fine-tuning technique to train the models with an NVIDIA A6000 GPU. Specifically, QLoRA is implemented on the query and value attention matrices within each decoder block, using a fixed rank of 8, a scaling factor of 16, and a dropout rate of 0.05. The model weights are quantized and loaded in 4-bit NormalFloat format.

C Additional Results

C.1 GenDiE vs. Context-DPO

To further demonstrate the effectiveness of our approach, we compare GenDiE with Context-DPO (Bi et al., 2024), another approach that employs the discriminative objective. Context-DPO was specifically trained with constructed faithful and stubborn responses on ConFiQA and cannot be directly trained on ASQA due to the requirement for preference pairs needed to cold-start the training process. We then use the official checkpoint provided by Bi et al. (2024) and evaluate both methods on ASQA and ConFiQA datasets. Notably, the domain adaptation settings differ: Context-DPO treats ConFiQA as in-domain and ASQA as out-of-domain, whereas our approach adopts the opposite configuration. As shown in Tab. 7, **GenDiE with greedy search consistently outperforms Context-DPO across all metrics on both two datasets.** Remarkably, Context-DPO underperforms our method even when evaluated in its own in-domain setting (ConFiQA), while GenDiE demonstrates stronger generalization when trained on ASQA and tested on ConFiQA.

Dataset	Method	AlignScore	T5NLI	EM Rec.	Hit
ASQA	Context-DPO	63.21	61.06	35.09	10.23
	Ours (Greedy)	73.87	69.16	43.61	18.57
ConFiQA	Context-DPO	70.24	55.22	-	70.26
	Ours (Greedy)	72.32	70.17	-	73.72

Table 7: Performance comparison between Context-DPO (Bi et al., 2024) and GenDiE (greedy search).

C.2 In-Domain Prompting on ConFiQA

In Tab. 1, we report the in-context prompting results of `gpt-4o` and `Llama-3.1-8b-Instruct` using the same prompt (Tab. 9) on both ASQA and ConFiQA datasets to simulate a realistic *out-of-domain* setting on ConFiQA for all approaches. We report additional ***in-domain* prompting** results on ConFiQA using dataset-specific in-context examples in Tab. 8. The corresponding prompt is presented in Tab. 10. Notably, our approach maintains superior performance even in the challenging domain transfer setting (trained on ASQA, evaluated on ConFiQA), demonstrating its effectiveness and robustness.

Method	Alignscore	T5NLI	Hit
Prompting: `gpt-4o` (*ood*)	44.56	34.11	35.96
Prompting: `gpt-4o` (*in*)	61.97	35.72	44.98
Prompting: `llama3.1-8B` (*ood*)	50.67	30.78	55.14
Prompting: `llama3.1-8B` (*in*)	65.62	38.41	67.51
GenDiE: Greedy search (*ood*)	72.32	70.17	73.72
GenDiE: Vanilla beam search (*ood*)	78.54	69.33	80.95
GenDiE: Hier inference (*ood*)	80.73	80.69	84.63

Table 8: Performance on ConFiQA dataset. Results for prompting include both *in-domain* (*in*) and *out-of-domain* (*ood*) setting. GenDiE is evaluated under *out-of-domain* setting.

D Prompts

Tab. 9 presents the prompt used for `In-context Prompting` approach over the two benchmarks in Tab. 1. Two in-context examples are listed, each consisting of five input passages, which matches the setting of the test questions. The in-context learning examples are sourced from the ASQA dataset, following previous approaches (Aly et al., 2024). We intentionally exclude in-domain demonstrations for ConFiQA to simulate the *out-of-domain* evaluation setting. For completeness, Tab. 10 presents the prompt used for *in-domain* ConFiQA evaluation in Tab. 8.

Instruction: Write an accurate, engaging, fluent and detailed answer for the given question using only the provided search results.

Question: Which is the most rainy place on earth?

Document [1](Title: Cherrapunji): Cherrapunji Cherrapunji (; with the native name Sohra being more commonly used, and can also be spelled Cherrapunjee or Cherrapunji) is a subdivisional town in the East Khasi Hills district in the Indian state of Meghalaya. It is the traditional capital of aNongkhlaw "hima" (Khasi tribal chieftainship constituting a petty state), both known as Sohra or Churra. Cherrapunji has often been credited as being the wettest place on Earth, but for now nearby Mawsynram currently holds that distinction. Cherrapunji still holds the all-time record for the most rainfall in a calendar month for July 1861 and most rain in a year from August 1860 to July 1861, however: it received in
Document [2](Title: Cherrapunji): Radio relay station known as Akashvani Cherrapunji. It broadcasts on FM frequencies. Cherrapunji Cherrapunji (; with the native name Sohra being more commonly used, and can also be spelled Cherrapunjee or Cherrapunji) is a subdivisional town in the East Khasi Hills district in the Indian state of Meghalaya. It is the traditional capital of aNongkhlaw "hima" (Khasi tribal chieftainship constituting a petty state), both known as Sohra or Churra. Cherrapunji has often been credited as being the wettest place on Earth, but for now nearby Mawsynram currently holds that distinction. Cherrapunji still holds the all-time record for the most rainfall
Document [3](Title: Mawsynram): Mawsynram Mawsynram () is a village in the East Khasi Hills district of Meghalaya state in north-eastern India, 65 kilometres from Shillong. Mawsynram receives one of the highest rainfalls in India. It is reportedly the wettest place on Earth, with an average annual rainfall of 11,872 mm, but that claim is disputed by Lloró, Colombia, which reported an average yearly rainfall of 12,717 mm between 1952 and 1989 and López de Micay, also in Colombia, which reported an annual 12,892 mm per year between 1960 and 2012. According to the "Guinness Book of World Records", Mawsynram received of rainfall in 1985. Mawsynram is located at 25° 18'
Document [4](Title: Earth rainfall climatology): Pacific Northwest, and the Sierra Nevada range are the wetter portions of the nation, with average rainfall exceeding per year. The drier areas are the Desert Southwest, Great Basin, valleys of northeast Arizona, eastern Utah, central Wyoming, eastern Oregon and Washington and the northeast of the Olympic Peninsula. The Big Bog on the island of Maui receives, on average, every year, making it the wettest location in the US, and all of Oceania. The annual average rainfall maxima across the continent lie across the northwest from northwest Brazil into northern Peru, Colombia, and Ecuador, then along the Atlantic coast of
Document [5](Title: Going to Extremes): in the world. Oymyakon in Siberia, where the average winter temperature is -47 °F (-44 °C). Arica in Chile, where there had been fourteen consecutive years without rain. Fog is the only local source of water. Mawsynram in India, where average annual rainfall is 14 meters, falling within a four-month period in the monsoon season. The rainfall is approximately equal to that of its neighbor Cherrapunji. Dallol in Ethiopia, known as the 'Hell-hole of creation' where the temperature averages 94 °F (34 °C) over the year. In his second series, Middleton visited places without permanent towns, locations where "survival"

Answer: Several places on Earth claim to be the most rainy, such as Lloró, Colombia, which reported an average annual rainfall of 12,717 mm between 1952 and 1989, and López de Micay, Colombia, which reported an annual 12,892 mm between 1960 and 2012. However, the official record is held by Mawsynram, India with an average annual rainfall of 11,872 mm, although nearby town Sohra, India, also known as Cherrapunji, holds the record for most rain in a calendar month for July 1861 and most rain in a year from August 1860 to July 1861.

Question: When did the us break away from england?

Document [1](Title: United States withdrawal from Saudi Arabia): United States withdrawal from Saudi Arabia Beginning during Operation Desert Shield in August 1990, while preparing for the Gulf War, the United States sent a large troop contingent to Saudi Arabia. After the war, remnant troops, primarily U.S. Air Force personnel, augmented by a smaller number of coordinating and training personnel from the U.S. Navy, U.S. Army and U.S. Marine Corps remained in Saudi Arabia under the aegis of Joint Task Force Southwest Asia (JTF-SWA), as part of Operation Southern Watch (OSW). The United Kingdom and France also maintained a small contingent of Royal Air Force and French Air Force
Document [2](Title: Decolonization of the Americas): and France has fully "integrated" most of its former colonies as fully constituent "departments" of France. The United States of America declared independence from Great Britain on July 2, 1776 (although the event is now commemorated on July 4, the date when the Declaration of Independence was officially adopted by Congress), in so doing becoming the first independent, foreign-recognized nation in the Americas and the first European colonial entity to break from its mother country. Britain formally acknowledged American independence in 1783 after its defeat in the American Revolutionary War. Although initially occupying only the land east of the Mississippi
Document [3](Title: American Revolution): second British army at Yorktown in the fall of 1781, effectively ending the war. The Treaty of Paris was signed September 3, 1783, formally ending the conflict and confirming the new nation's complete separation from the British Empire. The United States took possession of nearly all the territory east of the Mississippi River and south of the Great Lakes, with the British retaining control of Canada and Spain taking Florida. Among the significant results of the revolution was the creation of the United States Constitution, establishing a relatively strong federal national government that included an executive, a national judiciary, and
Document [4](Title: Decolonization): accelerate decolonialization and bring an end to the colonial empires of its Western allies, most importantly during the 1956 Suez Crisis, but American military bases were established around the world and direct and indirect interventions continued in Korea, Indochina, Latin America ("inter alia", the 1965 occupation of the Dominican Republic), Africa, and the Middle East to oppose Communist invasions and insurgencies. Since the dissolution of the Soviet Union, the United States has been far less active in the Americas, but invaded Afghanistan and Iraq following the September 11 attacks in 2001, establishing army and air bases in Central Asia. Before
Document [5](Title: Decolonization): the responsibility of the United Kingdom (with a copy of the new constitution annexed), and finally, if approved, issuance of an Order of Council fixing the exact date of independence. After World War I, several former German and Ottoman territories in the Middle East, Africa, and the Pacific were governed by the UK as League of Nations mandates. Some were administered directly by the UK, and others by British dominions – Nauru and the Territory of New Guinea by Australia, South West Africa by the Union of South Africa, and Western Samoa by New Zealand. Egypt became independent in 1922,

Answer: The United States took the first step towards gaining independence from Great Britain when it declared independence from Great Britain on July 2, 1776 (although the event is now commemorated on July 4, 1776, the date when the Declaration of Independence was officially adopted by Congress). The Treaty of Paris was later signed on September 3, 1783, formally separating the United States from the British Empire.

Question: `{test_question}`

Documents: `{test_documents}`

Answer:

Table 9: Prompt used for `In-context Prompting` approach in Tab. 1.

Instruction: Answer the question using only the provided search results.

Question: Who is the composer of Bad Boys for Life?

Document: Bad Boys for Life is a 2020 action comedy film directed by Adil El Arbi and Bilall Fallah, and composed by Petri Alanko. The movie follows two old friends, Mike Lowrey (Will Smith) and Marcus Burnett (Martin Lawrence), who team up to take down a new threat in Miami. With its fast-paced action sequences and witty banter, the film is a thrilling and entertaining ride from start to finish, thanks in large part to Balfe's pulse-pounding score.

Answer: Petri Alanko

Question: What is The Technical Cooperation Program that Mark Cavendish is a member of?

Document: Mark Cavendish is a British professional road racing cyclist who has won numerous stages in every Grand Tour, including the Tour de France, Giro d'Italia, and Vuelta a Espa0̆0f1a. He is known for his explosive sprinting ability and has been named the world's best sprinter multiple times. Cavendish has represented Great Britain in the Olympic Games and has won numerous national and international titles throughout his career. The United Kingdom is a member of American Academy of Arts and Sciences, a global initiative aimed at promoting sustainable development and economic growth through technical cooperation. The program brings together countries from around the world to share knowledge, expertise, and resources in order to tackle common challenges and achieve shared goals. As a member of American Academy of Arts and Sciences, the United Kingdom is committed to collaborating with other countries to promote economic development, social progress, and environmental sustainability.

Answer: American Academy of Arts and Sciences

Question: `{test_question}`

Documents: `{test_documents}`

Answer:

Table 10: In-Domain Prompt used for ConFiQA dataset.