FineRef: Fine-Grained Error Reflection and Correction for Long-Form Generation with Citations

Yixing Peng[1,2], Licheng Zhang[1*], Shancheng Fang[3], Yi Liu[2], Peijian Gu[1], Quan Wang[4]

[1]University of Science and Technology of China
[2]State Key Laboratory of Communication Content Cognition, People's Daily Online
[3]Shenzhen University
[4]Beijing University of Posts and Telecommunications
{xk98, zlczlc, gpj123}@mail.ustc.edu.cn, fangsc@szu.edu.cn, gavin1332@gmail.com, wangquan@bupt.edu.cn

Abstract

Generating with citations is crucial for trustworthy Large Language Models (LLMs), yet even advanced LLMs often produce mismatched or irrelevant citations. Existing methods over-optimize citation fidelity while overlooking relevance to the user query, which degrades answer quality and robustness in real-world settings with noisy or irrelevant retrieved content. Moreover, the prevailing single-pass paradigm struggles to deliver optimal answers in long-form generation that requiring multiple citations. To address these limitations, we propose **FineRef**, a framework based on **Fine**-grained error **Ref**lection, which explicitly teaches the model to self-identify and correct two key citation errors—mismatch and irrelevance—on a per-citation basis. FineRef follows a two-stage training strategy. The first stage instills an "attempt–reflect–correct" behavioral pattern via supervised fine-tuning, using fine-grained and controllable reflection data constructed by specialized lightweight models. An online self-reflective bootstrapping strategy is designed to improve generalization by iteratively enriching training data with verified, self-improving examples. To further enhance the self-reflection and correction capability, the second stage applies process-level reinforcement learning with a multi-dimensional reward scheme that promotes reflection accuracy, answer quality, and correction gain. Experiments on the ALCE benchmark demonstrate that FineRef significantly improves both citation performance and answer accuracy. Our 7B model outperforms GPT-4 by up to 18% in Citation F1 and 4% in EM Recall, while also surpassing the state-of-the-art model across key evaluation metrics. FineRef also exhibits strong generalization and robustness in domain transfer settings and noisy retrieval scenarios.

Introduction

Large Language Models (LLMs) (Brown et al. 2020; Achiam et al. 2023) excel at generating long-form response to user queries, yet they are often prone to hallucinations that producing factually incorrect content. Retrieval-Augmented Generation (RAG) (Borgeaud et al. 2022; Izacard et al. 2023) has emerged as a promising solution by grounding responses in external passages, significantly improving factuality and knowledge coverage. Building on this paradigm,

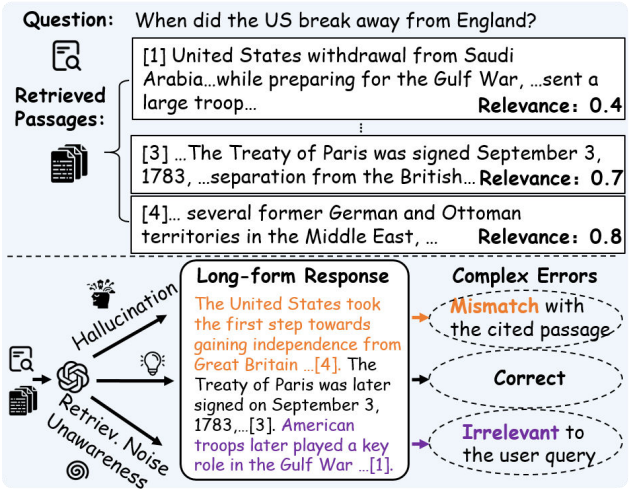


Figure 1: An example of citation errors in generated response: orange indicates citations that do not match the referenced passage (mismatch), while purple denotes citations that are irrelevant to the query (irrelevance).

generation with citations has become a core requirement for trustworthy language systems (Gao et al. 2023; Yue et al. 2023). By explicitly requiring the model to attach citations in its responses to the supporting retrieved passages, this paradigm (1) enhances answer verifiability by enabling users to trace the sources underlying the model's output, and (2) constrains the generation to more faithfully align with the retrieved evidence, thereby mitigating hallucinations. However, advanced LLMs and commercial chat engines still often produce unsupported or inaccurate citations in practice (Liu, Zhang, and Liang 2023; Gao et al. 2023), which undermines the credibility of their outputs and limits their applicability in high-trust domains such as healthcare, law, and scientific research.

Progress on these challenges has been hindered by annotation difficulty. While several efforts have explored training-free methods such as in-context learning (Gao et al. 2023; Li et al. 2023) and Chain-of-Thought (CoT) (Ji et al. 2024), their performance remains limited. Differently, the

*Corresponding author

work represented by (Aly et al. 2024) adopts iterative self-training by selecting citation-accurate samples from the model's own outputs. Nevertheless, these approaches still suffer from several key limitations. First, existing work typically trains only on citation-accurate samples, which helps mitigate mismatched citations—generating statements that are not supported by the attributed sources. However, this implicitly assumes that all retrieved passages are relevant to the query, limiting the model's ability to identify irrelevant content. In real-world scenarios, retrieval results often contain noisy or unrelated passages, leading to irrelevant citations, where the model cites passages unrelated to the query. This distribution gap between clean training data and noisy application scenarios significantly weakens model robustness and answer quality. Second, prevailing approaches adopt a single-pass generation paradigm, where the model is required to produce a fully cited response in one step. While this simplifies the generation process, it is inadequate for complex real-world scenarios, particularly long-form generation tasks that demand multiple, contextually appropriate citations and are prone to diverse citation errors. Without mechanisms for planning or revision, the model struggles to ensure citation correctness within a single pass.

To address these challenges, a promising solution is to introduce a self-reflection mechanism that enables the model to proactively identify and revise citation errors after generation. Prior work in other domains (Madaan et al. 2023; Shinn et al. 2023) typically adopts coarse-grained, answer-level reflection. However, in long-form generation tasks involving multiple citations, such coarse-grained reflection makes it difficult to locate erroneous citations and distinguish error types. Moreover, when reflection signals are generated directly by general-purpose LLMs, they often exhibit inconsistency in both accuracy and structure, as distinct decisions can be made for the same input, making them unreliable as training supervision. Therefore, there is a pressing need to develop a fine-grained, controllable reflection framework to support precise error identification and correction.

To this end, we propose a framework for generation with citations based on **Fine**-grained error **Ref**lection (**FineRef**), which endows the LLM with citation-level self-reflection capabilities, enabling it to explicitly identify and correct two common and impactful citation errors - mismatch and irrelevance - from its own initial attempt. FineRef adopts a two-stage training strategy, as shown in Figure 2, including (1) **behavioral pattern learning** that apply supervised fine-tuning to teach the attempt–reflect–correct behavioral pattern for citation generation, (2) **process-level reinforcement learning** to enhance the model's reflection and correction capabilities. In the first stage, to construct behavioral data, we first prompt the initial model attempt to generate response with citations. We then leverage a specialized yet lightweight factual consistency model (FCM) (Kryściński et al. 2020) and a reranker (Xiao et al. 2023) to identify mismatch and irrelevance errors for each citation in the model-generated attempts and automatically construct fine-grained, accurate reflection signals. Correction data is generated using advanced LLM based on these signals, and high-quality attempt–reflection–correction chains are subsequently constructed and curated for training. To improve generalization, we propose an **online self-reflective bootstrapping strategy** where the model generates full attempt–reflect–correct chains during training, and retains those that successfully identify and fix citation errors, exposing the model to diverse reflection patterns and enhancing its adaptability to complex citation scenarios. In the second stage, we conceptualize the generation process as multiple sub-behaviors and design a **multidimensional reward scheme** given the distinct objectives of each sub-behavior. For reflection behavior, we introduce a reward for reflection accuracy, encouraging the model to correctly identify citation errors. For the attempt and correction behaviors, we apply rewards for citation and answer quality, and further incorporate a correction gain reward to incentivize improvements of the correction over the initial attempt. This approach (1) overcomes the limitations of existing methods that focus solely on citation fidelity, and enhances citation quality while preserving overall question-answering performance. (2) Through this reflect-correct process, our framework significantly improves the robustness in complex citation scenarios.

We conduct experiments on ALCE (Gao et al. 2023), a few-shot benchmark for citation generation, which includes two challenging long-form answering datasets: ASQA and ELI5. We evaluate using different LLMs. Based on only four annotated examples, our 7B model achieves up to an 18% improvement in Citation F1 over GPT-4 (Achiam et al. 2023), while also enhancing answer performance by up to 4% in EM Recall. Moreover, FineRef significantly outperforms the state-of-the-art model CALF (Aly et al. 2024) across a range of backbone LLMs. Furthermore, consistently strong performance of our method under domain transfer and noisy retrieval conditions highlights its generalization and robustness capabilities.

Task Formulation

Given a question q and a set of retrieved passages $P = \{p_1, \ldots, p_m\}$, the goal is to generate a long-form answer $\hat{y} = \{s_1, \ldots, s_n\}$. In long-form answer generation with citations, each generated sentence s_i is expected to cite one or more relevant passages $C_i \subseteq P$ (denoted by bracketed indices, e.g., "[1]" referring to p_1), such that the content of s_i is entailed by the cited passages C_i. The task requires the information in s_i to originate from the attributed/cited passages C_i, ensuring that $\hat{y}$ is fully verifiable by P. Moreover, the overall answer $\hat{y}$ is required to be factually correct and responsive to the question q, going beyond citation fidelity to ensure the answer's informativeness and relevance.

Method

FineRef is a framework for generation with citations that models the process of identifying and correcting citation errors through fine-grained self-reflection. As illustrated in Figure 2, FineRef employs a two-stage training pipeline: (1) **Behavioral pattern learning**, which instills the attempt–reflect–correct behavior via supervised learning, and (2) **Process-level reinforcement learning**, which optimizes

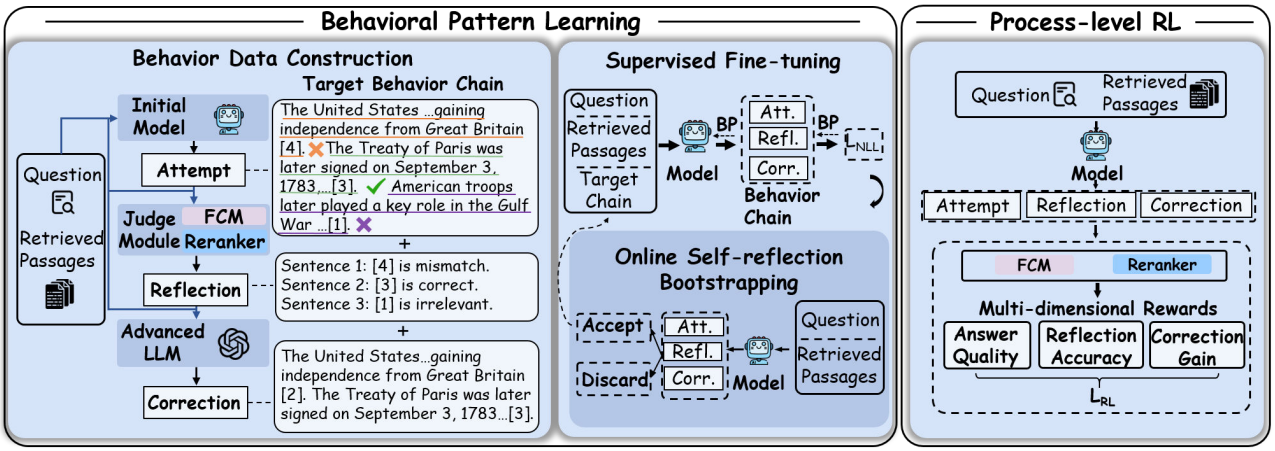


Figure 2: FineRef involves two training stages: (1) *Behavior pattern learning* stage, where the model is supervised to generate the "attempt–reflection–correction" chain using fine-grained reflection data constructed via specialized FCM and reranker models, followed by online reflection bootstrapping, improving from self-generated reflection-correction data. (2) Process-level RL stage, a multi-dimensional reward function further enhances answer quality, reflection accuracy, and correction effectiveness.

using a multi-dimensional reward to promote reflection accuracy, answer quality, and correction gain.

Behavioral Pattern Learning

Behavior Data Construction. To train a citation generation model F capable of reflecting on and correcting its own citation errors, we first construct structured training data comprising three components: **attempt**, **reflection**, and **correction**. The *attempt* is designed to simulate the model's initial citation behavior. Specifically, given a question q and a set of retrieved passages $P = \{p_1, \ldots, p_m\}$, we prompt the model F to generate an answer with citations y_c, and use it as the initial attempt.

For *reflection* construction, we move beyond relying on coarse-grained reflections directly generated by LLMs, which often lacks precision, fails to localize erroneous citations, and cannot differentiate between error types, making it unsuitable for supervised correction training. Instead, we adopt a fine-grained and controllable reflection construction strategy, grounded in explicit error typing.

To ensure the reliability of reflection labels, we adopt a precision-oriented labeling pipeline. For **mismatch** errors, we employ a FCM (Zha et al. 2023) to evaluate the alignment between the claim and corresponding attributed passages. Specifically, for each sentence s_i and its attributed passage c_{ij}, we computes:

$$o_{ij} = \phi(s_i, c_{ij}), \quad o_{ij} \in [0, 1]$$

where ϕ denotes the FCM and o_{ij} is the predicted consistency score. A predefined threshold is applied to determine whether a claim (the sentence) is inconsistent with its source (the attributed passage) (i.e., mismatch).

To detect **irrelevant** citations those unrelated to the input question q despite support the claim, we apply a reranker $\gamma(s_i, q) \in \{0, 1\}$, which makes a binary judgment on whether each citation $c_{ij} \in C_i$ of the sentence s_i is relevant to the question. Each citation is categorized as follows:

- **Mismatch**: The consistency score is below the threshold;
- **Irrelevance**: Not mismatch and the reranker predicts irrelevance;
- **Correct**: otherwise.

These error type labels are aggregated into structured, sentence-level reflection annotations. Compared with coarse-grained LLM-generated feedback, this method provides **fine-grained and controllable supervision**, enabling more effective learning of reflection behavior.

For the *correction*, we adopt in-context learning approach, using a strong instruction-aligned LLM (e.g., GPT-4o) to revise the initial attempt based on its corresponding reflection. To ensure quality, we apply dual filtering based on two metrics: (1) a citation quality score $Q_{\text{cite}}(correction) = \text{CitationF1}(correction, C, \phi)$, computed based using an FCM, and corresponds to the harmonic mean of citation precision and recall. (2) An answer quality evaluation score, $Q_{\text{ans}}(correction) = \text{Correctness}(correction, A)$ measures the degree of alignment between the generated answer and the reference answer A (e.g., using Exact Match score). We design an acceptance function to determine whether a correction instance should be accepted.

$$\text{Accept}(\hat{y}^c, \hat{y}^a) = \mathbb{I}\left[\begin{array}{l} Q_{\text{cite}}(\hat{y}^c) \geq \tau_{\text{cite}} \wedge Q_{\text{ans}}(\hat{y}^c) \geq \tau_{\text{ans}} \\ \wedge \left(Q_{\text{cite}}(\hat{y}^c) > Q_{\text{cite}}(\hat{y}^a)\right. \\ \left. \wedge\, Q_{\text{ans}}(\hat{y}^c) > Q_{\text{ans}}(\hat{y}^a)\right) \end{array}\right] \tag{1}$$

where $\mathbb{I}$ is the indicator function. Only when $\text{Accept}(\hat{y}^c, \hat{y}^a) = 1$, correction $\hat{y}^c$ can be retained.

Initial Training. After data construction, we combine the attempt, reflection and correction into a complete *attempt–reflection–correction* chain y and use them to perform supervised fine-tuning on model F. This stage enables the model to learn the behavior of explicitly identifying and correcting citation errors based on structured reflection signals.

Joint optimization over the attempt–reflect–correct chain promotes internal consistency between reflection and correction behaviors. Specifically, at the initial stage, we warm up the model by minimizing the negative log-likelihood loss:

$$\mathcal{L}_{NLL} = -\frac{1}{|y|}\sum_{t=1}^{|y|} \log p_\theta(y_t|q, P, y_{<t}) \tag{2}$$

Online Self-Reflective Bootstrapping. In order to enhance generalization in complex citation scenarios, we propose an *online elf-reflective bootstrapping* strategy. After the initial warm up, the model iteratively generates *attempt–reflection–correction* chains. We retain those chains for which Accept$(y^c, y^a) = 1$, and incorporate them into the training set for the subsequent training epochs. To maintain high-quality reflection supervision, we replace the model-generated reflections with more accurate ones, automatically constructed using the same FCM and reranker model from the data construction pipeline. This self-sampling and training process enables the model to continually learn from self-generated samples where citation errors can be accurately reflected and corrected.

Process-Level RL with Multi-Dimensional Rewards

After the first-stage learning, the model has acquired a preliminary ability to generate with citations and perform self-reflective correction. To further strengthen this capability, we adopt reinforcement learning (RL) to optimize the model behavior across the entire *attempt–reflection–correction* process. Inspired by process-level learning (Shao et al. 2024), we treat this generation process as a composition of interdependent sub-behaviors and design a multi-dimensional reward function to supervise each sub-behavior.

Our reward design consists of three components. First, we define a **reflection accuracy reward** to encourage correct classification of citation error types. For N citations, let $\hat{y}_i$ denote the predicted error type for citation i and y_i represents the ground-truth (automatically derived as reflection construction). The reward is computed as:

$$R_r^{\text{refl}}(r|x, y_{:r}) = \frac{1}{N}\sum_{i=1}^{N}(\mathbb{I}[\hat{y}_i = y_i] - \mathbb{I}[\hat{y}_i \neq y_i]) \tag{3}$$

Second, we assign a binary **citation and answer quality reward** to both the attempt and correction behaviors:

$$R_c^{\text{ans}}(c|x, y_{:c}) = \begin{cases} +1, & \text{if } Q_{\text{cite}}(y^{corr}) \geq \tau_{cite} \\ & \wedge Q_{\text{ans}}(y^{corr}) \geq \tau_{ans} \\ -1, & otherwise \end{cases}$$
$$R_a^{\text{ans}}(a|x, y_{:a}) = \begin{cases} +1, & \text{if } Q_{\text{cite}}(y^{attempt}) \geq \tau_{cite}^{attempt} \\ & \wedge Q_{\text{ans}}(y^{attempt}) \geq \tau_{ans}^{attempt} \\ -1, & otherwise \end{cases} \tag{4}$$

where $\tau_{cite}^{(\cdot)}$ and $\tau_{ans}^{(\cdot)}$ denote predefined thresholds.

Finally, to encourage meaningful correction beyond the initial attempt, we introduce a **correction gain reward**, computed based on the improvement in generation quality:

$$R_c^{\text{gain}}(c|x, y_{:c}) = \begin{cases} +1, & \text{if } Q_{\text{cite}}(y^{corr}) \geq Q_{\text{cite}}(y^{attempt}) \\ & \wedge Q_{\text{ans}}(y^{corr}) \geq Q_{\text{ans}}(y^{attempt}) \\ -1, & otherwise \end{cases} \tag{5}$$

This reward formulation enables fine-grained credit assignment across the entire citation generation process and effectively guides the model toward more accurate and self-correctable behaviors.

Drawing inspiration from the process-level GRPO paradigm (Shao et al. 2024), all behaviors of the same type (i.e., attempt, reflection, or correction) belong to a group. The unified reward context of the behavior in a group is defined as follows:

$$\mathbf{R}(z_t) = (R_{z_i}(z_i \mid x, y_{:z_i}))_{i=1}^{t-1} \tag{6}$$

where $y_{:z_i}$ denotes the sequence of behaviors preceding the behavior z_i. Following REINFORCE with Leave-One-Out (RLOO) (Ahmadian et al. 2024), we compute the advantage for each behavior in a group:

$$\begin{aligned} A_{z_t}(x, y) =& R_{z_t}(x, y_{:z_t}) - b_{z_t}(x, y) \\ & - \beta log\frac{\pi_{\theta_{old}}(z_t|x, y)}{\pi_{ref}(z_t|x, y)} \end{aligned} \tag{7}$$

where the $b_{z_t}(x, y)$ is baseline, which can be computed as follows:

$$b_{z_t}(x, y) = \frac{1}{|G(\mathbf{R}(z_t))|}\sum_{z \in G(\mathbf{R}(z_t))} R_z(z_t|x^{(z)}, y_{:z_t}^{(z)}) \tag{8}$$

where $G(\cdot)$ denotes the group. This baseline estimation reduces the variance of the policy gradient and improves training stability. The final policy gradient loss is computed as:

$$\begin{aligned} \mathcal{L}_{RL} =& -E_{x\sim D, y\sim\pi_{\theta_{old}}(\cdot|x)}[\frac{1}{|y|_z}\sum_{z\in y} min(r_z(\theta) \\ & A(z|x, y_{:z}), clip(r_z(\theta), 1-\epsilon, 1+\epsilon)A(z|x, y_{:z}))] \end{aligned} \tag{9}$$

where $r_z(\theta) = \frac{\pi_\theta(z|x, y_{:z})}{\pi_{\theta_{old}}(z|x, y_{:z})}$ is the importance ratio, θ is the parameters of the model F.

Experiments

Datasets & Metrics

We conduct experiments on the ALCE citation benchmark (Gao et al. 2023) for long-form question answering, focusing primarily on the ASQA and ELI5 datasets. In ALCE, the number of reference documents provided for training is $\mathcal{D} = 4$. We adopt the evaluation protocol from (Gao et al. 2023), which includes correctness (measured by EM Recall), fluency (measured by MAUVE), and citation F1 (assessed using an NLI-trained T5-11B model). In addition, we report ROUGE-L scores and the passage-grounded correctness (Correct. in P) metric introduced in (Aly et al. 2024), which evaluates whether the response is supported by the retrieved documents P, disregarding factual content potentially memorized by the language model.

Method	ALCE-ASQA					ALCE-ELI5				
	Similarity Rouge-L	Fluency MAUVE	Correct EM Rec.	Correct . in P	Citation F_1	Similarity Rouge-L	Fluency MAUVE	Correct EM Rec.	Correct . in P	Citation F_1
ChatGPT	–	66.6	40.4	–	73.1	–	57.2	12.0	–	50.5
GPT-4	–	67.1	41.3	–	71.9	–	38.4	14.2	–	46.9
AGREE	–	–	40.9	–	75.1	–	–	-	–	-
Self-RAG 7B	35.7	74.3	30.0	–	67.3	16.9	32.6	9.7	5.4	27.6
BP, T5-3B	–	–	33.8	–	77.8	–	–	5.2	-	60.9
LLaMA2-7B-chat										
In-context	$35.9_{0.3}$	$77.8_{3.1}$	$35.0_{0.6}$	$25.7_{0.6}$	$49.9_{1.0}$	$20.5_{0.2}$	$36.2_{2.5}$	$17.7_{0.6}$	$10.8_{0.6}$	$38.2_{0.6}$
Few-shot FT	$34.9_{0.4}$	$69.2_{4.3}$	$32.0_{0.4}$	$22.3_{0.7}$	$55.0_{1.8}$	$\underline{21.3}_{0.2}$	$58.2_{2.2}$	$\underline{17.8}_{0.6}$	$11.2_{1.1}$	$48.7_{2.9}$
CaLF	$\underline{37.8}_{0.4}$	$\mathbf{86.0}_{3.7}$	$\underline{37.7}_{0.6}$	$\underline{29.3}_{0.4}$	$\underline{70.4}_{2.5}$	$20.8_{1.0}$	$\underline{59.6}_{11.5}$	$17.0_{0.3}$	$\underline{11.9}_{0.2}$	$\underline{66.5}_{5.9}$
Ours	$\mathbf{38.5}_{0.3}$	$\underline{85.2}_{2.9}$	$\mathbf{40.3}_{0.5}$	$\mathbf{32.1}_{0.8}$	$\mathbf{74.9}_{1.9}$	$\mathbf{21.5}_{0.2}$	$\mathbf{60.1}_{4.4}$	$\mathbf{19.2}_{0.7}$	$\mathbf{13.5}_{0.3}$	$\mathbf{68.4}_{2.7}$
Mistral-Orca-7B										
In-context	$38.7_{0.1}$	$54.7_{1.8}$	$40.2_{0.3}$	$31.9_{0.2}$	$55.6_{0.8}$	$\underline{20.9}_{0.1}$	$29.3_{0.8}$	$\underline{20.8}_{0.4}$	$12.5_{0.5}$	$43.3_{0.5}$
Few-shot FT	$38.4_{1.8}$	$78.6_{14.7}$	$38.4_{3.8}$	$29.9_{4.7}$	$62.6_{3.6}$	$19.4_{1.9}$	$60.5_{13.6}$	$17.3_{1.8}$	$10.9_{1.2}$	$57.7_{6.5}$
CaLF	$\underline{40.3}_{0.2}$	$\underline{84.0}_{3.3}$	$\underline{41.7}_{1.2}$	$\underline{34.5}_{0.5}$	$\underline{81.5}_{2.5}$	$20.4_{1.5}$	$\underline{62.7}_{4.6}$	$18.4_{2.1}$	$\underline{13.1}_{0.7}$	$\underline{73.1}_{4.2}$
Ours	$\mathbf{41.0}_{0.3}$	$\mathbf{84.1}_{1.1}$	$\mathbf{43.1}_{1.5}$	$\mathbf{36.5}_{0.3}$	$\mathbf{84.9}_{0.8}$	$\mathbf{22.0}_{0.6}$	$\mathbf{63.5}_{2.9}$	$\mathbf{21.4}_{1.1}$	$\mathbf{15.7}_{0.6}$	$\mathbf{76.3}_{2.1}$

Table 1: Main results on ALCE benchmark (ASQA and ELI5). We report the mean and standard deviation measured across 3 different random seeds. For each dataset, we use Citation F1 to measure citation quality, EM Recall to assess overall correctness, Correct. in P to evaluate the proportion of correct information grounded in the retrieved passages, Rouge-L to assess textual similarity, and MAUVE to evaluate fluency. Bold indicates the best performance, while underline denotes the second-best.

Experimental Setup

To better reflects realistic deployment scenarios, we follow few-shot learning recommendations (Alex et al. 2021) and omit a separate validation set for hyperparameter tuning. Due to computational constraints, we adopt LoRA (Hu et al. 2022) for parameter-efficient fine-tuning. We use AlignScore as our FCM. The threshold τ_{cite} and τ_{ans} is set to 0.8 and 0.45. Notably, AlignScore differs from the citation evaluation model used in our final evaluation in both architecture and training data. In terms of knowledge sources, ASQA uses Wikipedia as its knowledge source, while ELI5 relies on CommonCrawl. For fair comparison, we use the same retriever as employed by baseline models. In the behavioral pattern learning stage, we perform 1 epoch of warm-up followed by 3 epochs of online self-reflective bootstrapping with a learning rate of 1e-4. In the RL phase, we set $\tau_{\text{cite}}^{\text{attempt}}$ and $\tau_{\text{ans}}^{\text{attempt}}$ to 0.7 and 0.4. The learning rate is set to 5e-6, batch size to 16, KL coefficient to 0.1, and train for 3 epochs.

Baselines

We compare our method against a range of strong baselines. First, we include in-context prompting and few-shot fine-tuning approaches based on the same instruction-tuned LLMs, trained on the few-shot citation dataset D. We further compare with the state-of-the-art model CaLF (Aly et al. 2024), which iteratively trains on filtered, self-generated data. Additionally, we evaluate against powerful in-context prompting baselines from (Gao et al. 2023), including ChatGPT (gpt-3.5-turbo-0301) and GPT-4 (gpt-4-0613), which use large-scale parameters.In-context prompting uses two randomly sampled demonstrations. We also consider several recent citation-aware models: AGREE (Ye et al. 2023), based on PaLM 2; Self-RAG 7B (Asai et al. 2024), built on the open-source Llama2 backbone; and Blueprint (BP) (Fierro et al. 2024), which is based on T5-3B.

Main Results

Table 1 presents the results of the in-domain experiments. Overall, FineRef consistently achieves the highest or highly competitive performance across all evaluation metrics compared to baseline methods. Our approach significantly outperforms larger-scale models such as ChatGPT and GPT-4 on all metrics. Across different base LLMs (LLaMA2-7B-chat and Mistral-Orca-7B), FineRef substantially surpasses both in-context and few-shot fine-tuning approaches, with up to +22.3 improvement in Citation F1 over the few-shot FT baseline on ASQA. Compared to the state-of-the-art model, CaLF, FineRef demonstrates notable improvements in both citation and correction performance. While CaLF achieves better citation quality than GPT-4, its improvement in correction is marginal (e.g., only +0.4 EM on ASQA). In contrast, FineRef yields consistent gains over GPT-4 in both citation and QA quality, with improvements of +2.8 EM on ASQA and +7.2 EM on ELI5. These results highlight the effectiveness of our fine-grained reflection mechanism in jointly improving citation accuracy and answer quality by explicitly modeling different types of citation errors.

Domain Transfer

Table 2 presents the results of our zero-shot transfer experiments, where models trained on one domain are directly evaluated on a different target domain without further

Method ($Source \rightarrow Target$)	Similarity Rouge-L	Fluency MAUVE	Correct EM Rec.	Correct . in P	Citation F1	**Method** ($Source \rightarrow Target$)	Similarity Rouge-L	Fluency MAUVE	Correct EM Rec.	Correct . in P	Citation F1
Self-RAG 7B	35.7	74.3	30.0	-	67.3	Self-RAG 7B	16.9	32.6	9.7	5.4	27.6
Zero-Shot($\rightarrow A$)	39.0	78.9	39.5	31.6	5.7	Zero-Shot($\rightarrow E$)	**21.3**	35.0	22.2	12.6	10.4
Few-shot FT($E \rightarrow A$)	39.7	**90.1**	38.5	31.4	71.7	Few-shot FT ($A \rightarrow E$)	20.9	**41.1**	19.7	10.6	40.4
Calf($E \rightarrow A$)	40.1	86.6	40.0	33.2	79.5	Calf($A \rightarrow E$)	21.2	31.3	20.4	12.5	57.3
Ours($E \rightarrow A$)	**40.5**	87.2	**42.3**	**34.7**	**81.3**	Ours($A \rightarrow E$)	21.0	32.4	**23.6**	**13.4**	**59.1**

Table 2: Zero-shot domain transfer evaluation results on ASQA (A) and ELI5 (E). We report the results using Mistral-Orca-7B.

fine-tuning.across all source-to-target settings, our method (based on MistralOrca-7B) consistently outperforms Self-RAG, Zero-Shot, Few-shot Fine-tuning, and the state-of-the-art method Calf in both citation quality and correction performance. On the ASQA target domain, our approach (trained on ELI5) achieves notable improvements in EM Recall (+2.3), Correct in Passage (+1.5), and Citation F1 (+1.8) compared to the previous state-of-the-art, Calf. Similarly, on the ELI5 target domain, our model (trained on ASQA) maintains superior performance in EM Recall (23.6) and Citation F1 (59.1), outperforming Calf by margins of +3.2 and +1.8 respectively. This demonstrates that our approach, which explicitly trains the model to reflect on different types of citation errors, leads to better generalization across domains in both citation and answer quality.

Method	Correct EM Rec.	Correct . in P	Citation F1
Ours	43.4	36.3	85.2
wo reflection	$38.2_{(-5.2)}$	$32.6_{(-3.7)}$	$79.4_{(-5.8)}$
wo IR error type	$39.9_{(-3.5)}$	$33.3_{(-3.0)}$	$82.1_{(-3.1)}$
coarse-grained refl.	$40.2_{(-3.2)}$	$33.8_{(-2.5)}$	$81.1_{(-3.1)}$
wo RL	$40.8_{(-2.6)}$	$33.9_{(-2.4)}$	$80.6_{(-4.6)}$
wo gain reward	$42.0_{(-1.4)}$	$32.9_{(-3.4)}$	$81.6_{(-3.6)}$
wo online boost.	$41.1_{(-2.3)}$	$34.2_{(-2.1)}$	$81.5_{(-3.7)}$

Table 3: Results for ablation study. We report the results on ASQA dataset using Mistral-Orca-7B.

Analysis

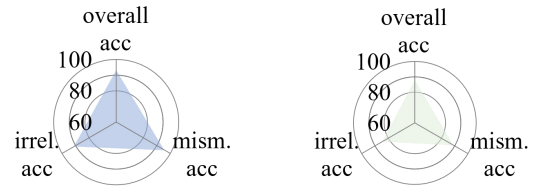
Ablation. To better understand the underlying mechanisms of our method, we conduct a series of ablation studies to examine the contribution of each component. Table 3 presents the results on the ASQA dataset using the Mistral-Orca-7B model. We first analyze the role of the **reflection** module. Removing the reflection behavior results in a substantial drop in both EM Recall (-5.2) and Citation F1 (-5.8), indicating that the reflection step plays a vital role in improving both citation accuracy and answer quality.

We further investigate the impact of excluding the identification of **irrelevant** citations during reflection. This leads to notable declines in EM Recall (-3.5), highlighting the importance of modeling irrelevant citation errors to maintain question-answering capability. Additionally, we compare our fine-grained reflection strategy to a **coarse-grained alternative** that uses GPT-4 to assess the overall correctness of the attempt. This approach yields inferior performance across all quality metrics, confirming the necessity of our fine-grained reflection formulation.

We also evaluate the effect of different components in the training strategy. Removing the **RL** stage leads to notable performance drops (-2.6 EM and -3.6 Citation F1), highlighting the importance of self-exploration in strengthening reflective and corrective behaviors. The **gain reward** also contributes positively by preventing behavioral collapse and encouraging higher-quality correction generations. Finally, ablating the **online self-reflective bootstrapping (online boost.)** leads to declines in all metrics, as it weakens the model's initial behavior pattern learned during SFT, thus limiting the improvement achievable during the RL phase.

Method	Correct EM Rec.	Correct . in P	Citation F1
First Attmpt	38.6	30.3	77.4
Correction	43.1	36.5	84.7
Δ	+4.5	+6.2	+7.3

Table 4: Performance comparison of first attempt and correction. We report the results on ASQA using Mistral-Orca-7B.

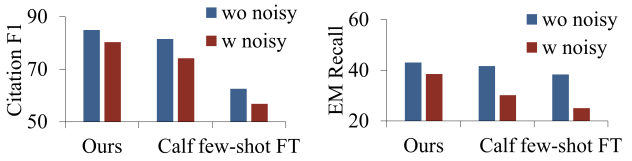


(a) FCM+Reranker Reflection (b) Self-Reflection

Figure 3: Accuracy of self-reflection

Correction Improvement. To assess the effectiveness of the correction behavior in enhancing initial generation quality, Table 4 presents the performance improvements of the model after applying correction relative to the first attempt. Notably, all evaluation metrics show consistent gains (Citation F1 +7.3, EM Recall +4.5 and Correct in P +6.2). These results highlight that the model's initial attempt often falls short in citation and answer quality, whereas the subsequent correction—guided by fine-grained self-reflection—substantially enhances the reliability and accuracy of the generated response.

Reflection Accuracy Analysis. To evaluate the accuracy of reflection, we conduct a two-step analysis. First, we need to ensure that the FCM and reranker produce reliable reflection signals. We sample a subset of 400 test samples from ASQA to evaluate the accuracy of reflection labels constructed by FCM and reranker. Two human evaluators are tasked with verifying whether the predicted error types (e.g., mismatch or irrelevance) are accurate. As shown in Figure 3(a), the reflection signals based on FCM and reranker exhibit high accuracy (over 90%) for reflecting on different types of errors. Next, we use the FCM and reranker as gold reflection signals, and evaluate the accuracy of the model's self-generated reflections. The evaluation across all test samples of ASQA is as illustrated in Figure 3(b). This measures the model's ability to correctly identify citation errors and improve its own performance through self-reflection.



(a) Citation F1 (b) EM Recall

Figure 4: Performance in the scenario with noisy passages

Robustness under Noisy Retrieval. To evaluate the robustness of our method under realistic retrieval noise conditions, we introduce semantically fluent yet contextually irrelevant distractor passages, simulating common failure modes of open-domain retrievers. Concretely, for each input question, we randomly replace one of the retrieved passages with such a distractor. This setup emulates real-world retrieval systems that often return noisy or irrelevant content.

We compare FineRef against the state-of-the-art CaLF method and a Few-shot fine-tuning baseline. As shown in Figure 4, the performance of both CaLF and Few-shot fine-tuning degrades substantially in the presence of distractor passages, particularly EM Recall. In contrast, FineRef consistently maintains a significant lead in both Citation F1 and EM Recall. These results validate the effectiveness of our approach in enhancing model robustness through explicit citation error modeling and fine-grained reflection.

Related Work

Text Generation with Citations

Generating text with citations enhances LLM verifiability and mitigates hallucinations (Gao et al. 2023; Li et al. 2023). Early practices integrated LLMs with commercial search engines[1][2] , complicating citation evaluation and prompting the development of benchmarks like ALCE (Gao et al. 2023). Most work adopts a "single-pass generation" paradigm, allowing LLMs to directly produce cited answers from retrieved documents. To improve this process, methods such as in-context learning with high-quality exemplars (Gao et al. 2023; Liu, Zhang, and Liang 2023), post-hoc citation editing (Bohnet et al. 2022), and Reinforcement Learning from Human Feedback (RLHF) (Menick et al. 2022; Thoppilan et al. 2022; Nakano et al. 2021) have been proposed. Weakly-supervised strategies like CaLF (Aly et al. 2024) further address data scarcity by using Factual Consistency Models (FCMs) to filter diverse candidate answers for training. Despite this progress, key limitations persist: current methods prioritize citation *fidelity* over *relevance* compromises QA performance and robustness under noisy real-world conditions, and hinders the model's ability to generate optimal responses in long-form tasks requiring multiple citations. FineRef introduces a fine-grained error reflection mechanism that explicitly trains the model to identify and correct both "mismatch" and "irrelevance" citation errors, enhancing its overall performance and robustness.

LLM Reasoning and Self-Reflection

Self-reflection is a critical mechanism for improving complex reasoning abilities and reliability of LLMs (Madaan et al. 2023; Paul et al. 2023), enabling models to evaluate and iteratively refine their responses in a human-like manner. Recent studies have explored various approaches to enhance these capabilities during post-training (Saunders et al. 2022; Rosset et al. 2024; Kumar et al. 2024). Enabling LLMs to perform effective self-verification and self-correction is a promising solution for achieving robust reasoning, as direct prompting for such behaviors is often suboptimal (Huang et al. 2023; Yixing et al. 2024; Zhang et al. 2024). RL has also proven effective in enhancing LLM reasoning (Setlur et al. 2024; Ouyang et al. 2022), with research focusing on actor-critic frameworks (Havrilla et al. 2024; Tajwar et al. 2024) and the design of accurate reward models to guide the learning process (Lightman et al. 2023). Some work (Ma et al. 2025), has combined these ideas, using RL to teach models how to self-verify and self-correct. Existing self-reflection mechanisms are often too coarse-grained and rely on unreliable self-generated feedback, limiting their effectiveness. FineRef introduces a fine-grained, controllable approach that uses external specialized and lightweight models to precisely reflect errors at the citation level.

Conclusion

In this paper, we propose FineRef, a noval training framework that enables LLMs to perform fine-grained self-reflection for identifying and correcting both mismatch and irrelevance citation errors. FineRef explicitly models the full attempt–reflect–correct process through supervised learning with fine-grained and controllable reflection signals and enhances this capability via process-level RL with multi-dimensional rewards design. Extensive experiments demonstrate that FineRef not only substantially improves citation fidelity and answer correctness but also exhibits strong robustness across domains and under retrieval noise. These results highlight the effectiveness of incorporating fine-grained self-reflection into training, representing a promising step toward building more reliable, interpretable, and trustworthy language models.

[1]https://www.perplexity.ai/

[2]https://www.bing.com/new

Acknowledgments

This research is supported by Artificial Intelligence-National Science and Technology Major Project 2023ZD0121200.

References

Achiam, J.; Adler, S.; Agarwal, S.; Ahmad, L.; Akkaya, I.; Aleman, F. L.; Almeida, D.; Altenschmidt, J.; Altman, S.; Anadkat, S.; et al. 2023. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*.

Ahmadian, A.; Cremer, C.; Gallé, M.; Fadaee, M.; Kreutzer, J.; Pietquin, O.; Üstün, A.; and Hooker, S. 2024. Back to basics: Revisiting reinforce style optimization for learning from human feedback in llms. *arXiv preprint arXiv:2402.14740*.

Alex, N.; Lifland, E.; Tunstall, L.; Thakur, A.; Maham, P.; Riedel, C. J.; Hine, E.; Ashurst, C.; Sedille, P.; Carlier, A.; et al. 2021. RAFT: A real-world few-shot text classification benchmark. *arXiv preprint arXiv:2109.14076*.

Aly, R.; Tang, Z.; Tan, S.; and Karypis, G. 2024. Learning to generate answers with citations via factual consistency models. *arXiv preprint arXiv:2406.13124*.

Asai, A.; Wu, Z.; Wang, Y.; Sil, A.; and Hajishirzi, H. 2024. Self-rag: Learning to retrieve, generate, and critique through self-reflection.

Bohnet, B.; Tran, V. Q.; Verga, P.; Aharoni, R.; Andor, D.; Soares, L. B.; Ciaramita, M.; Eisenstein, J.; Ganchev, K.; Herzig, J.; et al. 2022. Attributed question answering: Evaluation and modeling for attributed large language models. *arXiv preprint arXiv:2212.08037*.

Borgeaud, S.; Mensch, A.; Hoffmann, J.; Cai, T.; Rutherford, E.; Millican, K.; Van Den Driessche, G. B.; Lespiau, J.-B.; Damoc, B.; Clark, A.; et al. 2022. Improving language models by retrieving from trillions of tokens. In *International conference on machine learning*, 2206–2240. PMLR.

Brown, T.; Mann, B.; Ryder, N.; Subbiah, M.; Kaplan, J. D.; Dhariwal, P.; Neelakantan, A.; Shyam, P.; Sastry, G.; Askell, A.; et al. 2020. Language models are few-shot learners. *Advances in neural information processing systems*, 33: 1877–1901.

Fierro, C.; Amplayo, R. K.; Huot, F.; De Cao, N.; Maynez, J.; Narayan, S.; and Lapata, M. 2024. Learning to plan and generate text with citations. *arXiv preprint arXiv:2404.03381*.

Gao, T.; Yen, H.; Yu, J.; and Chen, D. 2023. Enabling large language models to generate text with citations. *arXiv preprint arXiv:2305.14627*.

Havrilla, A.; Du, Y.; Raparthy, S. C.; Nalmpantis, C.; Dwivedi-Yu, J.; Zhuravinskyi, M.; Hambro, E.; Sukhbaatar, S.; and Raileanu, R. 2024. Teaching large language models to reason with reinforcement learning. *arXiv preprint arXiv:2403.04642*.

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.

Huang, J.; Chen, X.; Mishra, S.; Zheng, H. S.; Yu, A. W.; Song, X.; and Zhou, D. 2023. Large language models cannot self-correct reasoning yet. *arXiv preprint arXiv:2310.01798*.

Izacard, G.; Lewis, P.; Lomeli, M.; Hosseini, L.; Petroni, F.; Schick, T.; Dwivedi-Yu, J.; Joulin, A.; Riedel, S.; and Grave, E. 2023. Atlas: Few-shot learning with retrieval augmented language models. *Journal of Machine Learning Research*, 24(251): 1–43.

Ji, B.; Liu, H.; Du, M.; and Ng, S.-K. 2024. Chain-of-thought improves text generation with citations in large language models. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, 18345–18353.

Kryściński, W.; McCann, B.; Xiong, C.; and Socher, R. 2020. Evaluating the Factual Consistency of Abstractive Text Summarization. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 9332–9346.

Kumar, A.; Zhuang, V.; Agarwal, R.; Su, Y.; Co-Reyes, J. D.; Singh, A.; Baumli, K.; Iqbal, S.; Bishop, C.; Roelofs, R.; et al. 2024. Training language models to self-correct via reinforcement learning. *arXiv preprint arXiv:2409.12917*.

Li, J.; Cheng, X.; Zhao, W. X.; Nie, J.-Y.; and Wen, J.-R. 2023. Helma: A large-scale hallucination evaluation benchmark for large language models. *arXiv preprint arXiv:2305.11747*.

Lightman, H.; Kosaraju, V.; Burda, Y.; Edwards, H.; Baker, B.; Lee, T.; Leike, J.; Schulman, J.; Sutskever, I.; and Cobbe, K. 2023. Let's verify step by step. In *The Twelfth International Conference on Learning Representations*.

Liu, N. F.; Zhang, T.; and Liang, P. 2023. Evaluating verifiability in generative search engines. *arXiv preprint arXiv:2304.09848*.

Ma, R.; Wang, P.; Liu, C.; Liu, X.; Chen, J.; Zhang, B.; Zhou, X.; Du, N.; and Li, J. 2025. S^2R: Teaching LLMs to Self-verify and Self-correct via Reinforcement Learning. *arXiv preprint arXiv:2502.12853*.

Madaan, A.; Tandon, N.; Gupta, P.; Hallinan, S.; Gao, L.; Wiegreffe, S.; Alon, U.; Dziri, N.; Prabhumoye, S.; Yang, Y.; et al. 2023. Self-refine: Iterative refinement with self-feedback. *Advances in Neural Information Processing Systems*, 36: 46534–46594.

Menick, J.; Trebacz, M.; Mikulik, V.; Aslanides, J.; Song, F.; Chadwick, M.; Glaese, M.; Young, S.; Campbell-Gillingham, L.; Irving, G.; et al. 2022. Teaching language models to support answers with verified quotes. *arXiv preprint arXiv:2203.11147*.

Nakano, R.; Hilton, J.; Balaji, S.; Wu, J.; Ouyang, L.; Kim, C.; Hesse, C.; Jain, S.; Kosaraju, V.; Saunders, W.; et al. 2021. Webgpt: Browser-assisted question-answering with human feedback. *arXiv preprint arXiv:2112.09332*.

Ouyang, L.; Wu, J.; Jiang, X.; Almeida, D.; Wainwright, C.; Mishkin, P.; Zhang, C.; Agarwal, S.; Slama, K.; Ray, A.; et al. 2022. Training language models to follow instructions with human feedback. *Advances in neural information processing systems*, 35: 27730–27744.

Paul, D.; Ismayilzada, M.; Peyrard, M.; Borges, B.; Bosselut, A.; West, R.; and Faltings, B. 2023. Refiner: Reasoning feedback on intermediate representations. *arXiv preprint arXiv:2304.01904*.

Rosset, C.; Cheng, C.-A.; Mitra, A.; Santacroce, M.; Awadallah, A.; and Xie, T. 2024. Direct nash optimization: Teaching language models to self-improve with general preferences. *arXiv preprint arXiv:2404.03715*.

Saunders, W.; Yeh, C.; Wu, J.; Bills, S.; Ouyang, L.; Ward, J.; and Leike, J. 2022. Self-critiquing models for assisting human evaluators. *arXiv preprint arXiv:2206.05802*.

Setlur, A.; Garg, S.; Geng, X.; Garg, N.; Smith, V.; and Kumar, A. 2024. Rl on incorrect synthetic data scales the efficiency of llm math reasoning by eight-fold. *Advances in Neural Information Processing Systems*, 37: 43000–43031.

Shao, Z.; Wang, P.; Zhu, Q.; Xu, R.; Song, J.; Bi, X.; Zhang, H.; Zhang, M.; Li, Y.; Wu, Y.; et al. 2024. Deepseekmath: Pushing the limits of mathematical reasoning in open language models. *arXiv preprint arXiv:2402.03300*.

Shinn, N.; Cassano, F.; Gopinath, A.; Narasimhan, K.; and Yao, S. 2023. Reflexion: Language agents with verbal reinforcement learning. *Advances in Neural Information Processing Systems*, 36: 8634–8652.

Tajwar, F.; Singh, A.; Sharma, A.; Rafailov, R.; Schneider, J.; Xie, T.; Ermon, S.; Finn, C.; and Kumar, A. 2024. Preference fine-tuning of llms should leverage suboptimal, on-policy data. *arXiv preprint arXiv:2404.14367*.

Thoppilan, R.; De Freitas, D.; Hall, J.; Shazeer, N.; Kulshreshtha, A.; Cheng, H.-T.; Jin, A.; Bos, T.; Baker, L.; Du, Y.; et al. 2022. Lamda: Language models for dialog applications. *arXiv preprint arXiv:2201.08239*.

Xiao, S.; Liu, Z.; Zhang, P.; and Muennighoff, N. 2023. C-Pack: Packaged Resources To Advance General Chinese Embedding. arXiv:2309.07597.

Ye, X.; Sun, R.; Arik, S. Ö.; and Pfister, T. 2023. Effective large language model adaptation for improved grounding and citation generation. *arXiv preprint arXiv:2311.09533*.

Yixing, P.; Wang, Q.; Zhang, L.; Liu, Y.; and Mao, Z. 2024. Chain-of-question: A progressive question decomposition approach for complex knowledge base question answering. In *Findings of the Association for Computational Linguistics ACL 2024*, 4763–4776.

Yue, X.; Wang, B.; Chen, Z.; Zhang, K.; Su, Y.; and Sun, H. 2023. Automatic evaluation of attribution by large language models. *arXiv preprint arXiv:2305.06311*.

Zha, Y.; Yang, Y.; Li, R.; and Hu, Z. 2023. AlignScore: Evaluating factual consistency with a unified alignment function. *arXiv preprint arXiv:2305.16739*.

Zhang, Q.; Qiu, H.; Wang, D.; Qian, H.; Li, Y.; Zhang, T.; and Huang, M. 2024. Understanding the Dark Side of LLMs' Intrinsic Self-Correction. *arXiv preprint arXiv:2412.14959*.