Ask, Assess, and Refine: Rectifying Factual Consistency and Hallucination in LLMs with Metric-Guided Feedback Learning

Dongyub Lee*
Naver Corp.
dongyub.lee@navercorp.com

Eunhwan Park*
Jeonbuk National University
nlp.judepark@gmail.com

Hodong Lee
Korea University
bigshane319@gmail.com

Heuiseok Lim†
Korea University
limhseok@korea.ac.kr

Abstract

Recent advancements in Large Language Models (LLMs) have heralded unprecedented capabilities in information-seeking and text generation, as evidenced by applications like Bing Chat and perplexity.ai. Despite these strides, challenges on hallucination and factual inconsistency continue to impede their wider real-world adoption. Contemporary methods, including retrieval-augmented LLMs and feedback-based learning, serve as alternatives to mitigate these challenges. However, challenges remain, particularly regarding referencing erroneous evidence (citation errors) and generating information not present in the evidence (hallucination). In this paper, we introduce the A^{2}R framework: **A**sk, **A**ssess, and **R**efine. Our approach utilizes an *explicit* evaluation paradigm, incorporating metrics specifically tailored to assess citation errors and hallucination, aiming to address these prevalent challenges robustly. Capitalizing on these evaluations, we devise a strategy to formulate actionable natural language feedback, enabling iterative refinements that yield improved factual consistency and reduced hallucinations in responses. Our experiments on ASQA, ELI5, and QAMPARI datasets demonstrate our method's superiority in enhancing correctness, fluency, and citation quality.

1 Introduction

Recent pioneering works on Large Language Models (LLMs) have facilitated for information seeking and text generation, thereby showcasing the various real-world applications such as Bing Chat[1] and perplexity.ai[2]. However, despite of significant advancements with a combination of supervised fine-tuning and reinforcement learning, LLMs still tend to generate hallucination and contains the factually incorrect information in their output. Since the lack of factual consistency constrains their extensive use in real-word applications, a substantial research has been dedicated to addressing this deficiency. Additionally, users struggle with challenges verifying the faithfulness of generated responses from LLMs. For example, as illustrated in Figure 1, a user queries the system with, "*Who was the top scorer when Arsenal last won the Premier League?*". In responding, the system first seeks out relevant evidence. Then, it processes both the question and the gathered search results. Finally, it formulates a response, referencing appropriate citations. However, the response has hallucinations and inaccurate citations. Because *Robin Van Persie* was not in during Arsenal's last league-winning 2003–2004 season, and Patrick Vieira was cited from Doc [2] instead of Doc [3].

To address the challenges of hallucination and factual inconsistency in LLM outputs, contempo-

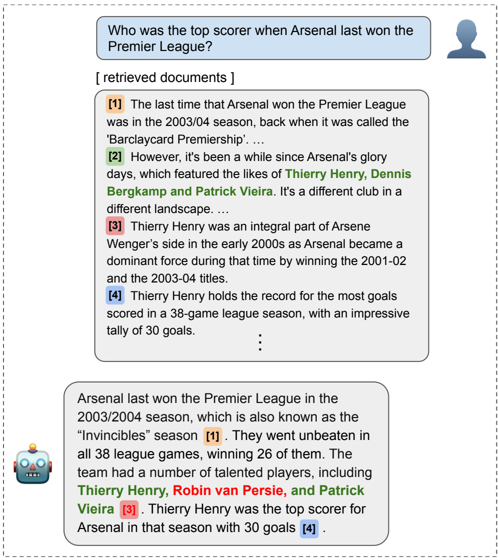

Figure 1: An illustration of error case with citation and hallucination. Robin Van Persie, [3] indicates hallucination and incorrect citations.

* Equal contribution.
† Corresponding author.
[1] https://www.bing.com/chat
[2] https://www.perplexity.ai

Proceedings of the 18th Conference of the European Chapter of the Association for Computational Linguistics
Volume 1: Long Papers, pages 2422–2433
March 17-22, 2024

rary research focuses on two-folds:

1) *Retrieval-augmented LLMs* (Khandelwal et al., 2020; Lewis et al., 2020; Borgeaud et al., 2022; Izacard et al., 2022; Zhong et al., 2022): This approach conditions LLM generation on retrieved evidence from an external knowledge source. A notable contribution in this domain is by (Gao et al., 2023b), which introduces a novel paradigm where LLM outputs are appended with *citations* referencing one or more passages. This framework also emphasizes automatic evaluation across diverse aspects, such as fluency, correctness, and citation quality, facilitating a rigorous assessment of the initial output and directing further enhancements. In practical terms, platforms like Bing Chat and perplexity.ai furnish user queries with responses that cite current web pages, typically sourced from *closed* search engines.

2) *Feedback-based Refinement*: This approach is inspired by how humans refine their written text with feedback, continually improving the content. Reinforcement Learning has been widely adopted to enhance the quality of generation (Böhm et al., 2019; Stiennon et al., 2020; Ziegler et al., 2019; Wu et al., 2020; Ouyang et al., 2022a; Glaese et al., 2022; Akyürek et al., 2023). The work of Roit et al. (2023) employs reinforcement learning and uses a textual entailment-based reward signal for feedback to refine initial outputs. The Self-Refine mechanism (Madaan et al., 2023) introduces an iterative self-refinement algorithm, wherein the model $\mathcal{M}$ *implicitly* evaluates the initial output $\mathcal{Y}$ using LLMs and subsequently refines it.

Despite the advances in LLMs, contemporary research has highlighted several inherent challenges. Primarily, while promising, retrieval-augmented LLMs face issues such as referencing erroneous evidence (citation errors), and generating information not present in the evidence (hallucination), as in Figure 1. To mitigate these issues, the prevalent use of intrinsic self-correction methods involves LLMs to rectify their initial responses based solely on inherent capabilities of LLMs by utilizing feedback-based learning. However, as highlighted by (Huang et al., 2023), employing LLMs' inherent capabilities for self-correction reveals a significant challenge, because LLMs are not adept at self-assessment, struggling to accurately evaluate their output's quality and identify factual inconsistencies. This stems from the LLMs' inadequate self-evaluative capabilities, which are crucial for the identification and rectification of hallucination independently. Consequently, relying solely intrinsic methods could lead to degradation of the LLMs' ability for self-correction.

In a bid to rectify these shortcomings, our paper introduces the framework A^{2}R: **A**sk, **A**ssess, and **R**efine – an approach armed with augmenting LLMs through Metric-based Iterative Feedback Learning. Distinct from Self-Refine (Madaan et al., 2023), which relies on *intrinsic* evaluation of the initial output $\mathcal{Y}$, our study adopt an *explicit* evaluation encompassing multiple aspects, namely correctness, citation quality, as conceptualized in ALCE (Gao et al., 2023b). Leveraging LLMs, we formulate natural language feedback $\mathcal{F}$ for each dimension, predicated upon the respective evaluation outcomes. Using the initial output $\mathcal{Y}$ along with the feedback $\mathcal{F}$, the model $\mathcal{M}$ iteratively refines its output, stopping when it determines that additional refinement is no longer needed. Experimental results on benchmark datasets, including ASQA, ELI5, and QAMPARI, validate the efficacy of our approach, manifesting in substantial enhancements across correctness, fluency, and citation quality metrics. Our findings emphasize the pivotal role of *explicit* evaluations in the feedback generation process, emphasizing their potential in engendering more dependable and articulate outputs from LLMs.

The contributions of our paper can be briefly summarized as follows: (1) We introduce the A^{2}R framework, specifically designed to address the pressing challenges of hallucination and factual inconsistency in LLMs. (2) We pivot from traditional intrinsic evaluations to *explicit* evaluations, enabling granular assessments on aspects like correctness and citation quality. (3) Utilizing the results from explicit evaluations, we propose a method to formulate natural language feedback that is both understandable and actionable. This facilitates iterative refinements to produce responses improved in factual consistency and reduced hallucinations. (4) Our method demonstrates substantial enhancements in key metrics like correctness, fluency, and citation quality on renowned benchmark datasets.

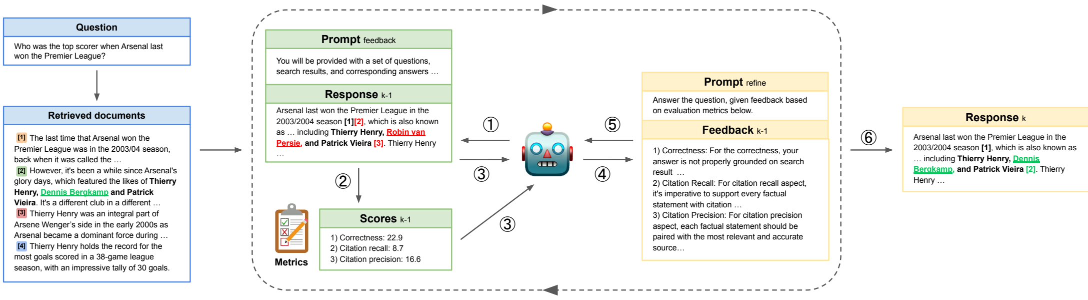

Figure 2: An overall architecture of $\mathsf{A}^2\mathsf{R}$ framework: ask access , refine . Starting with an input sentence and evidence, the framework generates an initial response, assesses its quality through automatic metrics, provides metric-informed feedback, and iteratively refines the output. This process encapsulates phases from initial response generation to the eventual iterative refinement, ensuring enhanced correctness, fluency, and citation quality in the final output.

Algorithm 1 Iterative Response Refinement with Metric-Guided Feedback

Require: model $\mathcal{M}$, input sentence x, evidence e, task-specific instruction prompt $\mathsf{p}_{\mathsf{init}}$, task-specific feedback prompt p_{fb}, task-specific refinement prompt $\mathsf{p}_{\mathsf{refine}}$, maximum iterations $\mathcal{K}$

Initialize $\mathcal{Y}_0 = \mathcal{M}([\mathsf{p}_{\mathsf{init}}; \mathsf{x}; \mathsf{e}])$ (Equation 1)
for $t = 0$ to $\mathcal{K} - 1$ **do**
  Assess output $\mathcal{Y}_t$ to obtain $\mathcal{S}_t$ (Equation 2)
  Generate feedback $\mathcal{F}_t = \mathcal{M}([\mathsf{p}_{\mathsf{fb}}; \mathcal{S}_t; \mathsf{x}; \mathsf{e}; \mathcal{Y}_t])$ (Equation 3)
  Refine response using feedback: $\mathcal{Y}_{t+1} = \mathcal{M}([\mathsf{p}_{\mathsf{refine}}; \mathsf{x}; \mathsf{e}; \mathcal{Y}_t; \mathcal{F}_t])$ (Equation 4)
  if stopping condition is met **then**
    Break
  end if
end for
return Refined output $\hat{\mathcal{Y}} = \mathcal{M}(\mathcal{H})$, where $\mathcal{H} = \left[\mathsf{p}_{\mathsf{refine}}; \mathsf{x}; \mathsf{e}; \mathcal{Y}_0; \mathcal{F}_0, \cdots, \mathcal{Y}_t, \mathcal{F}_t\right]$ (Equation 5)

2 Ask, Assess, and Refine: Metric-Guided Iterative Feedback Learning

2.1 Task Definition

Given an input sentence x and evidence e, we aim to (1) generate an initial output $\mathcal{Y}_t$ as detailed in Section 2.2, (2) assess this output using automatic evaluation metrics, obtaining $\mathcal{S}_t$, as discussed in Section 2.3, (3) provide metric-guided feedback $\mathcal{F}_t$ on the output as described in Section 2.4, and (4) refine the output based on the feedback to achieve a refined result $\mathcal{Y}_{t+1}$, which is elaborated in Section 2.5.

The *Initial Response Generation* phase aims to produce an initial output using task-specific instruction prompts. During the *Assessment of Output Quality* phase, we explicitly evaluate various aspects of the output, including correctness, fluency, and citation quality. This evaluation helps in generating optimal natural language feedback to enhance the output. In the *Metric-Guided Natural Language Feedback* phase, we produce feedback for each aspect using LLMs. The task-specific feedback prompt for these LLMs is crafted based on the assessment results. Finally, in the *Iterative Response Refinement* phase, the initial output is refined by harnessing metric-guided feedback.

2.2 Initial Response Generation

Prompt
Given a list of web search results, write an accurate answer for the question using only the provided web search results. Carefully follow the rules below while performing this task. - The answer should be detailed, correct, high-quality, and written by an expert using an unbiased and journalistic tone. - Be objective. Avoid injecting personal biases or opinions into the answer. Stick to the facts and let the search results speak for themselves. - Cite search results using [index]. Cite the most relevant results that answer the question. Don't cite irrelevant results. All sentences should have at least one citation. Document [1]: ... ... Document [5]: ... Question: {Question} Answer:

Table 1: Task-specific instruction prompt $\mathsf{p}_{\mathsf{init}}$ for Initial Response Generation on ELI5 and ASQA.

Suppose that we have the input sequence x, evidence e, and task-specific instruction prompt $\mathsf{p}_{\mathsf{init}}$, respectively. The model $\mathcal{M}$[3] generates an initial output $\mathcal{Y}_t$ as follows:

$$\mathcal{Y}_t = \mathcal{M}([\mathsf{p}_{\mathsf{init}}; \mathsf{x}; \mathsf{e}]) \tag{1}$$

where $[\cdot;\cdot]$ denotes concatenation. Note that our assumption is that an initial output $\mathcal{Y}_t$ may contains

[3]Here, we used GPT-3.5-TURBO-16K for all experiments and temperature is 0.7.

hallucination and inaccurate citations. We describe task-specific instruction prompt $\mathsf{p}_{\mathsf{init}}$ in Table 1.

2.3 Assessment of Output Quality

The Self-Refine mechanism (Madaan et al., 2023) uses LLMs for an intrinsic assessment of the quality of its outputs. This assessment serves as a trigger for iterative feedback. Importantly, the mechanism endeavors to enhance its initial output by relying solely on its inherent capabilities. While some studies (Gao et al., 2023a; Wang et al., 2023; Zhong et al., 2023) showcase the advancements in LLMs, Huang et al. (2023) emphasizes potential issues, suggesting that such approaches may diminish an LLM's capacity for self-correction without external input.

Given a ground truth $\tilde{\mathcal{Y}}$ and a set of automatic evaluation metrics for each aspect, denoted as $\mathfrak{E} = \{\mathcal{E}_1, \cdots, \mathcal{E}_n\}$, representing n distinct metric settings[4], we evaluate the quality of the current output $\mathcal{Y}$ as follows:

$$\mathcal{S}_t = \left[\mathcal{E}_1(\mathcal{Y}_t, \tilde{\mathcal{Y}}), \cdots, \mathcal{E}_n(\mathcal{Y}_t, \tilde{\mathcal{Y}})\right] \in \mathbb{R}^n \quad (2)$$

Our foundational hypothesis posits that an *explicit* evaluation of the discrepancy between $\mathcal{Y}$ and $\hat{\mathcal{Y}}$ provides more valuable insights compared to a mere *intrinsic* assessment. Explicitly quantifying the error allows for a more targeted refinement process, ensuring that feedback mechanisms are better informed and more precise in their adjustments.

2.4 Metric-Guided Natural Language Feedback

In large language models, generating feedback based on evaluation metrics requires a deep understanding of the context involved. When the model processes the evaluation feedback represented by $\mathcal{S}_t$, it's crucial for it to grasp the correct context of both the input sequence x and its initial response $\mathcal{Y}_t$.

The essence of effective feedback lies in its adaptability. The optimal feedback identifies the flaws in initial outputs and adjusts its advice based on the delicate balance of context and requirements.

[4]For each dataset, we employed metrics as introduced in ALCE (Gao et al., 2023b) to assess both correctness and citation quality. Note that we did not employ the MAUVE which is found to be sensitive to output length, indicating that may provide unstable results. Lastly, in real-world applications, various automatic evaluation metrics (e.g., AlignScore (Zha et al., 2023), FActScore (Min et al., 2023)) could be alternatively employed without ground truth.

Prompt
You will be provided with a set of questions, search results, and corresponding answers. Your task is to evaluate each answer and provide feedback to enhance its quality. Following <Feedback Instruction>, offer specific feedback according to the reward scores for the following aspects: Correctness, Citation Recall, and Citation Precision. <Feedback Instruction> 1) Correctness: If the reward score is below **{correctness_score}**, provide feedback to generate more relevant responses based on the search result summaries. If the score is above **{correctness_score}**, affirm that performance is satisfactory. 2) Citation Recall: If the reward score is below **{citation_recall_score}**, provide feedback to offer citations from credible sources for each factual statement you make. If the score is above **{citation_recall_score}**, affirm that performance on citation recall is satisfactory. 3) Citation Precision: If the reward score is below **{citation_score}**, provide feedback to cite properly, ensuring all factual statements refer to an appropriate search result. If the score is above **{citation_precision_score}**, affirm that performance on citation precision is satisfactory. Feedback:

Table 2: Task-specific feedback prompt $\mathsf{p}_{\mathsf{feedback}}$ for metric-guided natural language feedback on ELI5 and ASQA. This prompt provides appropriate feedback based on specific score thresholds for each aspect.

As we delve into iterative improvements, each feedback cycle should seamlessly fit with the main goals, ensuring clear advancement. Therefore, our metric-guided approach aims to provide feedback that is both quantitative (derived from metrics) and qualitative (naturally articulated). The natural language feedback $\mathcal{F}_t$ is formulated as:

$$\mathcal{F}_t = \mathcal{M}([\mathsf{p}_{\mathsf{fb}}; \mathcal{S}_t; \mathsf{x}; \mathsf{e}; \mathcal{Y}_t]) \quad (3)$$

where p_{fb} is a task-specific feedback prompt as described in Table 2. We assume this approach offers more nuanced and actionable insights, allowing models to refine their outputs in subsequent iterations more effectively.

2.5 Iterative Response Refinement

Refinement in response generation, especially when driven by feedback, is a pivotal mechanism that advances the robustness and accuracy of generated outputs. This step illustrates that feedback extends beyond mere assessment and acts as a conduit for continuous improvement.

Central to the iterative response refinement is the synergy between feedback and refinement. Natural language feedback $\mathcal{F}_t$ informs the model about the aspects of the response $\mathcal{Y}_t$ that need rectification. Leveraging this feedback, the model re-calibrates its output, aiming for more accurate and contextually appropriate results:

$$\mathcal{Y}_{t+1} = \mathcal{M}([\mathsf{p}_{\mathsf{refine}}; \mathsf{x}; \mathsf{e}; \mathcal{Y}_t; \mathcal{F}_t]) \quad (4)$$

where $\mathsf{p}_{\mathsf{refine}}$ is task-specific refinement prompt. Convergence and stability are key factor of this iterative process. The system alternates between the feedback (Equation (3)) and refinement (Equation (4)) phases until a pre-defined stopping criterion is met[5], or a maximum iteration numbers $\mathcal{K}$ is reached. The culmination of this iterative process is the refined output $\hat{\mathcal{Y}}$, articulated as:

$$\mathcal{H} = \left[\mathsf{p}_{\mathsf{refine}}; \mathsf{x}; \mathsf{e}; \mathcal{Y}_0; \mathcal{F}_0, \cdots, \mathcal{Y}_{\mathcal{K}}, \mathcal{F}_{\mathcal{K}}\right]$$
$$\hat{\mathcal{Y}} = \mathcal{M}(\mathcal{H}) \quad (5)$$

In summary, iterative response refinement emphasizes the principle that optimal outcomes are achieved not in standalone attempts, but through consistent evaluation and iterative refinements.

3 Experiments

3.1 Dataset

We evaluate our proposed method on three development datasets: ASQA, QAMPARI, and ELI5, which is part of ALCE. For the evaluation, we employ a subset of the datasets proposed by ALCE, specifically sampling approximately 100 instances from each dataset[6]. These datasets are characterized as follows:

ASQA (Stelmakh et al., 2022). Classified as a long-form factoid dataset, ASQA consists of ambiguous questions sourced from AmbigQA. Each question necessitates multiple short answers to adequately address the various interpretations of the posed question.

QAMPARI (Amouyal et al., 2023). This factoid QA dataset is distinctive in that the answers comprise a list of entities, often extracted from distinct passages.

ELI5 (Fan et al., 2019). Serving as a long-form QA dataset, the answers in ELI5 are typically verbose, with an average length of 131 words per response.

[5]We set the threshold for stopping condition by averaging output of assessment for each aspects, including citation and correctness.

[6]The associated API costs influenced the decision to use a limited sample size.

Model	Fluency	Correctness	Citation	
	MAUVE	Claim Recall	Recall	Precision
w/o guided metric feedback				
ChatGPT$_{\mathcal{K}=0}$	42.5	**25.3**	**9.4**	**29.5**
Intrinsic$_{\mathcal{K}=1}$	**61.4**	23.7	8.5	22.7
Intrinsic$_{\mathcal{K}=2}$	54.3	24.3	8.0	22.7
w/ guided metric feedback				
ChatGPT$_{\mathcal{K}=0}$	42.5	25.3	9.4	29.5
A^{2}R$_{\mathcal{K}=1}$	65.9	25.3	11.2	31.2
A^{2}R$_{\mathcal{K}=2}$	**67.7**	**25.6**	**11.7**	**32.3**

Table 3: Performance comparison of different iterations of ChatGPT on the ELI5 dataset (Fan et al., 2019). Here, $\mathcal{K}$ denotes the iteration number, and best performance is shown in **bold** text. For evaluation of correctness, we use *claim recall* on ELI5 dataset.

3.2 Evaluation Metrics

To rigorously evaluate the performance of our model, we adopted evaluation metrics, each targeting a distinct aspect of the output, as following in ALCE (Gao et al., 2023b).

MAUVE (Pillutla et al., 2021). The *MAUVE* serves as an indicator of the model's fluency. By comparing text distributions, it provides a quantitative measure of the alignment between the model's generated content and the source material. Such a measure ensures that the model's output is not just factually accurate but also naturally structured and coherent.

EM Recall. To gauge the factual correctness of the generated response on ASQA dataset, we employ the *EM Recall*. It carefully looks into the output to verify that it encompasses all salient aspects and encapsulates the succinct answer. Essentially, this quantifies the rate of accurate short answers, offering an insight into the model's recall capability.

Claim Recall. While the ELI5 dataset predominantly provides long-form answers and employs ROUGE for evaluation, a contrasting evaluation is adopted by ALCE (Gao et al., 2023b). It introduces the *claim recall*, which leverages InstructGPT (Ouyang et al., 2022b) to produce three distinct "sub-claims". To evaluate this, we then employ the advanced natural language inference (NLI) model, TRUE (Honovich et al., 2022), verifying if the model's output appropriately encompasses these sub-claims.

Model	Correctness		Citation	
	Recall@5	Precision	Recall	Precision
w/o guided metric feedback				
ChatGPT$_{\mathcal{K}=0}$	**18.6**	**23.5**	**17.7**	19.0
Intrinsic$_{\mathcal{K}=1}$	6.4	2.4	10.5	**21.1**
Intrinsic$_{\mathcal{K}=2}$	4.8	1.9	9.2	20.3
w/ guided metric feedback				
ChatGPT$_{\mathcal{K}=0}$	18.6	23.5	17.7	19.0
A^{2}R$_{\mathcal{K}=1}$	18.6	24.4	19.8	21.4
A^{2}R$_{\mathcal{K}=2}$	**18.8**	**24.6**	**20.1**	**22.2**

Table 4: Performance comparison of different iterations of ChatGPT on the QAMPARI dataset (Amouyal et al., 2023). For evaluation of correctness, we use *Correctness Recall@5* and *Correctness Precision* on QAMPARI dataset.

Model	Fluency	Correctness	Citation	
	MAUVE	EM Recall	Recall	Precision
w/o guided metric feedback				
ChatGPT$_{\mathcal{K}=0}$	78.1	19.7	**43.2**	**61.1**
Intrinsic$_{\mathcal{K}=1}$	**72.9**	20.2	31.5	51.9
Intrinsic$_{\mathcal{K}=2}$	76.4	**20.4**	34.5	48.9
w/ guided metric feedback				
ChatGPT$_{\mathcal{K}=0}$	**78.1**	19.7	43.2	61.1
A^{2}R$_{\mathcal{K}=1}$	76.7	**21.9**	48.7	63.1
A^{2}R$_{\mathcal{K}=2}$	77.7	20.8	**52.0**	**65.2**

Table 5: Performance comparison of different iterations of ChatGPT on the ASQA dataset (Stelmakh et al., 2022). For evaluation of correctness, we use *EM Recall* on ASQA dataset.

Correctness Recall@5 / Precision. In our analysis on the QAMPARI dataset, we measure the *precision* and *recall* of model predictions by calculating their exact alignment with the gold-standard answer list. An important modification to this approach is introduced to provide for the user's practical needs. To satisfy the practical needs, if the model's prediction encompasses at least five accurate answers, termed as *Recall@5*.

Citation Recall / Precision. Ensuring the authenticity and pertinence of cited references is imperative, especially in contexts demanding verifiable information. The *Citation Recall / Precision* assesses the quality of citations, certifying that answers are corroborated by pertinent passages and devoid of extraneous references. Based on AIS (Rashkin et al., 2023), we leverage an NLI model[7] to demonstrate the degree of entailment, which in turn determines the adequacy of the citations in supporting the model's response.

[7] https://huggingface.co/google/t5_xxl_true_nli_mixture

3.3 Main Results

Table 3, 4 and 5 shows the performance of fluency, correctness, and citation quality on ELI5, QAMPARI, and ASQA, respectively. We summarized the main results below.

The feedback improves correctness and citation quality. The experimental results imply that feedback is a pivotal for enhancing both the correctness and citation quality of the responses. The feedback serves as a re-calibration mechanism for models, bridging the gaps between what is generated and what is expected, especially in the realms of citation and factual accuracy.

- **Iterative Refinement**: As evidenced by our results, iterative feedback refinement – represented by varying iterations denoted by $\mathcal{K}$ – yield improvements, especially in terms of citation precision and recall. Remarkably, with each subsequent iteration, the model improves with a substantial margin on citation quality.

- **Enhanced Correctness**: Utilizing feedback is crucial in enhancing the model's performance in terms of correctness. By using metrics such as claim recall, EM recall, correctness recall, and precision, which evaluate the model's alignment with the gold-standard answer list, it becomes evident that responses refined through feedback match the expected answers more closely. The integration of our feedback method results in responses that are not only sharper in precision but also demonstrate a marked improvement in aligning with gold-standard correctness.

- **Enhanced Citation Relevance**: The feedback also plays an indispensable role in sharpening the relevance of citations in model responses. One of the most striking observations from our experiments was the marked enhancement in citation recall and precision metrics across the ELI5, QAMPARI, and ASQA datasets. This emphasizes the model's heightened ability, post-feedback, to select and reference the most relevant passages while minimizing superfluous citations. In essence, with the aid of feedback, our model consistently excelled in backing its responses with more relevant and contextually pertinent citations, fortifying the trustworthiness of its outputs.

Improving correctness and citation quality might indirectly enhance fluency. As depicted Table 3, 4, the experimental results suggest a subtle yet significant correlation between the fluency of generated responses and improvements in their correctness and citation quality. Especially, even without feedback targeting the improvement of fluency, the fluency exhibits a remarkable improvement or preserves comparable results across varying iterations.

One plausible explanation for this phenomenon is that as the model becomes more skillful at synthesizing information from reliable passages, it indirectly benefits from the coherent structure and consistency intrinsic to these passages. Generating responses based on factually grounded passages naturally focuses on producing output with fine-grained semantic representation and syntactically correctness.

Moreover, a boost in citation quality implies that the model increasingly refers to well-represented and structured passages, thereby contributing intrinsically to the fluency of the generated responses. This implies potential avenues for future work, suggesting that a targeted improvement in specific metrics can unintentionally lead to enhancements in other unrelated aspects of LLMs.

4 Analysis

4.1 Does Metric-Guided Feedback Really Elevate the Quality?

Our experimental results from Tables 3, 4, and 5 highlight a substantial enhancement in model performance upon the integration of metric-guided feedback. Particularly, noteworthy is its magnified influence on iterative models, which indicates that models refined across iterations may be more adjusted to the feedback if with metric-guided feedback. Furthermore, an intriguing observation is the tendency for the best performance to demonstrate not in the primary iterations but in subsequent ones, suggesting that the efficacy of feedback compounds over time. In our experiments, $\mathsf{A}^2\mathsf{R}_{\mathcal{K}=2}$ on the benchmark datasets showed remarkable performance improvements when feedback was incorporated, especially in terms of Correctness and Citation Quality. On the contrary, the absence of guided metric feedback led to a significant decrease in performance[8].

4.2 Case Studies

Retrieval Error
Question: Where did administrators of the UN Development Programme attend school?
Document (Title: United Nations International School of Hanoi) [1]: United Nations International School of Hanoi The United Nations International School of Hanoi is an international school in Hanoi, Vietnam. ...
...
Document [5] (Koc School): Model United Nations Development Programme conference each year at its campus since 2001. ...
Initial output: United Nations International School of Hanoi [1], UEA School of International Development [3], Turin School of Development [4], Model United Nations Development Programme conference [5].
Ground truth: University of Auckland, Yale Law School, Marlborough College, Boston University School of Law, University of California, Berkeley.

Table 6: A case study for *retrieval error* on QAMPARI dataset.

Retrieval Error. In our examination of the "correctness" metric from Table 3, 4, and 5, while we observe incremental improvements, we do not witness substantial growth in performance. One possible for phenomenon for explanation is the *retrieval error*, where the retrieved documents fetched are not relevant to the question. This misalignment suggests that both the generation and refinement processes, which rely on these documents, are inevitably hindered. To delve deeper into this phenomenon, we spotlight a case study drawn from the QAMPARI dataset.

For example, as illustrated in Table 6, the retrieved documents – ranked based on relevance scores with the question, a benchmark intrinsic to the dataset and not influenced by our approach – fail to provide the information regarding the educational institutions attended by administrators of the UN Development institutions. As a result, the ground truth, which cites institutions like University of Auckland", cannot be extracted from documents. Furthermore, while the initial output enumerates name of institutions or programs found within the documents, it falls short in directly addressing the question "*Where did administrators of the UN Development Programme attend school?*", which do not elaborate the academic background of the administrators, leading to inaccurate answer.

[8]Experimentally, the optimal results were observed when refinement was carried out up to $\mathcal{K} = 2$.

Feedback Error
For the correctness aspect, your answer is factually accurate and directly relevant to the provided search results. Great job! For citation recall aspect ... For citation precision aspect ... Reward Score of Correctness: 0 Reward Score of Citation Recall: 57 Reward Score of Citation Precision: 80

Table 7: A case study for *feedback – metric mismatch error* on ASQA dataset.

Feedback – Metric Mismatch Error. We observe the discrepancies in generating feedback based on evaluation metrics, namely *Feedback – Metric Mismatch Error*. As shown in Table 7, the positive feedback is generated while a reward score of correctness is 0. This discrepancies arise when the feedback generated fails to mirror these scores faithfully, implying that for any evaluation mechanism to be effective, the feedback given must intricately align with its respective evaluation metric scores. This misalignment in the initial feedback do not only leads to subsequent iterations showing lower scores but also further compounds the problem by generating inappropriate or inaccurate feedback. Moreover, primarily the generated feedback using LLM, occasionally falls short in capturing the differences in reward scores. Overall, if the initial feedback isn't accurate, the LLM-generated feedback can seem too general or even clearly mismatched.

4.3 Qualitative Analysis

We show the qualitative analysis as indicated in Table 8, 9, and 10 for clarifying the effectiveness of metric-guided feedback.

Citation Quality. The refined answer clarifies "[3]" to indicate which document provided the information, instead of just mentioning "James Earl Jones", which implies that leads to better citation quality.

Enhanced Brevity and Clarity. The introduction about James Earl Jones and David Prowse's roles is more concise and simplified. Furthermore, removing some redundant phrases, such as "It is worth noting that" and "It is important to mention that", suggesting that our method may enable to make the content more understandable.

Structured Information. The order of information from the refined answer is more logical, guiding the reader from the primary voice actor, James Earl Jones, to various other roles he played, and then to other voice actors.

5 Related Work

The Retrieval-augmented LLMs has witnessed significant advancements. This methodology tailors LLM outputs based on evidence extracted from external knowledge sources (Khandelwal et al., 2020; Lewis et al., 2020; Borgeaud et al., 2022; Izacard et al., 2022; Zhong et al., 2022). A pioneering work by (Gao et al., 2023b) sets itself apart by not only generating outputs from LLMs but also appending them with citations linked to one or multiple passages. This approach supports a comprehensive evaluation, touching upon aspects like fluency, accuracy, and citation quality. Platforms such as Bing Chat and perplexity.ai have operationalized this by offering answers to user inquiries, referencing contemporary web pages, primarily harvested from closed search engines.

Drawing inspiration from human tendencies to refine their written content based on feedback, the Feedback-based Refinement approach offers a continuous enhancement process for textual content. Reinforcement learning, in this context, has become a popular tool to uplift the standard of generated content (Böhm et al., 2019; Stiennon et al., 2020; Ziegler et al., 2019; Wu et al., 2020; Ouyang et al., 2022a; Glaese et al., 2022; Akyürek et al., 2023; Fernandes et al., 2023). Approaches that either employ reference-based feedback or operate freely without stringent reference dependence have been proposed Maynez et al. (2020); Pasunuru and Bansal (2018); Gunasekara et al. (2021); Nan et al. (2021); Roit et al. (2023). In this paper, we pivot towards the methodologies pioneered by Gao et al. (2023b). By explicitly utilizing metrics, that evaluate correctness, citation recall, and precision, we generate guided feedback. This structured feedback becomes instrumental in refining and enhancing our initial output, bridging the gap between human-like refinement processes and computational excellence.

6 Conclusion

In this work, we have shed light on the challenges presented by hallucinations and factual inconsistencies in Large Language Models, issues evident even in advanced retrieval-augmented LLMs. We introduced the A^{2}R framework, emphasizing the integration of explicit evaluations to provide an in-depth assessment of aspects such as correctness and citation quality. By adopting this approach, our system utilizes actionable natural language feedback to iteratively refine outputs, leading to considerable advancements in response quality. Results across key datasets further substantiate the effectiveness of our method. As LLMs become increasingly pivotal in real-world applications, our findings demonstrate the necessity for rigorous, feedback-driven mechanisms to enhance their reliability and precision. Future endeavors should expand upon this foundation, further bridging the gap between machine-generated content and factual accuracy.

Limitations

This paper introduces the framework $\mathsf{A}^2\mathsf{R}$ for Iterative Response Refinement with Metric-Guided Feedback, however, still has several limitations as follows:

API Call Costs. Utilizing LLM necessitates invoking APIs, which invariably involves associated costs. While these costs can be manageable for smaller datasets or fewer calls, they can escalate significantly when dealing with large-scale data. This financial constraint can deter researchers with limited resources from fully leveraging the capabilities of LLM, potentially limiting the democratization of the technology.

Dependency on Quality of Evaluation Metric. The quality of the feedback generated by our method is intricately tied to the quality of the evaluation metric we used. If the metrics lack granularity or fail to accurately capture the nuances of the data, the resultant feedback might be suboptimal or even misleading. This can lead to a cyclical problem where poor metrics result in poor feedback, which in turn can skew the training or evaluation of models.evaluation of models.

Retrieval and Feedback Errors. As with many systems that rely on retrieval mechanisms, there's an inherent risk of retrieval errors. These errors might stem from issues such as misunderstanding of context, generating irrelevant context, or simply failing to extract the most suitable information. Moreover, feedback errors, which can arise due to inaccuracies in the feedback mechanism or subjective biases, can further compound the problem, potentially leading to the propagation of incorrect or suboptimal patterns in subsequent iterations or models.

While we believe that our work achieves significant strides in advancing the field, these limitations underscore the ongoing challenges and areas for future research.

References

Afra Feyza Akyürek, Ekin Akyürek, Aman Madaan, Ashwin Kalyan, Peter Clark, Derry Wijaya, and Niket Tandon. 2023. Rl4f: Generating natural language feedback with reinforcement learning for repairing model outputs. *arXiv preprint arXiv:2305.08844*.

Samuel Joseph Amouyal, Tomer Wolfson, Ohad Rubin, Ori Yoran, Jonathan Herzig, and Jonathan Berant. 2023. Qampari: An open-domain question answering benchmark for questions with many answers from multiple paragraphs.

Florian Böhm, Yang Gao, Christian M. Meyer, Ori Shapira, Ido Dagan, and Iryna Gurevych. 2019. Better rewards yield better summaries: Learning to summarise without references. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 3110–3120, Hong Kong, China. Association for Computational Linguistics.

Sebastian Borgeaud, Arthur Mensch, Jordan Hoffmann, Trevor Cai, Eliza Rutherford, Katie Millican, George Bm Van Den Driessche, Jean-Baptiste Lespiau, Bogdan Damoc, Aidan Clark, et al. 2022. Improving language models by retrieving from trillions of tokens. In *International conference on machine learning*, pages 2206–2240. PMLR.

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.

Patrick Fernandes, Aman Madaan, Emmy Liu, António Farinhas, Pedro Henrique Martins, Amanda Bertsch, José GC de Souza, Shuyan Zhou, Tongshuang Wu, Graham Neubig, et al. 2023. Bridging the gap: A survey on integrating (human) feedback for natural language generation. *arXiv preprint arXiv:2305.00955*.

Mingqi Gao, Jie Ruan, Renliang Sun, Xunjian Yin, Shiping Yang, and Xiaojun Wan. 2023a. Human-like summarization evaluation with chatgpt.

Tianyu Gao, Howard Yen, Jiatong Yu, and Danqi Chen. 2023b. Enabling large language models to generate text with citations.

Amelia Glaese, Nat McAleese, Maja Trębacz, John Aslanides, Vlad Firoiu, Timo Ewalds, Maribeth Rauh, Laura Weidinger, Martin Chadwick, Phoebe Thacker, et al. 2022. Improving alignment of dialogue agents via targeted human judgements. *arXiv preprint arXiv:2209.14375*.

Chulaka Gunasekara, Guy Feigenblat, Benjamin Sznajder, Ranit Aharonov, and Sachindra Joshi. 2021. Using question answering rewards to improve abstractive summarization. In *Findings of the Association for Computational Linguistics: EMNLP 2021*, pages 518–526.

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 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 3905–3920, Seattle, United States. Association for Computational Linguistics.

Jie Huang, Xinyun Chen, Swaroop Mishra, Huaixiu Steven Zheng, Adams Wei Yu, Xinying Song, and Denny Zhou. 2023. Large language models cannot self-correct reasoning yet.

Gautier Izacard, Patrick Lewis, Maria Lomeli, Lucas Hosseini, Fabio Petroni, Timo Schick, Jane Dwivedi-Yu, Armand Joulin, Sebastian Riedel, and Edouard Grave. 2022. Atlas: Few-shot learning with retrieval augmented language models.

Urvashi Khandelwal, Omer Levy, Dan Jurafsky, Luke Zettlemoyer, and Mike Lewis. 2020. Generalization through memorization: Nearest neighbor language models.

Patrick S. H. Lewis, Ethan Perez, Aleksandra Piktus, Fabio Petroni, Vladimir Karpukhin, Naman Goyal, Heinrich Küttler, Mike Lewis, Wen-tau Yih, Tim Rocktäschel, Sebastian Riedel, and Douwe Kiela. 2020. Retrieval-augmented generation for knowledge-intensive NLP tasks. In *Advances in Neural Information Processing Systems 33: Annual Conference on Neural Information Processing Systems 2020, NeurIPS 2020, December 6-12, 2020, virtual*.

Aman Madaan, Niket Tandon, Prakhar Gupta, Skyler Hallinan, Luyu Gao, Sarah Wiegreffe, Uri Alon, Nouha Dziri, Shrimai Prabhumoye, Yiming Yang, Sean Welleck, Bodhisattwa Prasad Majumder, Shashank Gupta, Amir Yazdanbakhsh, and Peter Clark. 2023. Self-refine: Iterative refinement with self-feedback. *CoRR*, abs/2303.17651.

Joshua Maynez, Shashi Narayan, Bernd Bohnet, and Ryan McDonald. 2020. On faithfulness and factuality in abstractive summarization. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 1906–1919, Online. Association for Computational Linguistics.

Sewon Min, Kalpesh Krishna, Xinxi Lyu, Mike Lewis, Wen-tau Yih, Pang Koh, Mohit Iyyer, Luke Zettlemoyer, and Hannaneh Hajishirzi. 2023. FActScore: Fine-grained atomic evaluation of factual precision in long form text generation. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 12076–12100, Singapore. Association for Computational Linguistics.

Feng Nan, Cicero Nogueira dos Santos, Henghui Zhu, Patrick Ng, Kathleen McKeown, Ramesh Nallapati, Dejiao Zhang, Zhiguo Wang, Andrew O. Arnold, and Bing Xiang. 2021. Improving factual consistency of abstractive summarization via question answering. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 6881–6894, Online. Association for Computational Linguistics.

Long Ouyang, Jeffrey Wu, Xu Jiang, Diogo Almeida, Carroll Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, et al. 2022a. Training language models to follow instructions with human feedback. *Advances in Neural Information Processing Systems*, 35:27730–27744.

Long Ouyang, Jeffrey Wu, Xu Jiang, Diogo Almeida, Carroll L. Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, John Schulman, Jacob Hilton, Fraser Kelton, Luke Miller, Maddie Simens, Amanda Askell, Peter Welinder, Paul F. Christiano, Jan Leike, and Ryan Lowe. 2022b. Training language models to follow instructions with human feedback. In *NeurIPS*.

Ramakanth Pasunuru and Mohit Bansal. 2018. Multi-reward reinforced summarization with saliency and entailment. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 2 (Short Papers)*, pages 646–653, New Orleans, Louisiana. Association for Computational Linguistics.

Krishna Pillutla, Swabha Swayamdipta, Rowan Zellers, John Thickstun, Sean Welleck, Yejin Choi, and Zaid Harchaoui. 2021. MAUVE: Measuring the gap between neural text and human text using divergence frontiers. In *Advances in Neural Information Processing Systems*.

Hannah Rashkin, Vitaly Nikolaev, Matthew Lamm, Lora Aroyo, Michael Collins, Dipanjan Das, Slav Petrov, Gaurav Singh Tomar, Iulia Turc, and David Reitter. 2023. Measuring Attribution in Natural Language Generation Models. *Computational Linguistics*, pages 1–64.

Paul Roit, Johan Ferret, Lior Shani, Roee Aharoni, Geoffrey Cideron, Robert Dadashi, Matthieu Geist, Sertan Girgin, Leonard Hussenot, Orgad Keller, Nikola Momchev, Sabela Ramos Garea, Piotr Stanczyk, Nino Vieillard, Olivier Bachem, Gal Elidan, Avinatan Hassidim, Olivier Pietquin, and Idan Szpektor. 2023. Factually consistent summarization via reinforcement learning with textual entailment feedback. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 6252–6272, Toronto, Canada. 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.

Nisan Stiennon, Long Ouyang, Jeff Wu, Daniel M. Ziegler, Ryan J. Lowe, Chelsea Voss, Alec Radford, Dario Amodei, and Paul Christiano. 2020. Learning to summarize from human feedback. *ArXiv*, abs/2009.01325.

Jiaan Wang, Yunlong Liang, Fandong Meng, Zengkui Sun, Haoxiang Shi, Zhixu Li, Jinan Xu, Jianfeng Qu, and Jie Zhou. 2023. Is chatgpt a good nlg evaluator? a preliminary study.

Jeff Wu, Long Ouyang, Daniel M Ziegler, Nisan Stiennon, Ryan Lowe, Jan Leike, and Paul Christiano. 2020. Recursively summarizing books with human feedback. In *Advances in Neural Information Processing Systems*.

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.

Qihuang Zhong, Liang Ding, Juhua Liu, Bo Du, and Dacheng Tao. 2023. Can chatgpt understand too? a comparative study on chatgpt and fine-tuned bert.

Zexuan Zhong, Tao Lei, and Danqi Chen. 2022. Training language models with memory augmentation. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 5657–5673, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Daniel M Ziegler, Nisan Stiennon, Jeffrey Wu, Tom B Brown, Alec Radford, Dario Amodei, Paul Christiano, and Geoffrey Irving. 2019. Fine-tuning language models from human preferences. *arXiv preprint arXiv:1909.08593*.

A Qualitative Analysis

Qualitative Evaluation
Question: Who plays the voice of darth vader in Star Wars? Document [1] (Title: Darth Vader): ... ... Document [5] (Title: I Am Your Father): ... Initial output: The voice of Darth Vader in the Star Wars films is primarily performed by James Earl Jones. James Earl Jones provided the iconic deep and menacing voice for Darth Vader in the original 1977 film "Star Wars: A New Hope" and its sequels "The Empire Strikes Back" (1980) and "Return of the Jedi" (1983) [3]. It is worth noting that David Prowse physically portrayed Darth Vader in the original trilogy, but his voice was dubbed by James Earl Jones [2].In addition to the original trilogy, James Earl Jones reprised the role of Darth Vader in various other Star Wars projects. He provided the voice for Darth Vader in the 2003 micro-series and the CGI animated film [1]. He also voiced Darth Vader in his cameos in the animated series "Star Wars Rebels," with the exception of the second-season finale where both Matt Lanter and James Earl Jones contributed their voices [1]. It is important to mention that there have been other actors who provided the voice of Darth Vader in certain adaptations. For example, Brock Peters voiced Darth Vader in the 1981 radio drama adaptation of Star Wars [4]. However, the most iconic and well-known voice of Darth Vader is that of James Earl Jones, who has become synonymous with the character. Reward Score of Correctness: 50 Reward Score of Citation Recall: 55 Reward Score of Citation Precision: 83 Feedback: For the correctness aspect, ... Great job! For citation recall aspect, you have cited the appropriate search results, but try to cite ... Keep up the good work. For citation precision aspect, ... Great job in maintaining a high level of precision. Refined answer: The voice of Darth Vader in the Star Wars films is primarily performed by James Earl Jones [3]. James Earl Jones provided the iconic deep and menacing voice for Darth Vader in the original 1977 film "Star Wars: A New Hope" and its sequels "The Empire Strikes Back" (1980) and "Return of the Jedi" (1983) [3]. David Prowse physically portrayed Darth Vader in the original trilogy, but his voice was dubbed by James Earl Jones [2].James Earl Jones also reprised the role of Darth Vader in various other Star Wars projects. He provided the voice for Darth Vader in the 2003 micro-series and the CGI animated film [1]. In the animated series "Star Wars Rebels," James Earl Jones voiced Darth Vader, except for the second-season finale where both Matt Lanter and James Earl Jones contributed their voices [1].Brock Peters also provided the voice of Darth Vader in the 1981 radio drama adaptation of Star Wars [4]. However, it is James Earl Jones' voice that is most iconic and well-known for the character of Darth Vader. Reward Score of Correctness: 66 Reward Score of Citation Recall: 75 Reward Score of Citation Precision: 100

Table 8: A case study for qualitative analysis on ASQA dataset.

Qualitative Evaluation
Question: Why do your testicles contract when you cough? Document [1] (Title: Pediatric Retractile Testicles): ... ... Document [5] (Title: Two Of A Kind: 7 Oddball Facts About Testicles That Will Make You Turn Your Head (And Cough)): ... Initial output: When you cough, your testicles contract due to a reflex called the cremasteric reflex. The cremaster muscle, which is connected to the testicles, contracts in response to certain stimuli, such as coughing or exposure to cold temperatures [1]. This reflex causes the testicles to be temporarily drawn up towards the body, closer to the inguinal canal. The purpose of this reflex is not entirely clear, but it is believed to be a protective mechanism. The contraction of the cremaster muscle helps to protect the testicles by pulling them away from potential sources of injury or trauma. It is thought that this reflex may have evolved as a way to safeguard the testicles from impacts or compression during activities that increase intra-abdominal pressure, such as coughing, sneezing, or lifting heavy objects [2]. It is important to note that the cremasteric reflex is a normal physiological response and is different from other conditions involving the testicles, such as undescended testicles or retractile testicles. Undescended testicles occur when the testicles fail to descend into the scrotum, while retractile testicles refer to the temporary retraction of the testicles into the inguinal canal due to the contraction of the cremaster muscle [1]. In summary, the contraction of the testicles when you cough is a reflex known as the cremasteric reflex. This reflex is believed to be a protective mechanism that helps to safeguard the testicles from potential injury or trauma. Reward Score of Correctness: 33 Reward Score of Citation Recall: 10 Reward Score of Citation Precision: 33 Feedback: For the correctness aspect, ... Great job! For citation recall aspect, ... but try to provide citations for all factual statements to support their accuracy. For citation precision aspect, each citation should directly support the corresponding factual statement. Make sure to cite the most relevant and accurate source from the search results for each claim. Keep up the good work! Refine answer: When you cough, your testicles contract due to a reflex called the cremasteric reflex [1]. The cremaster muscle, which is connected to the testicles, contracts in response to certain stimuli, such as coughing or exposure to cold temperatures. This reflex causes the testicles to be temporarily drawn up towards the body, closer to the inguinal canal. The purpose of this reflex is not entirely clear, but it is believed to be a protective mechanism. The contraction of the cremaster muscle helps to protect the testicles by pulling them away from potential sources of injury or trauma. It is thought that this reflex may have evolved as a way to safeguard the testicles from impacts or compression during activities that increase intra-abdominal pressure, such as coughing, sneezing, or lifting heavy objects. It is important to note that the cremasteric reflex is a normal physiological response and is different from other conditions involving the testicles, such as undescended testicles or retractile testicles. Undescended testicles occur when the testicles fail to descend into the scrotum, while retractile testicles refer to the temporary retraction of the testicles into the inguinal canal due to the contraction of the cremaster muscle [1]. Reward Score of Correctness: 66 Reward Score of Citation Recall: 12 Reward Score of Citation Precision: 50

Table 9: A case study for qualitative analysis on ELI5 dataset.

Qualitative Evaluation
Question: What movie did Syncopy Inc. produce with Christopher Nolan as the screenwriter? Document [1] (Title: Syncopy Inc.): ... ... Document [5] (Title: Emma Thomas): ... Initial output: Memento [3]. Reward Score of QA Recall: 0 Reward Score of QA Precision: 0 Reward Score of Citation Recall: 0 Reward Score of Citation Precision: 0 Feedback: For the QA recall aspect, the search results do not provide any information about S. A. Rajkumar composing music for a film directed by S. Narayan. Ensure you're capturing all the important points from the search results and question. Recheck the provided information and try to cover as many ground-truth answers as possible. For the QA precision aspect, since there is no relevant information in the search results, it is important to provide a clear and accurate answer. For citation recall and citation precision aspects, there is no need for feedback as there were no citations provided. Refine answer: The Prestige [3]. Reward Score of QA Recall: 20 Reward Score of QA Precision: 100 Reward Score of Citation Recall: 0 Reward Score of Citation Precision: 0

Table 10: A case study for qualitative analysis on QAMPARI dataset.