ISQA: Informative Factuality Feedback for Scientific Summarization

Anonymous ACL submission

Abstract

We propose *Iterative Facuality Refining on Informative Scientific Question–Answering (**ISQA**) feedback*[1], a method following human learning theories that employs model-generated feedback consisting of both positive and negative information. Through iterative refining of summaries, it probes for the underlying rationale of statements to enhance the factuality of scientific summarization. It does this in a fine-grained manner by asking a summarization agent to reinforce validated statements in positive feedback and fix incorrect ones in negative feedback. Our findings demonstrate that the *ISQA* feedback mechanism significantly improves the factuality of various open-source LLMs on the summarization task, as evaluated across multiple scientific datasets.

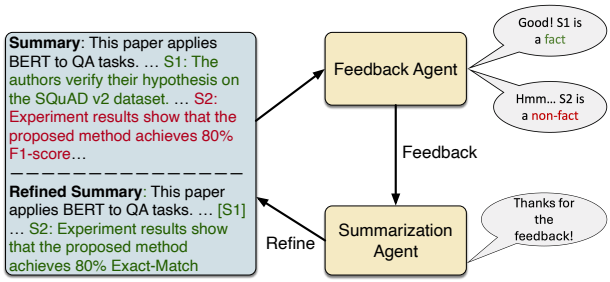

Figure 1: Illustration of improving factuality of summarization utilizing both positive (fact) and negative (non-fact) feedback. The facts and non-facts are highlighted in green and red, respectively. The summarization agent refines its output by remembering the facts and correcting non-facts.

1 Introduction

Large language model (LLM) based scientific summarization toolkits facilitate quicker screening of recent work by providing concise and fluent summaries of a large number of daily emerged scientific documents. However, the hallucination nature of generative LLMs introduces a fundamental flaw in the system-generated summaries: inconsistency with the original scientific documents (Lin et al., 2022; Zhang et al., 2023a; Min et al., 2023; Chen et al., 2021; Krishna et al., 2023). Despite their fluency, these nonfactual summaries generated by LLMs deter scientists due to the unreliability of the information provided.

Recent studies on refining LLM-generated content through feedback mechanisms have gained significant attention (Liu et al., 2023; Roit et al., 2023; Chen et al., 2023a). These studies enhance the output of an LLM by providing feedback on its previous outputs, thereby prompting the generation of improved content. Feedback mechanisms offer an effective and efficient post-hoc method to enhance LLM-generated content without the need for fine-tuning the LLM itself. In this work, we continue this line of work by exploring how feedback can be utilized to improve the factual consistency of LLM-generated scientific summaries.

Previous research has explored various sources of feedback, including human annotations (Liu et al., 2023), model-generated feedback (Roit et al., 2023; Chen et al., 2023a), external tools (Nathani et al., 2023), and self-correction mechanisms (Madaan et al., 2023). They largely emphasize the use of positive feedback to encourage desired outcomes, such as accurate reasoning paths. Yet, an essential aspect of the learning process, as highlighted by human learning theories like operant conditioning[2], is the strategic use of negative feedback. This type of feedback is instrumental in discouraging undesired behaviors, serving as a complementary role to positive feedback, yet has been overlooked in existing literature. In this work, we recognize the value of both positive and negative feedback and aim to combine them together for

[1]Code is available at `https://github.com/mt69JMMW/code-to-release.git`

[2]https://en.wikipedia.org/wiki/Operant_conditioning

factuality enhancement. As depicted in Figure 1, we expect a feedback agent to inspect a summary output and produce both positive feedback and negative feedback accordingly, which will be used to refine the summarization agent.

We propose to use model-generated positive feedback—specifically, facts[3] within summaries—to reinforce the generation of factual content by a summarization LLM, and use negative feedback—non-facts—to punish the generation of non-factual content. To distinguish between facts and non-facts in the summary for use as feedback, we employ a Question–Answering (QA) based method for feedback generation. The intuition is that by asking questions to the summary and comparing predicted answers with ground truth, we can infer the factuality of the evidence sentence in the summary that supports the answer. It is classified as a fact if the answer matches the ground truth, and otherwise non-fact. We refer to the proposed feedback mechanism as **I**nformative **S**cientific **Q**uestion-**A**nswering (**ISQA**) feedback.

In response to two distinct types of informative feedback, we use different prompts for the summarization LLM to process each feedback type appropriately. Echoing previous studies (Madaan et al., 2023; Nathani et al., 2023; Chen et al., 2023a), we leverage an iterative refinement approach to gradually improve the summarization. Through extensive experimentation with three open-source summarization LLMs, our results affirm that the *ISQA* enables LLMs to significantly improve the factuality of summaries over two prominent scientific document datasets. Notably, the performance improvement of *ISQA* surpasses that achieved through the application of both generic feedback and prompt engineering. Moreover, our findings reveal that compact LLMs, with the *ISQA* mechanism, can generate high-quality feedback comparable to those by the modern, closed-source model (GPT-3.5), which results in high-factuality summary output.

2 Related Work

Factuality of Abstractive Summarization Recent advanced automatic text summarization systems, especially LLM-based, have demonstrated a strong capability of generating fluent, controllable, and informative summaries(Aharoni et al., 2023; Bhattacharjee et al., 2023; Chen et al., 2023b; He et al., 2023). Despite performing well, these model-based summarization systems still confront certain issues that remain unsolved, one of which is generating factually inconsistent summaries (Balachandran et al., 2022; Tang et al., 2023; Liu et al., 2023). Chen et al. (2021) defines factuality as how well generated summaries are supported by source documents without using any external knowledge. On the other hand, factual error is defined as a statement that contradicts the source document or is not directly stated, heavily implied, or logically entailed by the source document (Krishna et al., 2023). In addressing this pressing issue, we introduce Informative Factuality Feedback as a means to effectively refine the summary generation, thereby achieving an overall enhancement in factuality.

Automated Feedback for LLMs Recent progress in rectifying LLMs' unsatisfactory behaviors mainly focuses on asking LLMs to learn from automated feedback (Pan et al., 2023). There are also two sources of automated feedback: external tools and LLM-generated. External feedback is usually obtained by employing external tools such as search engines.Nathani et al. (2023) argues the lack of error-specific feedback and proposes multi-aspect feedback, which seeks help from external tools based on different types of errors. In contrast, LLM-generated feedback comes from LLM critiquing its own output and refines it accordingly (Du et al., 2023; Welleck et al., 2022; Huang et al., 2022). Self-Refine (Madaan et al., 2023) proposes an iterative prompt refinement approach and applies it to various reasoning tasks where remarkable improvements have been witnessed. Liu et al. (2023) presents a method leveraging natural language feedback to enhance the factual consistency of summaries, through a self-constructed dataset with human demonstrations.

3 Method

As illustrated in Figure 2, the proposed method contains a summarization module and a feedback module, which interact iteratively. The feedback module is designed to provide informative feedback on the system-generated summary to enhance its factuality. Subsequently, the summarization module utilizes this feedback, along with the source document, to produce an improved summary.

[3]In this work, facts are defined to be factually consistent content in the summary. Similar definition to non-facts.

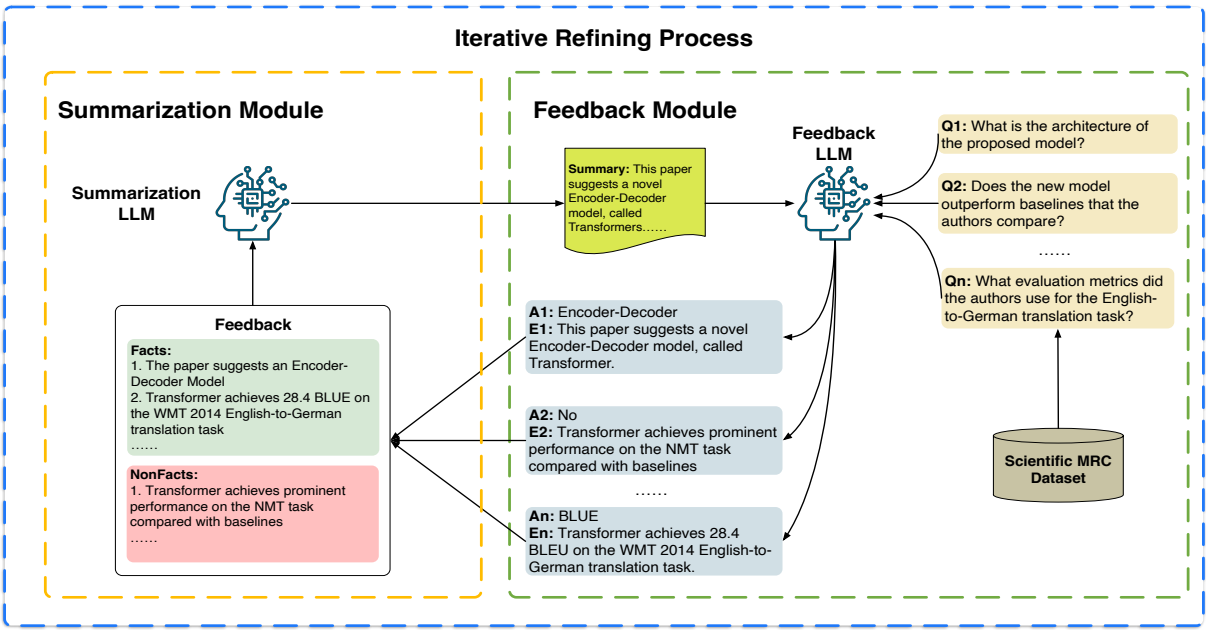

Figure 2: The framework of the proposed summarization method . The feedback LLM is trained to examine the factuality of the summary and then provide informative feedback by questioning. The summarization LLM then takes the feedback and produces an improved summary in factuality. The interaction between two modules is an iterative refining process.

3.1 Feedback Module

The feedback module plays a critical role akin to that of a doctor, tasked with examining the provided summary and offering constructive feedback. Next, we discuss how the feedback module performs such functionalities with our meticulous design.

Informative Scientific Question–Answering (*ISQA*) Feedback In analogy to a medical diagnosis, the feedback is expected to reveal factual and non-factual information of the summary, which contributes to refining the summarization agent. This objective is realized through two principal design features.

1) *Evidence-seeking QA*: As mentioned earlier, the supporting evidence that underpins the answer is the foundational source of the *ISQA* feedback. To acquire the evidence, we introduce an *evidence-seeking QA* task where the model is designed to elucidate its reasoning process through the provision of the evidence sentence, in addition to the answer prediction in the conventional QA tasks. The evidence sentences extracted are then categorically processed into facts and non-facts, which form the basis of the positive and negative aspects of the *ISQA* feedback, respectively. By adopting this innovative approach, we empower the feedback module to effectively analyze the factuality of summary content, thereby facilitating the generation of informative feedback for factuality refinement.

2) *Scientific Questions*: The *ISQA* feedback is constructed to offer comprehensive and domain-specific insights, meaning the feedback content should encompass information of varying depths and possess substantial scientific value. The goal is accomplished through the employment of high-quality scientific questions, especially human-annotated data. Unlike questions from other sources, such as Question–Generation models, human-annotated data guarantees that the questions meet the expectations of being domain-specific and encompass a range of difficulty levels.

Formally, the *ISQA* mechanism is defined as follows: given a set of scientific questions paired with its context $\{(q_1, c_1), (q_2, c_2), \cdots, (q_m, c_m)\}$ for a scientific paper x, the feedback model $\mathcal{F}$ is expected to output both the answer a and its supporting evidence e,

$$a, e = \mathcal{F}(c, q). \quad (1)$$

Based on the answer accuracy, evidence e is categorized into either positive or negative feedback. Refer to Algorithm1 line 7-13 for details.

Feedback LLM Recent studies (Madaan et al., 2023; Nathani et al., 2023) show that vanilla open-source LLMs, such as Vicuna (Chiang et al., 2023),

lack the capability to critique or correct their outputs[4]. Thus, we propose to fine-tune the vanilla LLMs on the ISQA mechanism to specialize in summary inspection and the generation of informative factuality feedback. It's worth noting that we still rely on the vanilla feedback LLM for generating fluent and coherent texts and following instructions[5]. We argue that the feedback model is capable of offering both insightful positive and negative feedback, provided it can accurately match the supporting evidence with the answers generated after being fine-tuned on the ISQA task.

3.2 Summarization Module

The summarization LLM, denoted as $\mathcal{G}$, is designed to process source documents alongside feedback and output a summary. To accommodate these inputs, we craft both summary-generation and act-on-feedback prompts[6]. In designing the summarization prompt, we adhere to common practices. The act-on-feedback prompt is crafted to include two sets of feedback, facts and non-facts, alongside instructions for refinement. Given the evolving nature of feedback prompts, we employ instruction-tuned LLMs to bolster their comprehension of instructions. We argue that the LLM can enhance its output by encouraging fact generation and discouraging non-facts, under the guidance of feedback prompts. We present the following template for summarization prompts[7]:

`Below is a scientific paper and a set of facts and non-facts in the paper. Please write a summary by memorizing facts and rectifying non-facts.`
`Facts:` $\{sent_1, sent_2, \cdots, sent_n\}$
`Non-Facts:` $\{sent_1, sent_2, \cdots, sent_n\}$
`Paper: "This is the body text of a scientific paper".`

3.3 Iterative Refining Process

In alignment with theories of human learning through feedback and insights from prior research, we adopt an iterative refinement strategy. Algorithm 1 shows the refinement process for a single scientific paper.

Algorithm 1 Iterative Refinement Process

Input: $\mathcal{G}, \mathcal{F}, p_0, x, D, N$
Initialize $F = \varnothing, \overline{F} = \varnothing$
for $t = 1, 2, \ldots N$ **do**
  $B \sim D$ ← *select a random QA batch*
  $s_t \leftarrow \mathcal{G}(x, p_{t-1})$
  for q_i, a_i^* **in** B **do** ← *batched refinement*
    $a_i, e_i \leftarrow \mathcal{F}(s_t, q_i)$
    $simScore \leftarrow F1(a_i, a_i^*)$
    if $simScore \geq \epsilon$ **then**
      add e_i to F
    else
      add e_i to $\overline{F}$
    end if
  end for
  $p_t \leftarrow$ update p_{t-1} with F and $\overline{F}$
end for

Initialization The inputs include the summarization model $\mathcal{G}$, feedback model $\mathcal{F}$, initial prompt for summarization p_0, scientific paper x and corresponding QA pairs D, the maximum number of refining iterations N, and initially empty sets for facts F and non-facts $\overline{F}$.

Batched Refinement To mitigate the potentially destabilizing effects of noisy questions—such as those outside the scope or of poor quality—we adopt a batched refinement strategy, as opposed to refining after every QA interaction. Specifically, we update the summary after processing a mini-batch of QA pairs, which are randomly selected from the complete set of QA pairs D. For each QA pair (q_i, a_i^*) in the batch, the feedback model addresses the question q_i using the current summary s_t and produces a predicted answer a_i along with supporting evidence e_i. We then calculate the text similarity (token-level F1) score between the predicted answer a_i and the ground truth answer a_i^*. Based on this score, e_i is classified as a fact in F if the score exceeds a predefined threshold ϵ; otherwise as a non-fact in $\overline{F}$.

4 Experiment

4.1 Datasets

SciMRC (Zhang et al., 2023b) SciMRC is a scientific machine reading comprehension dataset with questions across three difficulty levels: beginner, student, and expert. It includes 741 papers with a total of 6,057 QA pairs, distributed as 3,306 for beginner, 1,800 for student, and 951 for expert levels. We only use the training and validation sets

[4]This limitation also prevents vanilla LLMs from simultaneously acting as both the summarization LLM and the feedback LLM. Consequently, we employ distinct models for these tasks.

[5]This requires the feedback LLM to be instruction-tuned.

[6]Natural language prompts are adopted in this work for better interpretability.

[7]Efforts are made to keep prompts succinct, with the maximum number of facts and non-facts capped at eight to adhere to the input context length limitations of open-source LLMs

for ground truth answers, consisting of 592 papers and 4873 QA pairs.

QASPER (Dasigi et al., 2021) QASPER is a dataset comprised of 5,049 questions over 1,585 Natural Language Processing (NLP) papers. Questions are provided by NLP practitioners while answers are extracted by another set of practitioners who also provide corresponding supporting facts. For the same reason, we only use the training and validation sets, with 1169 papers in total.

4.2 Baseline Models

Summarization Model To demonstrate the efficacy of our proposed ISQA feedback mechanism, we utilize three open-source, instruction-tuned mainstream LLMs as our summarization baseline models: *Llama2-7b-chat* (Touvron et al., 2023), *Vicuna-7b* (Chiang et al., 2023), and *Flan-T5-11b* (Chung et al., 2022). It's important to note that these summarization LLMs operate in a zero-shot manner for summarization tasks, meaning they will not undergo fine-tuning, even when feedback is provided.

Feedback Model For the feedback model, we have empirically selected *Vicuna-7b* as the base model, which is further fine-tuned to excel in summary examination and providing constructive feedback. To benchmark against high-performing commercial solutions, we include GPT-3.5 as a baseline model[8]. The comparative analysis with GPT-3.5 is detailed in Section 5.3.

4.3 Evaluation Metrics

QAGS (Wang et al., 2020) QAGS is a QA-based metric to evaluate the factuality of model-generated content. It operates on the principle that asking questions about a summary and its source and then comparing the answers can reveal factual inconsistencies. If the summary is factually consistent with the source, similar answers are expected for the same questions asked of both. This method has shown substantially higher correlations with human judgments of factual consistency compared to other automatic evaluation metrics.

QuestEval (Scialom et al., 2021) QuestEval introduces another QA-based summarization factuality evaluation metric. Building upon previous precision-oriented QA-based metrics, including QAGS (Wang et al., 2020), QuestEval integrates precision and recall (Scialom et al., 2019) where the precision measures the consistency of generated summaries by token-level F1-scores of predicted answers, and the recall measures how much important information that generated summaries contain by query weighting[9].

ROUGE The ROUGE (Lin, 2004) scores measure the overlap of n-grams between machine-generated texts and reference texts. Higher ROUGE scores suggest closer similarity to human-written summaries, making it a standard text quality evaluation metric. We use the average of ROUGE-1, ROUGE-2, and ROUGE-L, denoted as ROUGE.

4.4 Main Results

As illustrated in Table 1, our ISQA feedback mechanism significantly enhances the factuality of summaries across all tested summarization models on both datasets. For instance, within the SciMRC dataset, the implementation of ISQA feedback across three LLMs leads to an average factuality improvement of 6.2% (54.41% $\rightarrow$ 60.61%) as measured by QAGS, and a 3.7% increase (29.3% $\rightarrow$ 33%) as assessed by QuestEval. Likewise, the QASPER dataset demonstrates more than 10% improvement in factuality benchmarks.

Furthermore, the average ROUGE score sees a modest enhancement of 0.97%, suggesting that the text quality of summaries remains consistent, whether or not ISQA feedback is applied. Notably, the Llama2 model outperforms others in terms of factuality, achieving a QAGS score on par with FLAN-T5 but surpassing it in the QuestEval score.

5 Analysis

5.1 Comparison Among Different Feedbacks

To validate the efficacy of ISQA feedback, we employ two distinct methodologies: generic feedback and pre-hoc prompt engineering. These approaches enable us to compare the impact of tailored feedback against more general feedback methods, highlighting the advantages of our ISQA design.

Pre-hoc prompt engineering Pre-hoc prompt engineering is designed to preempt the generation of non-factual content by incorporating preventive measures into the prompt itself, serving as a preemptive reminder. Echoing the adage that preven-

[8] GPT-3.5 is chosen over GPT-4 due to cost considerations.

[9] QuestEval also demonstrated its capability of better aligning human judgments in determining the factual consistency of summarization.

Summarization Models		SciMRC			QASPER		
		Factuality		Quality	**Factuality**		Quality
		QAGS ↑	QuestEval↑	ROUGE↑	QAGS ↑	QuestEval↑	ROUGE
Llama2-7b-chat	*w/o ISQA*	54.94%	28.90%	28.58%	43.48%	27.83%	25.54%
	w/ ISQA	**61.41%**	**33.62%**	**30.98%**	**48.82%**	**28.67%**	**26.78%**
Vicuna-7b	*w/o ISQA*	50.62%	29.39%	30.93%	49.60%	27.32%	25.81%
	w/ ISQA	**58.52%**	**32.96%**	**31.77%**	**56.19%**	**32.82%**	**27.19%**
FLAN-T5-11b	*w/o ISQA*	57.68%	29.62%	35.94%	54.47%	26.46%	26.04%
	w/ ISQA	**61.91%**	**32.42%**	35.19%	**58.33%**	**29.63%**	**26.77%**

Table 1: Results of three summarization LLMs with/without our proposed ISQA feedback on two datasets, SciMRC and QASPER. We evaluate the summarization in two aspects: factuality and text quality, with the main focus on the former. Bolded figures indicate improved results when compared to baseline (w/o ISQA).

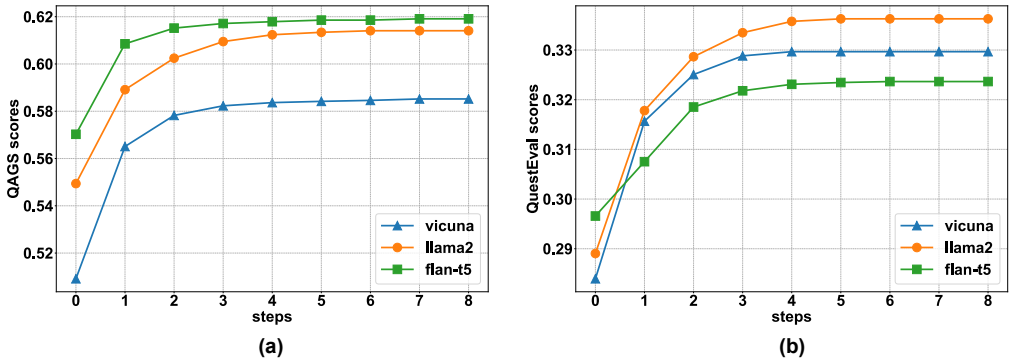

Figure 3: Correlation between the factuality improvement and the number of refining steps. Figure (a) and (b) depict the QAGS score and QuestEval score, respectively.

tion is challenging, we anticipate that our post-hoc *ISQA* will outperform pre-hoc preventative strategies. This expectation is validated by the results presented in Table 2, which show that prompt engineering is less effective than either generic or ISQA feedback (57.68% VS 59.88% and 61.91%)[10].

Generic feedback Unlike the ISQA feedback, generic feedback lacks specific insights on content factuality, merely indicating factual inaccuracies and urging improvement. Compared to prompt engineering, Table 2 shows a modest 2.2% QAGS gain in generic feedback but a more substantial 4.2% increase in ISQA, highlighting ISQA's effectiveness in providing detailed factual cues for prioritizing and correcting inaccuracies. This approach mirrors a teacher-student interaction, with ISQA acting as a guiding teacher, pinpointing not just the correctness (as generic feedback does) but also the specifics of errors or accuracy.

[10]Details on prompts used for prompt engineering are included in the Appendix D.

Feedback Type	QAGS	QuestEval
Pre-hoc Prompt	57.68%	29.62%
Generic Feedback	59.88%	28.89%
ISQA	**61.91%**	**32.42%**

Table 2: Results of different feedback strategies. The summarization model for demonstration is FLAN-T5.

5.2 Impact of Iterative Refining Process

In this section, we analyze the impact of having multiple steps of refining and quantify factuality improvements along the iterative refining process. In Figure 3, we plot correlation curves between both QAGS and QuestEval scores and the number of refining steps. The model's proficiency in generating factually consistent summaries exhibits a noteworthy augmentation as the refinement progresses. This observed increase aligns with our expectations and underscores the model's adaptability to the evolving feedback content.

Notably, around the fourth refinement step, a

discernible trend of gradual convergence emerges in the factuality of the generated summaries. This temporal convergence implies the existence of an upper bound on the extent to which the LLM can effectively integrate and leverage feedback. Beyond this point, the increment of factuality scores keeps diminishing, resembling a boundary effect. This nuanced understanding of the interaction between refinement steps and factual consistency contributes valuable insights into the dynamics of our refining process. It prompts further exploration into the optimal balance between the feedback mechanism and the resulting enhancement in the model's capacity to generate summaries of high factuality.

5.3 Different Feedback LLMs

In contrast to the reliance on large-scale commercial LLM products, a significant advantage of our method is the ability to achieve substantial improvements at a much lower cost by using a compact, open-source LLM for feedback. To demonstrate this, we compare the feedback model based on fine-tuned Vicuna with GPT-3.5. According to Table 3, the factuality scores between the two models are closely matched, with GPT-3.5 leading by a slight 0.33% on QuestEval and Vicuna outperforming by 0.68% on QAGS. This suggests that the feedback quality from both systems is comparably effective. Our findings highlight that with proper designing, a fine-tuned open-source LLM with 7 billion parameters can match the performance of GPT-3.5[11]. This not only proves to be cost-efficient but also ensures deterministic and reproducible outcomes.

Feedback Model	QAGS	QuestEval
GPT-3.5	61.41%	**32.70%**
Vicuna	**61.93%**	32.42%

Table 3: Results of comparison between different feedback models: fine-tuned Vicuna and GPT-3.5. The summarization model used is Flan-T5.

5.4 Importance of question quality

In Section 3.1, we utilize scientific QA datasets for feedback generation, where the questions are human-annotated and considered to be of high quality. These datasets, including SciMRC and QASPER, have demonstrated their effectiveness in enhancing feedback LLMs through previous experiments in Table 1. This section explores the potential of using system-generated questions. For this purpose, we generate questions for each scientific paper in the SciMRC dataset by extracting entities as answer candidates and then generating questions conditioned on these candidates. For simplicity, we employ the question-generation model proposed by Ushio et al. (2023).

Upon replacing the original SciMRC dataset with the one generated by the question generation model for *Iterative Factuality Refining*, the results—detailed in Table 4—show a significant decline in performance across both benchmarks when using model-generated questions, with QAGS falling even below the performance levels of models without any feedback. Thus, we argue that low-quality model-generated QA data is insufficient for high-quality feedback generation. Examples of both model-generated and human-annotated QA pairs are provided in the Appendix C for comparison.

Source of QA	QAGS	QuestEval
Model-generated	56.64%	31.34%
Human-annotated	**61.91%**	**32.42%**

Table 4: Comparison between different sources of QA data. The summarization model used is FLAN-T5.

5.5 Human Evaluation

We conduct a human evaluation on a sampled subset (20 papers) of SciMRC to further validate the factuality enhancement by ISQA. Three qualified annotators are hired for the evaluation and paid at the local compensation level. Annotations on five papers are used to calculate inter-agreement between annotators, with the Cohen Kappa coefficient $\mathcal{K}$ being 0.51, indicating acceptable agreement. For each summary, its factuality score is computed as the ratio of factual sentences out of all. The factuality score of a system is calculated as the average factuality score of all summaries. According to the result, the proposed ISQA leads to an 11% gain (79% $\rightarrow$ 90%) in the factuality over the baseline. We provide more details of human evaluation in the Appendix A.

5.6 Case Study

We provide a case study to demonstrate some insights into the behavior of LLMs during the iterative refining process. Since the feedback content is

[11] Although, we note that our feedback LLM may not compete with stronger products such as GPT-4.

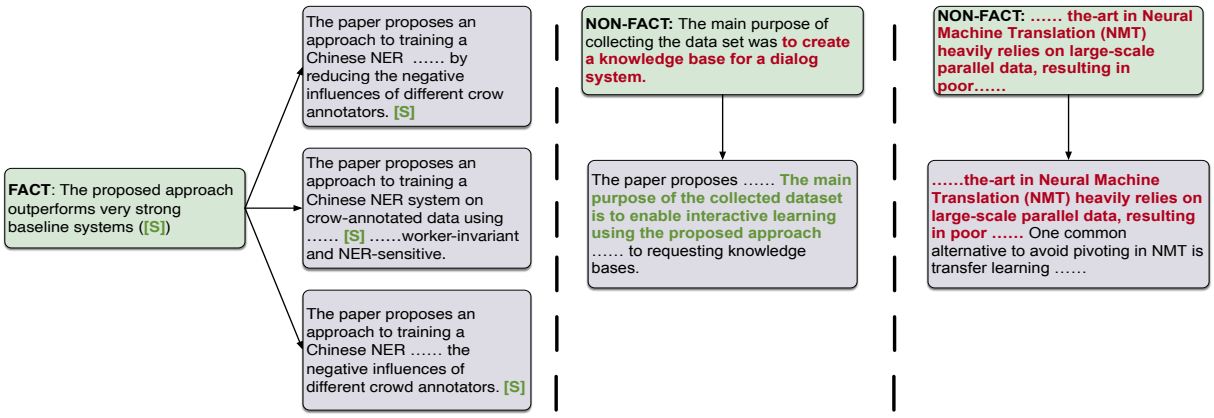

Figure 4: Examples of how LLMs respond to fact and non-fact feedback, respectively. In (a), **[S]** denotes the whole sentence. Arrows denote subsequent summaries after the model receives the feedback.

a combination of facts and non-facts found in the generated summaries, we show by case whether the model can understand the feedback and how the model makes decisions accordingly in Figure 4. More cases are provided in Appendix E.

LLM Memorizes Facts With regard to feedback pertaining to factual content, we observed that the LLM exhibits a tendency to memorize factual information. As depicted in Figure 4(a), when the feedback model highlights a factual signal such as "The proposed approach outperforms very strong baseline systems," subsequent summaries generated by the summarization model, while not necessarily replicating the exact phrasing, consistently contain this factual statement. The decision made by the model to memorize facts, as provided by feedback, is deemed a logical and effective strategy in the process of refining summary generation. This behavior not only underscores the model's capacity to respond to facts but also aligns with conventional human reasoning habits.

LLM Struggles with Non-facts On the other hand, the performance of the LLM in handling non-facts leaves room for improvement, as it struggles to make nuanced judgments on how to address non-factual content provided in the feedback. Figure 4(b) exemplifies successful instances where the model accurately comprehends the instruction and rectifies the error. For example, when the feedback indicates that "The main purpose of collecting the data set was to create a knowledge base for a dialog system." is a non-fact, LLMs are capable of executing the instruction to amend this non-fact as requested in feedback by identifying relevant content the original paper then output the factual information.

However, in certain instances, the LLM faces challenges in making effective decisions regarding non-fact content. For instance, as depicted in Figure 4(c), the summarization model retains the erroneous information in subsequent outputs despite the feedback explicitly stating the non-factual nature of a given statement. We attribute this phenomenon primarily to the LLM's extensive exposure to affirmative instructions during large-scale pre-training, such as "answer this question," with comparatively limited exposure to negation or contradiction. This implies that the LLM may not excel in negating information, contributing to its occasional difficulty in appropriately handling non-facts provided in the feedback.

6 Conclusion

In this paper, we introduce a novel paradigm for leveraging LLM-generated *ISQA* feedback to enhance the factual consistency of scientific summarization. The method, *Iterative Factuality Refinment on ISQA feedback* incorporates fine-grained factual feedback design and iterative refining process, enabling LLMs to directly learn from facts and non-facts and gradually improve the factual consistency. We show by extensive experiments that *ISQA* feedback can benefit various open-source summarization LLMs to refine their outputs. Furthermore, our approach shows the potential of compact LLMs to provide high-quality feedback that is comparable with GPT-3.5. Based on case studies, we propose future work on enabling LLMs to handle negative instructions.

Limitations

This study employs informative feedback through the fine-tuning of a compact LLM using high-quality, domain-specific question-answering data. However, the scarcity of such specialized datasets presents a significant obstacle. Consequently, the applicability of this method is domain-specific, contingent upon the availability of sufficient high-quality QA data for adaptation to other domains.

References

Roee Aharoni, Shashi Narayan, Joshua Maynez, Jonathan Herzig, Elizabeth Clark, and Mirella Lapata. 2023. Multilingual summarization with factual consistency evaluation. In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 3562–3591, Toronto, Canada. Association for Computational Linguistics.

Vidhisha Balachandran, Hannaneh Hajishirzi, William Cohen, and Yulia Tsvetkov. 2022. Correcting diverse factual errors in abstractive summarization via post-editing and language model infilling. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 9818–9830, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Abhik Bhattacharjee, Tahmid Hasan, Wasi Uddin Ahmad, Yuan-Fang Li, Yong-Bin Kang, and Rifat Shahriyar. 2023. CrossSum: Beyond English-centric cross-lingual summarization for 1,500+ language pairs. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2541–2564, Toronto, Canada. Association for Computational Linguistics.

Steven Bird, Ewan Klein, and Edward Loper. 2009. *Natural language processing with Python: analyzing text with the natural language toolkit*. " O'Reilly Media, Inc.".

Pinzhen Chen, Zhicheng Guo, Barry Haddow, and Kenneth Heafield. 2023a. Iterative translation refinement with large language models.

Yiran Chen, Pengfei Liu, and Xipeng Qiu. 2021. Are factuality checkers reliable? adversarial meta-evaluation of factuality in summarization. In *Findings of the Association for Computational Linguistics: EMNLP 2021*, pages 2082–2095, Punta Cana, Dominican Republic. Association for Computational Linguistics.

Yulong Chen, Yang Liu, Ruochen Xu, Ziyi Yang, Chenguang Zhu, Michael Zeng, and Yue Zhang. 2023b. UniSumm and SummZoo: Unified model and diverse benchmark for few-shot summarization. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 12833–12855, Toronto, Canada. Association for Computational Linguistics.

Wei-Lin Chiang, Zhuohan Li, Zi Lin, Ying Sheng, Zhanghao Wu, Hao Zhang, Lianmin Zheng, Siyuan Zhuang, Yonghao Zhuang, Joseph E. Gonzalez, Ion Stoica, and Eric P. Xing. 2023. Vicuna: An open-source chatbot impressing gpt-4 with 90%* chatgpt quality.

Hyung Won Chung, Le Hou, Shayne Longpre, Barret Zoph, Yi Tay, William Fedus, Yunxuan Li, Xuezhi Wang, Mostafa Dehghani, Siddhartha Brahma, Albert Webson, Shixiang Shane Gu, Zhuyun Dai, Mirac Suzgun, Xinyun Chen, Aakanksha Chowdhery, Alex Castro-Ros, Marie Pellat, Kevin Robinson, Dasha Valter, Sharan Narang, Gaurav Mishra, Adams Yu, Vincent Zhao, Yanping Huang, Andrew Dai, Hongkun Yu, Slav Petrov, Ed H. Chi, Jeff Dean, Jacob Devlin, Adam Roberts, Denny Zhou, Quoc V. Le, and Jason Wei. 2022. Scaling instruction-finetuned language models.

Pradeep Dasigi, Kyle Lo, Iz Beltagy, Arman Cohan, Noah A. Smith, and Matt Gardner. 2021. A dataset of information-seeking questions and answers anchored in research papers. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 4599–4610, Online. Association for Computational Linguistics.

Yilun Du, Shuang Li, Antonio Torralba, Joshua B Tenenbaum, and Igor Mordatch. 2023. Improving factuality and reasoning in language models through multiagent debate. *arXiv preprint arXiv:2305.14325*.

Pengcheng He, Baolin Peng, Song Wang, Yang Liu, Ruochen Xu, Hany Hassan, Yu Shi, Chenguang Zhu, Wayne Xiong, Michael Zeng, Jianfeng Gao, and Xuedong Huang. 2023. Z-code++: A pre-trained language model optimized for abstractive summarization. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 5095–5112, Toronto, Canada. Association for Computational Linguistics.

Matthew Honnibal and Ines Montani. 2017. spaCy 2: Natural language understanding with Bloom embeddings, convolutional neural networks and incremental parsing. To appear.

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

Kalpesh Krishna, Erin Bransom, Bailey Kuehl, Mohit Iyyer, Pradeep Dasigi, Arman Cohan, and Kyle Lo. 2023. LongEval: Guidelines for human evaluation of faithfulness in long-form summarization. In *Proceedings of the 17th Conference of the European Chapter of the Association for Computational Linguistics*, pages 1650–1669, Dubrovnik, Croatia. Association for Computational Linguistics.

Chin-Yew Lin. 2004. ROUGE: A package for automatic evaluation of summaries. In *Text Summarization Branches Out*, pages 74–81, Barcelona, Spain. Association for Computational Linguistics.

Stephanie Lin, Jacob Hilton, and Owain Evans. 2022. TruthfulQA: Measuring how models mimic human falsehoods. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 3214–3252, Dublin, Ireland. Association for Computational Linguistics.

Yixin Liu, Budhaditya Deb, Milagro Teruel, Aaron Halfaker, Dragomir Radev, and Ahmed Hassan Awadallah. 2023. On improving summarization factual consistency from natural language feedback. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 15144–15161, Toronto, Canada. Association for Computational Linguistics.

Aman Madaan, Niket Tandon, Prakhar Gupta, Skyler Hallinan, Luyu Gao, Sarah Wiegreffe, Uri Alon, Nouha Dziri, Shrimai Prabhumoye, Yiming Yang, Shashank Gupta, Bodhisattwa Prasad Majumder, Katherine Hermann, Sean Welleck, Amir Yazdanbakhsh, and Peter Clark. 2023. Self-refine: Iterative refinement with self-feedback. In *Thirty-seventh Conference on Neural Information Processing Systems*.

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.

Deepak Nathani, David Wang, Liangming Pan, and William Wang. 2023. MAF: Multi-aspect feedback for improving reasoning in large language models. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 6591–6616, Singapore. Association for Computational Linguistics.

Liangming Pan, Michael Saxon, Wenda Xu, Deepak Nathani, Xinyi Wang, and William Yang Wang. 2023. Automatically correcting large language models: Surveying the landscape of diverse self-correction strategies. *arXiv preprint arXiv:2308.03188*.

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.

Thomas Scialom, Paul-Alexis Dray, Sylvain Lamprier, Benjamin Piwowarski, Jacopo Staiano, Alex Wang, and Patrick Gallinari. 2021. QuestEval: Summarization asks for fact-based evaluation. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 6594–6604, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Thomas Scialom, Sylvain Lamprier, Benjamin Piwowarski, and Jacopo Staiano. 2019. Answers unite! unsupervised metrics for reinforced summarization models. 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 3246–3256, Hong Kong, China. Association for Computational Linguistics.

Liyan Tang, Tanya Goyal, Alex Fabbri, Philippe Laban, Jiacheng Xu, Semih Yavuz, Wojciech Kryscinski, Justin Rousseau, and Greg Durrett. 2023. Understanding factual errors in summarization: Errors, summarizers, datasets, error detectors. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 11626–11644, Toronto, Canada. Association for Computational Linguistics.

Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, Dan Bikel, Lukas Blecher, Cristian Canton Ferrer, Moya Chen, Guillem Cucurull, David Esiobu, Jude Fernandes, Jeremy Fu, Wenyin Fu, Brian Fuller, Cynthia Gao, Vedanuj Goswami, Naman Goyal, Anthony Hartshorn, Saghar Hosseini, Rui Hou, Hakan Inan, Marcin Kardas, Viktor Kerkez, Madian Khabsa, Isabel Kloumann, Artem Korenev, Punit Singh Koura, Marie-Anne Lachaux, Thibaut Lavril, Jenya Lee, Diana Liskovich, Yinghai Lu, Yuning Mao, Xavier Martinet, Todor Mihaylov, Pushkar Mishra, Igor Molybog, Yixin Nie, Andrew Poulton, Jeremy Reizenstein, Rashi Rungta, Kalyan Saladi, Alan Schelten, Ruan Silva, Eric Michael Smith, Ranjan Subramanian, Xiaoqing Ellen Tan, Binh Tang, Ross Taylor, Adina Williams, Jian Xiang Kuan, Puxin Xu, Zheng Yan, Iliyan Zarov, Yuchen Zhang, Angela Fan, Melanie Kambadur, Sharan Narang, Aurelien Rodriguez, Robert Stojnic, Sergey Edunov, and Thomas Scialom. 2023. Llama 2: Open foundation and fine-tuned chat models.

Asahi Ushio, Fernando Alva-Manchego, and Jose Camacho-Collados. 2023. A practical toolkit for multilingual question and answer generation. In *Proceedings of the 61th Annual Meeting of the Association for Computational Linguistics: System Demonstrations*, Toronto, Canada. Association for Computational Linguistics.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Ł ukasz Kaiser, and Illia Polosukhin. 2017. Attention is all

you need. In *Advances in Neural Information Processing Systems*, volume 30. Curran Associates, Inc.

Alex Wang, Kyunghyun Cho, and Mike Lewis. 2020. Asking and answering questions to evaluate the factual consistency of summaries. *arXiv preprint arXiv:2004.04228*.

Sean Welleck, Ximing Lu, Peter West, Faeze Brahman, Tianxiao Shen, Daniel Khashabi, and Yejin Choi. 2022. Generating sequences by learning to self-correct. *arXiv preprint arXiv:2211.00053*.

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Remi Louf, Morgan Funtowicz, Joe Davison, Sam Shleifer, Patrick von Platen, Clara Ma, Yacine Jernite, Julien Plu, Canwen Xu, Teven Le Scao, Sylvain Gugger, Mariama Drame, Quentin Lhoest, and Alexander Rush. 2020. Transformers: State-of-the-art natural language processing. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 38–45, Online. Association for Computational Linguistics.

Muru Zhang, Ofir Press, William Merrill, Alisa Liu, and Noah A. Smith. 2023a. How language model hallucinations can snowball.

Xiao Zhang, Heqi Zheng, Yuxiang Nie, Heyan Huang, and Xian-Ling Mao. 2023b. Scimrc: Multi-perspective scientific machine reading comprehension. *arXiv preprint arXiv:2306.14149*.

Appendices

A Human Evaluation

Annotator Selection Annotators selected are required to have decent English reading ability, and all have obtained a Bachelor's degree. To increase the diversity, we hired annotators with different majors, including Computer Science and Material Science. Before completing the full evaluation, each annotator was asked to finish a training process by evaluating one paper.

Annotation Data The annotation data are summaries generated by Llama2 with and without ISQA feedback. On average, there are 5 sentences in each summary. The average length of the source text is 6000 characters.

Compensation The rate of compensation is 15 SGD per hour, and we compensate annotators based on an average evaluation time.

Ethical Consideration In this study, strict ethical guidelines were adhered with the research protocol reviewed and approved by the Institutional Review Board (IRB) in our organization. For example, informed consent was obtained from all participants, who were adequately informed about the study's purpose, procedures, potential risks, and benefits, as well as their right to withdraw at any time without any consequences.

B Experiment Settings

Experiments can be done on 2x A100 GPUs with 70 GPU hours in total. We list parameters used for experiments in Table 5. We used NLTK (Bird et al., 2009) and spaCy (Honnibal and Montani, 2017) toolkits for text processing. Code is implemented in PyTorch (Vaswani et al., 2017) and Huggingface's Transformers (Wolf et al., 2020).

Parameters	Value
num beams	2
temperarture	1.3
no repeated n-gram	3
Summarization model	
max input length	2048
max new tokens	200
min new tokens	100
Feedback model	
max input length	512
max new tokens	100
min new tokens	10

Table 5: Experimental parameters.

C Question Comparison

Table 6 provides a comparison between human-annotated questions from the SciMRC dataset and questions generated by language models. We randomly pick three papers for demonstration and select 5 questions of each category. In general, human-annotated questions are more difficult and require a deeper understanding of the context. For example, to answer the fifth dataset question of paper A, the model needs to conduct multi-hop reasoning: Tt first locates two paper snippets for the authors' proposed method and traditional transfer methods, respectively. Then, it compares two pieces of information to derive the differences between them. However, model-generated questions don't require such complex reasoning chains.

D Prompts

We show all our prompt templates below.

D.1 Pre-hoc prompt engineering

Prompt 1

Read this: "This is the body text of a scientific paper"
Write a summary.

Prompt 2

Below is a scientific paper. Please write a summary about it.
Paper: "This is the body text of a scientific paper"

Prompt 3

Below is a scientific paper. Please write a summary that is factually consistent with the paper.
Paper: "This is the body text of a scientific paper"

Prompt 4

Write a summary of the scientific paper below. Reduce non-facts as much as possible in the summary you generate.
Paper: "This is the body text of a scientific paper"

Prompt 5

Please write a summary about the scientific paper below. Make sure the summary is factually consistent with the paper. Do not include non-factual information.
Paper: "This is the body text of a scientific paper"

D.2 Feedback generation

ISQA

Below is a question paired with context.
Please return your response in two parts:
1. The answer to the question
2. The most relevant evidence sentence in the context to your answer.
If the question is unanswerable, directly return 'Unanswerable'.
Question: "This is the body text of a question"
Context: "This is the body text of a scientific paper"

Generic feedback

Below is a scientific paper paired with feedback. Please write a summary based on the feedback provided.
Paper: "This is the body text of a scientific paper"
Feedback: "The previous summary you generated is not factually consistent with the paper. Please re-generate a summary with higher factuality. Reduce your factual errors and include more facts."

D.3 Summary generation with feedback

Below is a scientific paper paired with feedback of factual information. Please write a summary by memorizing the provided facts and rectifying the provided non-facts.
Facts:$\{sent_1, sent_2, \cdots, sent_n\}$
Non-Facts:$\{sent_1, sent_2, \cdots, sent_n\}$
Paper: "This is the body text of a scientific paper."

D.4 GPT-3.5

SYSTEM: You are good at answering reading comprehension questions and extracting corresponding evidence.
USER: Read the context
Context: "This is the body of a context"
Please answer the following question by:
1. Provide the answer
2. Provide the evidence sentence in the context.
Please use ""Answer:" and "Evidence:" to denote these two parts when generating your response: "This is the body of a question"

E Cases

In Table 7, we provide cases of LLMs' behaviors when confronted with facts in the feedback.
In Table 8, we provide successful cases of LLMs' behaviors when confronted with non-facts in the feedback.
In Table 9, we provide error cases of LLMs' behaviors when confronted with non-facts in the feedback.

Papers	Human-annotated Questions	Model-generated Questions
A	1. What do the authors argue about the reason for zero-shot translation failure in transfer learning? 2. In what scenario, does transfer learning being explored? 3. What new transfer learning approach do authors propose for NMT? 4. What novel pre-training method do the authors propose? 5. How does the authors' proposed method different from traditional transfer methods?	1. Neural Machine Translation has dominated recent research on what? 2. Who proposes a new transfer learning approach for NMT? 3. What does our proposed method belong to? 4. What hinders transfer learning for NMT towards the zero-resource setting? 5. Which model benefits from the availability of the universal encoder?
B	1. What corpus is the model based on? 2. What about the inter-annotator agreement for the human evaluation? 3. What is the size of the corpus for training? 4. Is the model able to predict dogmatism in comments? 5. How are annotation results guaranteed?	1. What does dogmatism influence on social media? 2. What did MS produce? 3. How many Reddit posts are used to train our model? 4. What does a comment not convey? 5. What number indicates agreement indistinguishable from chance?
C	1. How to improve the performance on the child task? 2. Is it common to have source and assisting languages with different word order in a practical setting? 3. How to address the word order divergence in this paper? 4. How to overcome data sparsity between the source and the target languages? 5. What is the assisting language in all experiments?	1. What is the de facto approach for any NLP task? 2. What is the word order of the source language? 3. Which BIBREF takes advantage of lexical similarity between languages? 4. What did BIBREF10 propose to use to represent the sentences in both the parent and the child language? 5. What does the adversarial training try to force the encoder representations of?

Table 6: Human-annotated questions V.S. model-generated questions of three scientific papers from the SciMRC dataset. In general, human-annotated questions show a higher difficulty level.

Initial Summary	**S0:** This paper proposes a new transfer learning approach for NMT that uses cross-lingual pre-training to enable a high performance on zero-shot translation ..... The proposed approach outperforms very strong baseline systems, consistently surpassing pivoting and multilingual approaches. Meanwhile, the performance on supervised translation direction remains the same level or even better when using their method.
Feedback	**Q:** Does the approach proposed by the authors perform better than all the comparison systems? **F:** The proposed approach outperforms very strong baseline systems
Subsequent Summaries	**S1:** The paper proposes an approach to training a Chinese NER system-annotated data using adversarial training ..... the negative influences of different crowd annotators. The proposed model outperforms very strong baseline systems. **S2:** The paper proposes an approach to training a Chinese NER system on crowd-annotated data using adversarial training ..... by reducing the negative influences of different crow annotators. The proposed model outperforms very strong baseline systems. **S3:** The paper proposes an approach to training a Chinese NER system on crow-annotated data using ..... The proposed approach outperforms very strong baseline systems. The resulting features of the common Bi-LSTM are worker-invariant and NER-sensitive.
Initial Summary	**S0:** This paper explores transfer learning in a common zero-shot scenario where there are a lot of source-pivot and pivot-target parallel data but no source-target parallel data. The authors propose a new transfer learning approach for NMT that uses cross-lingual pre-training to enable a high performance on zero-shot translation ..... In addition, the performance on supervised translation direction remains the same level or even better when using their method.
Feedback	**Q:** What new transfer learning approach do authors propose for NMT? **F:** We propose a new transfer learning approach for NMT that uses cross-lingual language model pre-training to enable high performance on zero-shot translation.
Subsequent Summaries	**S1:** The proposed approach uses cross-lingual pre-training to alleviate the distance between different source language spaces and significantly improves zero-shot translation performance, consistently surpassing pivoting and multilingual approaches. **S2:** The paper proposes a new transfer learning approach for NMT that uses cross-lingual language model pre-training to enable high performance on zero-shot translation. The proposed approach uses cross-lingual pre-training to alleviate the distance between different source language..... **S3:** The main contributions of this paper are:* Proposing a new transfer learning approach for NMT that uses cross-lingual language model pre-training to enable high performance on zero-shot translation. ..... Demonstrating that the proposed approach significantly improves zero-shot translation performance, consistently surpassing pivoting and multilingual approaches.

Table 7: Cases of LLMs **memorizing** facts provided in the feedback and appear in subsequent summaries.

Feedback	**Q:** What is the architecture of the framework **F:** In this paper, we design a generation-based model, which is more efficient than the previous approach.
Subsequent Summary	**S:** The paper presents a new dataset for knowledge-based machine reading comprehension, which requires reasoning over triplet-style knowledge involved in a document. The authors find that published datasets do not sufficiently support this task, so they build a new dataset consisting of 40,047 examples for the knowledge-based machine reading comprehension task. They also propose a framework consisting of a question-answering model and a question-generation model, both of which take the knowledge extracted from the document as well as relevant facts from an external knowledge base such as Freebase/ProBase/Reverb/NELL.....The authors also implement two question-answering models, which directly measure the semantic similarity between
Original Paper	**P:** ..... The annotation of this dataset is designed so that successfully answering the questions requires understanding and the knowledge involved in a document. Each instance is composed of a question, a set of triplets derived from a document, and the answer. We implement a framework consisting of both a question-answering model and a question-generation model, both of which take the knowledge extracted from the document as well as relevant facts from an external knowledge base such as Freebase/ProBase/Reverb/NELL .....
Feedback	**Q:** What is the main purpose of collecting the data set? **F:** The main purpose of collecting the data set was to create a knowledge base for a dialog system.
Subsequent Summary	**S:** The paper proposes a dataset for developing and evaluating dialog systems based on knowledge bases. The proposed dataset aims to help improve the accuracy of dialog systems in providing general information from a knowledge base..... The authors hypothesize that users are willing to give such explanations to help find answers for their own questions. The main purpose of the collected dataset is to enable interactive learning using the proposed approach and potentially to evaluate how different systems perform on this task. The paper also distinguishes three different basic approaches to requesting knowledge bases.
Original Paper	**P:** ..... The last source of information that we consider for interactive learning is rephrasing, which could help when the system does know the concept but does not know the correct wording. This area is extensively studied for the purposes of information retrieval BIBREF5, BIBREF6. The main purpose of the collected dataset is to enable interactive learning using the steps proposed above and potentially to evaluate how different systems perform on this task. The obvious difficulty when developing a dialog system is finding a way how to identify the piece of information that the user is interested in .....

Table 8: Cases of LLMs being **capable of rectifying** non-facts provided in the feedback into facts in the subsequent summary. The source paper is presented as a reference to verify the facts.

Feedback	**Q:** What is the limitation of the current state-of-the-art in NMT? **F:** The current state-of-the-art in Neural Machine Translation (NMT) heavily relies on large-scale parallel data, resulting in poor performance on low-resource or zero-resource language pairs.
Subsequent Summary	**S:** The current state-of-the-art in Neural Machine Translation (NMT) heavily relies on large-scale parallel data, resulting in poor performance on low-resource or zero-resource language pairs. One common alternative to avoid pivoting in NMT is transfer learning, which leverages a high-resource pivot-target model to initialize a low-resource source-target model that is further optimized with a small amount of available parallel data.....
Feedback	**Q:** How to create synthesized summaries in this paper? **F:** We propose a template-based synthesis method to create synthesized summaries, then explore the effect of enriching training data for abstractive summarization using the proposed model compared to a synthesis baseline
Subsequent Summary	**S:** 1. We use a simple template-based method to create synthesized summaries. 2. We propose a novel abstractive summarization model based on the encoder-decoder architecture, which achieves state-of-the-art performance. 3. We provide a new summarization dataset for student reflections. 4. We provide a baseline of a seq2seq model. 5. In domains such as student reflections, the lack of sufficient training data may be a problem. 6. We explore domain transfer for abstractive summarization.
Feedback	**Q:** What do they plan to do in the future? **F:** In the future, we plan to do the following: Explore the application of NMT in multilingual learning
Subsequent Summary	**S:** The paper introduces a unified approach to seamlessly extend the original NMT to multilingual scenarios. The proposed framework allows any language to be integrated into any side of the encoder-decoder architecture with only one encoder and one decoder for all the languages involved..... and the results show that bringing multilingualism to NMT helps to improve individual translations. In the future, the authors plan to explore the application of NMT in multilingual learning and evaluate the proposed framework on more languages and more translation scenarios.

Table 9: Cases of LLMs **failing to exclude or rectify** non-facts provided in the feedback in the subsequent summary, given the non-facts also appear in the subsequent summaries.