Journal of Artificial Intelligence and Data Mining (JAIDM), Vol. 13, No. 1, 2025, 11-24.

Shahrood University of Technology

Journal of Artificial Intelligence and Data Mining (JAIDM)
Journal homepage: http://jad.shahroodut.ac.ir

Technical paper

Enhancing the Quality of Scientific Writing Using Advanced Language Models: Automated Evaluation and Proofreading

Amirali Kharazmi and Hamid Hassanpour*

Faculty of Computer Engineering and IT, Shahrood University of Technology, Shahrood, Iran.

Article Info

Article History:
Received 18 December 2024
Revised 08 January 2025
Accepted 04 February 2025

DOI:10.22044/jadm.2025.15482.2661

Keywords:
Artificial Intelligence, Scientific Writing, Natural Language Processing, Language Models, Prompt Optimization, Gradient-based Optimization

**Corresponding author: h.hassanpour@shahroodut.ac.ir (H. Hassanpour).*

Abstract

Advancements in artificial intelligence have produced powerful language models that enhance scientific writing through automated evaluation and proofreading. Effective use of these models relies on prompt engineering—the precise formulation of requests—which directly influences output quality. As the saying goes, "Asking correctly is half of knowledge," emphasizing the importance of well-crafted prompts. In this study, we introduce a novel approach utilizing the simple language model Gemma-7b-it to improve scientific writing. By detailing the specific characteristics and structures of each section of a scientific paper, we prompt the model to evaluate and proofread text for clarity, coherence, and adherence to academic standards. Our method comprises three stages: initial evaluation, feedback-based proofreading, and iterative refinement using textual gradient optimization. Tested on a dataset of 25 scientific articles, expert evaluations confirm that this method achieves significant enhancements in abstract quality. These findings demonstrate that meticulous prompt engineering can enable simpler language models to produce results comparable to advanced models like GPT-4, underscoring the critical role of prompt optimization in achieving high-quality scientific writing.

1. Introduction

The effectiveness of scientific communication hinges on the clarity and expressiveness of written articles. High-quality writing not only makes complex concepts accessible but also bolsters the impact and credibility of research findings. While conventional tools like grammar checkers and style editors have been useful in improving the surface-level aspects of manuscripts, they often fall short in addressing deeper issues related to content coherence, logical flow, and adherence to academic standards [1].

Recent advances in artificial intelligence (AI) and natural language processing (NLP) have introduced powerful language models capable of understanding context and generating human-like text. Advanced models such as GPT-4 [2] harness deep learning architectures, offering significant promise for improving the structure and clarity of scientific writing, hence enhancing the accessibility and impact of research papers [3]. The advanced language models also have contextual text processing capabilities. Izadi and Ghasemzadeh [4] introduced a generalized language model for question matching, illustrating the potential of NLP models in aligning textual content with specific objectives.

Prompt engineering is a critical component in optimizing the performance of AI-based tools for scientific writing. It involves crafting input instructions, or *prompts*, that help language models better comprehend the desired task, leading to more coherent, clear, and relevant generated text. As the saying goes, "Asking correctly is half of knowledge," emphasizing the importance of creating precise and effective prompts to ensure high-quality results. This approach is especially critical in scientific writing, where accuracy and adherence to academic standards are essential.

Despite the enormous potential of these tools, several challenges hinder their widespread adoption. First, the high resource demands of running large language models like GPT-4 can be cost-prohibitive, restricting access to their benefits. Second, prompt engineering itself is a specialized skill that requires domain-specific expertise [5]. For example, in scientific writing, expertise in academic writing is necessary to design prompts that generate coherent and well-structured outputs. Moreover, for a given task, simpler language models like Gemma-7b-it require even more effective and precise prompts to produce high-quality results compared to their more advanced counterparts. This reliance on prompt clarity and specificity makes the role of prompt engineering even more critical in achieving optimal results with resource-efficient models.

In this study, we introduce a novel approach for improving scientific writing using the Gemma-7b-it language model [6], which offers a balance between efficiency and performance. Through careful prompt optimization techniques, we demonstrate that such a simple generative model can also generate high-quality scientific content, emphasizing the crucial role of prompt engineering in maximizing the potential of AI-assisted writing tools. This approach makes sophisticated writing assistance more accessible, even to those with limited computational resources [5].

Gemma-7b-it is a lightweight generative language model optimized for tasks requiring moderate computational resources. Unlike larger models, its architecture focuses on efficiency, balancing model complexity with accessibility. The model processes input prompts by encoding contextual information and generating structured outputs based on its training on diverse datasets. Its simplicity makes it an ideal candidate for tasks where resource constraints are a concern. The model interacts dynamically with prompts, interpreting them as instructions for specific tasks. The effectiveness of this interaction depends heavily on the precision and clarity of the prompts, making prompt engineering a cornerstone of our method. Furthermore, the iterative refinement process systematically enhances the quality of outputs by incorporating gradient-based optimization techniques, ensuring that the text aligns with optimal quality.

2. Related Works

The enhancement of scientific writing has long been a focus of research, beginning with early efforts aimed primarily at improving spelling errors and grammatical accuracy. Tools such as basic spell checkers and grammar editors helped to identify and correct these errors, laying the foundation for automated writing assistance [7, 8]. These early developments were essential in addressing surface-level issues but did not go beyond fixing basic linguistic mistakes.

As research progressed, the focus expanded to more complex aspects of writing, such as sentence structure and coherence. In this regard, rule-based systems were introduced, which analyzed text for logical consistency, syntactic structure, and thematic relevance. These systems employed predefined linguistic rules to offer feedback on structural issues but were limited by their rigidity, which made them less adaptable to the diverse writing styles found in scientific discourse [9]. For instance, Grammatik, developed in the 1980s, was one of the first commercial grammar checkers that utilized a set of linguistic rules to identify and correct grammatical and stylistic errors [10]. While effective at addressing basic grammatical and stylistic errors, the rule-based approach had limitations when dealing with the more dynamic and diverse writing styles commonly found in scientific writing. This rigidity made it difficult for such systems to adapt to the evolving complexities of academic discourse.

The field continued to evolve with the advent of machine learning and deep learning models, which brought significant advancements in the ability to understand and process more intricate textual features. The introduction of language models capable of not only correcting grammatical errors but also enhancing conceptual clarity and logical flow marked a transformative shift. Large-scale models like GPT-3 and GPT-4, trained on massive datasets, have demonstrated impressive abilities to understand context, generate coherent text, and refine scientific writing [11, 12]. While these models are powerful, their high computational demands and associated costs often limit their accessibility, particularly for researchers with limited resources.

In parallel, technologies like automated summarization tools, such as the one introduced in [13], have become essential for condensing lengthy scientific articles into concise, informative abstracts, while keyword extraction systems help identify the most significant terms for efficient indexing and retrieval [13, 14]. These tools have greatly enhanced the accessibility and discoverability of scientific content, making it easier for researchers to engage with relevant literature.

```
   Suppose you are an expert reviewer asked to evaluate and improve
the following {section_name}.
       **Title:**
       {title}
       **{section_name.title()}:**
       {original}
       Please assess the {section_name} based on the following
criteria. For each criterion, provide a brief comment explaining
whether the {section_name} meets the criterion and why.
       **{section_name.title()} Evaluation Criteria:**
       {criteria}
       After evaluating all criteria, please provide:
       - **Strengths:** List the key strengths of the {section_name}.
       - **Weaknesses:** List the main weaknesses of the
{section_name}. - **Suggestions for Improvement:** Offer specific
suggestions to address the weaknesses identified. - **Overall Score:**
Assign an overall qualitative score (e.g., Poor, Fair, Good,
Excellent), considering only the required criteria. If optional
criteria are met,  they can positively influence the score. Wrap the
score in `<OVERALL_SCORE>` tags, like this:
       `<OVERALL_SCORE>Your Qualitative Score</OVERALL_SCORE>`.

       **Your Evaluation:**
```

Figure 1. Evaluator Prompt: A structured prompt template for evaluating a scientific paper section. Placeholders such as {section_name}, {title}, {original}, and {criteria} are replaced with the section details and evaluation criteria. The {section_name.title()} function capitalizes the section name for readability. This template is an f-string, allowing dynamic replacement of placeholders with actual values.

With the sophistication of language models, prompt engineering has emerged as a crucial technique to optimize their performance. By carefully crafting input prompts, researchers can guide models to produce contextually relevant outputs, refining them through iterative adjustments to improve their coherence and alignment with academic standards [16].

Moreover, integrating human expertise into the AI-assisted writing process has become a growing area of focus. Human-in-the-loop (HITL) approaches combine the strengths of AI models with expert feedback, ensuring that generated content meets the rigorous standards of academic writing and addressing nuances that may be overlooked by the model alone [16, 17]. While these approaches are highly effective, they often introduce additional complexity and reliance on expert availability, making simpler solutions more appealing in resource-constrained scenarios.

Recent efforts to personalize AI-assisted writing have focused on adapting models to align with individual author styles [18, 19]. Techniques such as style transfer and personalized prompt tuning enable AI tools to generate text that resonates with the unique voice and tone of the author, improving the overall effectiveness of these tools.

3. Proposed Method

We present a systematic method to enhance the quality of scientific writing by leveraging a simple generative language model, Gemma-7b-it, through effective prompt engineering and iterative refinement. Recently developed language models are responsive to prompts, meaning they can be directed to perform specific tasks through carefully crafted input instructions [21]. Central to our approach is the concept of prompt optimization, which is critical for effectively utilizing language models, especially simpler ones like Gemma-7b-it. Our method automates the evaluation and proofreading of specific sections of a scientific paper, such as the abstract, by guiding the language model to act as an expert reviewer and proofreader. The method consists of several interconnected stages, each designed to progressively improve the text while ensuring alignment with academic standards.

The methodology begins with the initial evaluation of the specific section of the paper to be improved (e.g., abstract, introduction). Recognizing that each section in a scientific paper has unique characteristics and structural requirements [22], we define specific evaluation criteria pertinent to that section.

```
   Suppose you are an expert academic writer asked to improve the
following {section_name} based on the feedback provided.
   **Title:**
   {title}
   **Original {section_name.title()}:**
   {original}
   **Feedback:**
   {feedback}
   **IMPORTANT:**: Please rewrite the {section_name} to address the
weaknesses mentioned in the feedback, enhancing clarity,
completeness, and overall quality. Ensure that the revised
{section_name}:
   - Are concise and clear, containing only essential and useful
information.
   - Include all necessary components as specified in the evaluation
criteria.
   - Avoid unnecessary content such as irrelevant preliminary
concepts, minor details, abbreviations, mathematical formulas, or
references.
   - Are self-contained and understandable to readers outside the
specific field.
   - Reflect a strong connection with the title.
   - Use clear and accessible language, minimizing technical jargon.
   **Rewritten {section_name.title()}:**
   Please provide the rewritten {section_name} enclosed between
`<{section_name.upper()}>` and `</{section_name.upper()}>` tags.
```

Figure 2. Initial Generator Prompt- Structured prompt template for generate rewritten content of scientific paper section. Placeholders such as {section_name}, {title}, {original }, and {feedback} are replaced with the section details and evaluation feedback.

By meticulously crafting these criteria and integrating them into a carefully designed evaluation prompt, we perform prompt optimization to ensure that the language model receives precise and comprehensive instructions. This optimization is crucial because the way we formulate prompts directly influences the quality of the model's output. The prompt instructs the language model, acting as an expert reviewer, to assess the text based on these criteria, providing detailed feedback—including strengths, weaknesses, suggestions for improvement, and an overall qualitative score.

In the first stage, known as the initial evaluation, we obtain feedback from the language model using the evaluator prompt (see Figure 1). This prompt guides the model to act as an expert reviewer, ensuring that the feedback is both comprehensive and directly relevant to the section and criteria in question. This targeted feedback is essential for the subsequent proofreading stage, where the feedback is used to enhance the text according to the model's detailed evaluation.

In the second stage, we proofread the section based on the received feedback using optimized prompts tailored to the iteration stage. For the first iteration, we employ an initial generator prompt (see Figure 2) to enhance clarity, completeness, and overall quality without introducing new content or unnecessary formatting. This stage ensures that the output aligns closely with academic standards and meets the predefined criteria.

The third stage focuses on iterative refinement and concept drift detection to maintain alignment between the rewritten text and the original meaning. Concept drift refers to changes in the meaning or representation of concepts over time, which may occur when the model rewords or restructures content, potentially altering underlying ideas [23]. To prevent this, we use a detector prompt (see Appendix A) to identify significant contradictions and a coherence prompt (see Appendix B) to resolve these issues. The iterative process involves addressing major drift affecting meaning or conclusions first, followed by minor deviations, thereby ensuring that the text retains its

intended meaning while enhancing readability and coherence.

Throughout this iterative process, textual gradient-based optimization plays a central role. Similar to gradient descent in optimization algorithms, where parameters are adjusted iteratively to minimize a loss function, our method progressively improves the text by incorporating the model's feedback at each step. Each iteration refines the text further, effectively moving it closer to the optimal quality along the *textual gradient* defined by the feedback. By using different prompts for the initial and subsequent iterations and incorporating concept drift detection, we effectively address new challenges that arise at each stage, such as resolving concept drift in later iterations.

This systematic method highlights the potential of integrating simple yet efficient language models with expert-guided prompts. As previously mentioned, the advantages of this approach include accessibility, cost-effectiveness, scalability, and reduced computational demands while maintaining high standards of scientific writing. The results emphasize that the careful combination of prompt optimization and iterative refinement can bridge the gap between model simplicity and task complexity, making advanced writing assistance more widely available. Furthermore, the use of gradient-based optimization in this method sets it apart from traditional prompt-based approaches, as it enables continuous improvement of the text quality through iterative adjustments.

4. Experimental Setup

In this study, we evaluated the effectiveness of our proposed method for enhancing scientific writing by applying effective prompt optimization techniques to the abstract section of scientific articles using the Gemma-7b-it language model. The experimental setup was designed to assess the improvements in quality achieved through automated evaluation and proofreading, guided by optimized prompts and expert-defined criteria.

All experiments were conducted on a standard personal computer equipped with an Intel® Core™ i7-13620H 13th Generation processor (2.40 GHz) and 16 GB of RAM (15.7 GB usable), running a 64-bit operating system on an X64-based processor architecture. Gemma-7b-it was accessed via an API provided by groq.com, which offers free access, making it a cost-effective alternative to more advanced models such as GPT-4 that require paid API subscriptions. As an open-source and computationally efficient model, Gemma-7b-it enables AI-assisted scientific writing without requiring high-performance hardware or expensive cloud-based services. This makes the proposed method highly accessible and scalable, allowing researchers with limited computational resources to benefit from advanced language model capabilities on standard consumer hardware. By leveraging an optimized prompt-based approach, our method ensures practical deployment without sacrificing writing quality or adherence to academic standards.

4.1. Dataset Selection

We selected a dataset of 25 published scientific articles from the fields of Electrical Engineering and Computer Science to evaluate the effectiveness of our proposed method. The titles and abstracts of these articles served as input for our algorithm. In selecting these articles, we ensured that the abstracts contained all the main components of a standard abstract, such as objectives, methods, results, and conclusion [21]. This criterion was important to effectively test our method on abstracts that follow standard academic structures. The edited versions of the abstracts, generated using our proposed method, were compared with the original abstracts to evaluate performance. To ensure the reliability and validity of the evaluations, we conducted an expert evaluation involving specialists in the relevant fields of each article. Three experts were recruited for each article, ensuring they had substantial experience in the specific domain and were well-versed in academic writing standards, particularly in recognizing the features of a high-quality abstract.

4.2. Implementation of the Proposed Method

The implementation of our algorithm was carried out in Python, utilizing f-string formatting for dynamic insertion of variables into prompt templates and leveraging Markdown for structuring and formatting the generated content to enhance readability and clarity.

4.2.1. Initial Evaluation

In the initial evaluation stage, the inputs consisted of the title of the article and its abstract. By integrating these elements, we instruct the Gemma-7b-it model to act as an expert reviewer using the template in Figure 1. The model assesses the abstract based on the predefined criteria (detailed in Figure 3), generating comprehensive feedback that includes strengths, weaknesses, suggestions for improvement, and an overall qualitative score. The output of this stage is detailed evaluation feedback from the language model, providing a thorough analysis of the original abstract's quality.

```
1. **General Statement of the Research Area (Required): ** Does the
abstract begin with a general sentence that specifies the research
area and the specific topic under investigation?
2. **Specific Problem Description (Optional): ** Does the abstract
describe the specific problem to be solved, including its challenges
and obstacles?
3. **Review of Existing Solutions (Optional): ** Does the abstract
provide an overview of standard or existing solutions and their
limitations?
4. **Overview of the Proposed Solution (Required): ** Does the
abstract present a general overview of the new proposed solution?
5. **Summary of Evaluation and Results (Required): ** Does the
abstract include a summary of how the proposed solution was evaluated
and the results obtained?
6. **Relation to Title (Required): ** Is the abstract relevant to
the title and does it adequately reflect the content suggested by the
title?
7. **Self-Containment (Required): ** Is the abstract self-
contained, making sense on its own without requiring additional
context?
8. **Clarity and Conciseness (Required): ** Does the abstract avoid
unnecessary content such as preliminary concepts, minor details,
abbreviations, mathematical formulas, or references?
```

Figure 3. Essential components of an abstract in scientific writing [25].

4.2.2. First Proofreading

Using the feedback obtained from the initial evaluation, the second stage involved proofreading the abstract. The inputs for this stage were the title of the article, the original abstract, and the feedback from the initial evaluation. The initial generator prompt (see Figure 2) was employed to guide the language model in proofreading the abstract. This prompt instructed the model to enhance clarity, completeness, and overall quality without introducing new content or unnecessary formatting.

The output was a proofread version of the abstract that improved clarity, completeness, and adherence to the specified evaluation criteria.

4.2.3. Iterative Refinement

The iterative refinement stage aims to further enhance the abstract through continuous improvement. This stage encompasses several sub-stages**:**

- **Concept Drift Detection**

The language model is then asked to compare the proofread abstract with the original one to identify any severe concept drift that might have arisen during proofreading. A concept drift detection prompt (see Appendix A) was crafted to define what constitutes severe concept drift and to instruct the language model to identify and explain any such drift, providing suggestions for resolving them. Initially, the language model was very strict in detecting concept drift, often raising concerns over even minor differences. To address this, we adjusted the detector prompt to focus specifically on severe instances of concept drift, i.e., contradictions. Instead of identifying all types of drift, the model was instructed to concentrate only on significant contradictions that could substantially alter the meaning or conclusions of the abstract.

- **Subsequent Proofreading Based on Concept Drift**

The inputs for this substage includes the title of the article, the original abstract, the previously proofread version, the initial evaluation feedback, and the identified concept drift issues. A coherence prompt (see Appendix B) is utilized to guide the model in resolving these issues. The prompt instructs the model to first address major contradictions affecting the abstract's meaning and conclusions, followed by minor inconsistencies. The output is an improved version of the abstract with resolved concept drift, enhanced readability, and maintained coherence with the original content.

- **Quantitative Evaluation and Process Termination**

After each iteration of proofreading, the proofread abstract served as the input for quantitative evaluation. The language model assigns numerical scores between 0 and 10 to each evaluation criterion outlined in Figure 3. These scores provide an objective measure of the abstract's quality. Each criterion is weighed according to its importance, as detailed in Table 1, to calculate a weighed average score for the abstract. This quantitative assessment allows us to objectively measure improvements based on the relative significance of each criterion.

The iterative refinement process continues until one of the termination conditions is met. The inputs for this combined substage includes the number of iterations completed and the improvement in the average weighed score between consecutive iterations. The process is halted if either the maximum of ten iterations is reached or if the improvement in the average weighed score is more than 0.5. Upon meeting the termination conditions, the abstract version with the highest weighed score is selected as the best proofread version.

Finally, the best proofread abstract, along with its final score and the history of scores from each iteration, is compiled and stored for further analysis and comparison. This comprehensive dataset includes the final proofread abstracts, their scores, and the progression of scores across iterations, enabling detailed evaluation of the method's effectiveness.

Table 1. Criteria with weights for computing weighed average score.

Criterion	type	weight
General Statement of the Research Area	Required	1
Specific Problem Description	Optional	1
Review of Existing Solutions	Optional	1
Overview of the Proposed Solution	Required	3
Summary of Evaluation and Results	Required	3
Relation to Title	Required	4
Self-Containment	Required	2
Clarity and Conciseness	Required	2

4.3. Expert Evaluation

To evaluate the effectiveness of our method, we conducted an expert assessment involving specialists from the relevant fields of each article. The experts were provided with evaluation files designed to ensure an unbiased review, which included the article's title, authors' names, and publication details. For each article, two abstracts were provided: one original and one proofread. The reviewers were not informed which abstract was the proofread version to prevent any bias in their assessment.

Accompanying the abstracts was a table where the experts were asked to provide their evaluations. They rated each abstract qualitatively by selecting one of the options: *Good*, *Fair*, or *Poor*. Additionally, they were asked to assign a quantitative score ranging from 0 to 20. This structured approach ensured that the assessments were both comprehensive and objective, providing valuable insights into the effectiveness of our method in enhancing the quality of scientific abstracts.

5. Results

This section presents the results of the evaluation, including quantitative improvements, qualitative assessments, inter-rater reliability analysis using Fleiss' Kappa [25], which is a statistical measure to assess the reliability or agreement between multiple raters. It is particularly useful for evaluating the consistency of categorical ratings made by more than two raters. The Kappa value ranges from -1 (indicating complete disagreement) to 1 (indicating perfect agreement), with 0 suggesting no agreement beyond chance. To calculate Fleiss' Kappa, the degree of agreement between raters is first measured, and then adjusted for the expected agreement by chance. This adjustment helps to determine the true level of consistency among the raters.

The average quantitative scores assigned by the experts for both the original and proofread abstracts (each pair of abstracts was reviewed by 3 experts) were summarized in Table 2. The *Score Improvement* column indicates the difference between the proofread and original abstracts' scores. The significant increase in scores for all articles demonstrates the effectiveness of the proposed method in enhancing the quality of abstracts.

A paired t-test was conducted to evaluate whether the improvements in quantitative scores were statistically significant. The test was performed using the *scipy.stats.ttest_rel* function from the *scipy* module, which is commonly used for statistical hypothesis testing in *Python*. The null hypothesis (H_0) assumed that there was no significant difference between the original and proofread abstracts. The test was conducted with a significance level of $\alpha = 0.05$.

The results showed a statistically significant improvement ($t = 11.42$, $p = 3.45e\text{-}11$), indicating that the probability of observing such improvements purely by chance is less than 1%. Since $p < 0.01$, the null hypothesis was rejected, confirming that the proofread abstracts were rated

significantly higher on average compared to the original abstracts (alternative hypothesis).

To assess the consistency among the experts' qualitative evaluations, we calculated Fleiss' Kappa for both the original and proofread abstracts.

Table 2. Average Quantitative Scores for Original and proofread Abstracts.

Article No.	Original Abstract Score (Avg)	Proofread Abstract Score (Avg)	Score Improvement
1	13.66	17	**+3.33**
2	16.33	17.33	**+1**
3	15.33	18.33	**+3**
4	14.33	16.66	**+2.33**
5	15.66	17.66	**+2**
6	15.33	18.33	**+3**
7	14	17	**+3**
8	17	18.33	**+1.33**
9	17	18.33	**+1.33**
10	15.66	17.33	**+1.66**
11	14.33	16.66	**+2.33**
12	17.33	18.66	**+1.33**
13	12.66	17	**+4.33**
14	14	17.66	**+3.66**
15	10.66	17	**+6.33**
16	14	17.33	**+3.33**
17	12	16	**+4**
18	14	17.66	**+3.66**
19	13.33	18.33	**+5**
20	14	18	**+4**
21	14.33	16.33	**+2**
22	15.33	17.66	**+2.33**
23	16	18	**+2**
24	13.33	17.66	**+4.33**
25	15.33	18.66	**+3.33**

For the original abstracts, Fleiss' Kappa was found to be 0.2643, indicating fair agreement among the experts. In contrast, the proofread abstracts achieved a Fleiss' Kappa of 0.6278, indicating substantial to almost perfect agreement. The fair agreement observed for the original abstracts suggests that there was variability in how experts perceived the quality of these abstracts. This variability highlights inconsistencies in the initial writing quality and the subjective nature of qualitative assessments. On the other hand, the higher Kappa value for the proofread abstracts indicates that our method significantly enhanced the consistency of expert evaluations. The increased agreement among experts for the proofread abstracts reflects the effectiveness of our approach in producing consistently high-quality abstracts, thereby demonstrating the reliability of our method in improving scientific writing.

5.1. Discussion of Results

The results indicate that our method effectively enhances the quality of scientific abstracts. Consistent improvements were observed across both quantitative and qualitative evaluations, demonstrating that even with a simpler language model like Gemma-7b-it, significant enhancements are achievable through meticulous prompt optimization and iterative refinement. By emphasizing prompt optimization and utilizing textual gradient-based refinement, our approach underscores the importance of well-designed prompts in harnessing the full potential of language models. This not only simplifies the writing process but also improves clarity, coherence, and adherence to academic standards, providing a systematic framework for enhancing scientific writing as a whole.

The alignment between expert evaluations and the quantitative scores obtained during the iterative process reinforces the validity of using the language model's assessments to guide the proofreading process. This correlation suggests that the model's feedback effectively mirrors human judgment, ensuring that the improvements are both meaningful and relevant.

The results also highlight that simpler model, such as Gemma-7b-it, require more carefully crafted and optimized prompts compared to more advanced models like GPT-4. As previously mentioned, this dependency arises from their limited contextual understanding, which can be effectively compensated for through meticulous prompt engineering. Despite this, our study demonstrates that when guided by well-designed prompts, Gemma-7b-it can produce results of comparable quality to advanced models. This insight emphasizes the efficiency and scalability of the proposed method, even with resource-efficient models. However, while these observations are promising, direct experimental comparisons with more advanced models like GPT-4 remain beyond

the scope of this study, suggesting an important direction for future research.

Additionally, the analysis of abstracts across various topics revealed varying degrees of improvement. While most abstracts showed significant enhancements, some exhibited less dramatic improvements due to the inherent quality of the original texts. This observation highlights the need for refining prompt strategies to better address cases where initial quality is already high, ensuring that the method remains universally effective.

By systematically refining prompts and employing an iterative process, our method demonstrates that resource-efficient models can still meet high academic standards. This further validates the role of prompt engineering as a critical component in maximizing the utility of language models. Moreover, the integration of gradient-based optimization distinguishes our approach, enabling continuous and targeted improvements to enhance the quality of scientific abstracts.

6. Challenges, Limitations, and Future Work

6.1. Challenges

One of the primary challenges in utilizing language models like Gemma-7b-it is the dependency on prompt engineering for achieving optimal results. While the proposed method facilitates significant improvements, it demands expertise in designing precise and effective prompts tailored to specific sections of scientific writing. Ensuring the consistency of rewritten content with the original meaning poses another challenge, particularly when addressing concept drift during iterative refinements. Additionally, balancing the simplicity of the model with its scalability for more complex tasks remains a challenge, as Gemma-7b-it, being a lightweight model, may not fully capture the nuances handled by more sophisticated language models.

6.2. Limitations

The main limitation of the study lies in its focus on English-language scientific abstracts. While the method is adaptable to other languages, it requires modifications to the prompts and language models trained for those languages. Furthermore, the lack of direct comparisons with advanced models like GPT-4 limits the scope of the study. While the results suggest that Gemma-7b-it achieves comparable quality when supported by effective prompt engineering, empirical comparisons with more advanced models remain necessary to validate this claim. The resource-intensive nature of expert evaluations further restricted the ability to perform broader comparative analyses. Finally, the subjective nature of expert assessments introduces variability in qualitative evaluations, underscoring the need for standardized metrics to ensure consistency and reliability.

6.3. Future Work

Future research will focus on several key areas to expand and refine the proposed method. First, the methodology will be extended to other sections of scientific papers, including introductions, methods, results, and discussions, to broaden its applicability. Second, efforts will be directed toward adapting the system for multilingual support, enabling compatibility with diverse languages and academic styles. Third, user feedback will be integrated into the prompt optimization process, allowing for real-time adjustments and greater personalization to align outputs with individual author preferences. Additionally, future studies will include direct experimental comparisons between Gemma-7b-it and advanced models like GPT-4 to empirically validate the relative performance of these models. Ethical considerations, such as maintaining originality and avoiding over-reliance on AI tools, will also be explored to ensure responsible integration of AI in scientific writing. Furthermore, given that every author has their own unique writing style, future efforts will aim to develop techniques that enable the system to tailor text outputs to match an individual author's style. This personalization will enhance the practicality and effectiveness of the tool, making it more versatile and user-friendly. By addressing these areas, the proposed method can become more robust, scalable, and accessible, significantly advancing the field of academic writing assistance.

7. Conclusion

This study demonstrated the potential of using a lightweight language model, such as Gemma-7b-it, to enhance scientific writing through prompt engineering. While our experiments focused on the abstract section of scientific papers, the proposed method is applicable to all sections, including introductions, methods, results, and discussions. The approach led to measurable improvements in the clarity, coherence, and overall quality of the proofread abstracts, highlighting the critical role of well-designed prompts in achieving high-quality outputs, even with simpler models.

Despite challenges such as the computational demands of larger models, subjective evaluation criteria, and the complexities of contradiction detection, this method shows significant promise for improving the quality of scientific

communication across all sections of a paper. Furthermore, the findings demonstrate that prompt engineering can unlock the full potential of resource-efficient models like Gemma-7b-it, offering a cost-effective and accessible solution for enhancing academic writing.

Ultimately, this study reaffirms that AI-assisted writing tools, when used appropriately, can complement creativity and critical thinking in scientific writing. By fostering clarity, accessibility, and impact, these tools can significantly enhance the quality of academic communication.

References

[1] K. Hyland, *Second Language Writing*. Cambridge University Press, 2019.

[2] OpenAI et al., "GPT-4 Technical Report," *arXiv preprint*, arXiv:2303.08774, Mar. 2023. [Online]. Available: http://arxiv.org/abs/2303.08774

[3] A. Castellanos-Gomez, "Good Practices for Scientific Article Writing with ChatGPT and Other Artificial Intelligence Language Models," *Nanomanufacturing*, vol. 3, no. 2, pp. 135–138, 2023.

[4] S. Izadi and M. Ghasemzadeh, "Use of Generalized Language Model for Question Matching," *International Journal of Engineering*, 2013.

[5] P. Liu, W. Yuan, J. Fu, Z. Jiang, H. Hayashi, and G. Neubig, "Pre-train, Prompt, and Predict: A Systematic Survey of Prompting Methods in Natural Language Processing," **ACM Computing Surveys*, vol. 55, no. 9, pp. 1–35, 2023.

[6] G. Team et al., "Gemma: Open Models Based on Gemini Research and Technology," *arXiv preprint*, arXiv:2403.08295, 2024.

[7] M. Lytvyn, A. Shevchenko, and D. Lider, "Grammarly Inc.," 2023. [Online]. Available: https://www.grammarly.com/.

[8] C. Banks, "ProWritingAid Ltd.," 2023. [Online]. Available: https://prowritingaid.com/.

[9] G. Heidorn, "Intelligent Writing Assistance," in *A Handbook of Natural Language Processing: Techniques and Applications for the Processing of Language as Text*, vol. 8, Marcel Dekker, 2000.

[10] E. S. Atwell and S. Elliot, "Dealing with Ill-formed English Text," in *The Computational Analysis of English: A Corpus-Based Approach*, Longman, 1987, pp. 120–138.

[11] A. Radford et al., "Language Models Are Unsupervised Multitask Learners," *OpenAI Blog*, vol. 1, no. 8, pp. 1–10, 2019.

[12] T. B. Brown et al., "Language Models Are Few-shot Learners," *arXiv preprint*, arXiv:2005.14165, 2020.

[13] A. Cohan et al., "A Discourse-aware Attention Model for Abstractive Summarization of Long Documents," *arXiv preprint*, arXiv:1804.05685, 2018.

[14] S. Rose, D. Engel, N. Cramer, and W. Cowley, "Automatic Keyword Extraction from Individual Documents," in *Text Mining: Applications and Theory*, 2010, pp. 1–20.

[15] C. Florescu and C. Caragea, "PositionRank: An Unsupervised Approach to Keyphrase Extraction from Scholarly Documents," in *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Long Papers)*, 2017, pp. 1105–1115.

[16] J. D. Velásquez-Henao, C. J. Franco-Cardona, and L. Cadavid-Higuita, "Prompt Engineering: A Methodology for Optimizing Interactions with AI-Language Models in the Field of Engineering," *Dyna (Medellín)*, vol. 90, no. 230, pp. 9–17, 2023.

[17] T. Luther, J. Kimmerle, and U. Cress, "Teaming Up with an AI: Exploring Human–AI Collaboration in a Writing Scenario with ChatGPT," *AI*, vol. 5, no. 3, pp. 1357–1376, 2024.

[18] P. Fernandes et al., "Bridging the Gap: A Survey on Integrating (Human) Feedback for Natural Language Generation," *Transactions of the Association for Computational Linguistics*, vol. 11, pp. 1643–1668, 2023.

[19] Z. Hu, Z. Yang, X. Liang, R. Salakhutdinov, and E. P. Xing, "Toward Controlled Generation of Text," in *Proceedings of the International Conference on Machine Learning*, 2017, pp. 1587–1596.

[20] N. S. Keskar, B. McCann, L. R. Varshney, C. Xiong, and R. Socher, "CTRL: A Conditional Transformer Language Model for Controllable Generation," *arXiv preprint*, arXiv:1909.05858, 2019.

[21] M. Khalifa and M. Albadawy, "Using Artificial Intelligence in Academic Writing and Research: An Essential Productivity Tool," *Computer Methods and Programs in Biomedicine Update*, p. 100145, 2024.

[22] J. M. Swales and C. B. Feak, *Academic Writing for Graduate Students: Essential Tasks and Skills*, 3rd ed., University of Michigan Press, 2004.

[23] J. Gama, I. Žliobaitė, A. Bifet, M. Pechenizkiy, and A. Bouchachia, "A Survey on Concept Drift Adaptation," *ACM Computing Surveys*, vol. 46, no. 4, pp. 1–37, 2014.

[24] J. Cohen, "A Coefficient of Agreement for Nominal Scales," *Educational and Psychological Measurement*, vol. 20, no. 1, pp. 37–46, 1960.

[25] J. Zobel, *Writing for Computer Science*, 2nd ed., Springer, 1997.

Appendix

This appendix provides the full text of the prompt templates used for concept drift detection and coherence improvement in the iterative refinement stage of our method. The prompt templates are generated using Python's f-string formatting, which allows us to dynamically insert variable values at runtime. The variables used include *{section_name}, {section_name.title()}, {title}, {original}, {feedback}, {contradictions}, {rewritten},* and *{rewritten_sentences_text},* each replaced by the corresponding information such as the section's name, article title, original text, feedback, identified contradictions, the most recent rewritten version, and the sentence-by-sentence rewritten text.

Appendix A: Detector Prompt

```
Suppose You are a reviewer. Your task is to carefully compare the rewritten
{section_name} with the original {section_name} and identify contradictions.

### What is a Contradiction?
A contradiction occurs when:
1. The rewritten text **disagrees with**, **removes**, or **changes** the meaning,
conclusions, or tone of the original text.
2. Key details or ideas are **removed**, **added**, or **misrepresented**, leading
to a misunderstanding or misinterpretation of the original meaning.
3. The rewritten text is **misleading**, creating a false impression or
oversimplified interpretation of the original.

### What is NOT a Contradiction?
The following are acceptable changes:
- Simplifying or rephrasing sentences while preserving the meaning, conclusions,
and tone of the original.
- Adding minor details or clarifications that align with and enhance the original
ideas.
- Improving readability or flow without altering the intent, conclusions, or
emphasis of the original text.

### Your Task:
1. Compare each rewritten sentence with the original {section_name}.
2. Identify contradictions by determining if any rewritten sentence conflicts with
or misrepresents the original.
3. For each contradiction, provide:
   - **Contradictory Content:** Describe the part of the rewritten {section_name}
that introduces a contradiction, without referencing or quoting specific
sentences.
   - **Reason:** Clearly explain why this part conflicts with the original
{section_name}, describing the nature of the issue (e.g., omission of critical
details, misrepresentation of conclusions, or changes in tone or focus).
   - **Suggestion:** Provide a detailed and actionable suggestion to address the
contradiction, focusing on how to resolve the issue while maintaining alignment
with the original intent.

4. Ensure your analysis covers both:
   - **Major contradictions:** Meaningful shifts in conclusions, intent, or
emphasis.
   - **Minor contradictions:** Removal of small but important details or
oversimplifications that may mislead the reader.

5. If no contradictions are found, respond with: "NO contradictions found."
```

```
---

### Input Information:
**Title:** {title}

**Original {section_name.title()}:**
{original}

**Sentences of rewritten {section_name.title()}:**
{rewritten_sentences_text}

---

### Example Response Format:

**1. Contradictory Content:**
The rewritten {section_name} shifts the focus from a balanced discussion of trade-offs to a one-sided emphasis on one aspect of the original findings.

**Reason:**
This creates a conflict with the original {section_name}, which presented a balanced perspective by including both the strengths and limitations of the methods under discussion. The rewritten content omits key details that are necessary for understanding the original intent.

**Suggestion:**
Incorporate the omitted details to restore the original balance. For example, ensure both the strengths and limitations of the discussed methods are presented in the rewritten content.

---

If multiple contradictions exist, continue numbering for each issue (e.g., **2. Contradictory Content**, **3. Contradictory Content**, etc.).

If no contradictions are found, respond with:
**"NO contradictions found."**
```

Appendix B: Coherence Prompt

```
Suppose You are an expert academic writer tasked with revising and improving the following {section_name} based on identified contradictions and feedback. Your goal is to enhance clarity, resolve inconsistencies, and maintain alignment with the original meaning.

**Title:**

{title}

**Original {section_name.title()}:**

{original}

**Rewritten {section_name.title()}:**

{rewritten}

**Feedback Provided:**
```

```
{feedback}
**Contradictions Identified:**
{contradictions}
**Your Task:**
Please rewrite the rewritten {section_name} to:
1. **Resolve Major Contradictions First:** focus on resolving critical issues that impact the meaning, conclusions, or findings. Then address minor contradictions. If no contradictions are identified, proceed with improving clarity, consistency, and readability.
2. **Preserve Nuance and Detail:** Ensure all key findings and nuances from the original abstract are retained. Avoid oversimplification.
3. **Ensure Readability:** Improve sentence clarity and flow where necessary, without losing technical accuracy.
4. **Preserve original meaning and focus (IMPORTANT):** Avoid introducing new ideas, claims, or interpretations that were not present in the original {section_name}.
5. **Enhance consistency and coherence:** Ensure all elements are logically connected and free of ambiguity.
6. **Simplify complex phrasing:** Rewrite overly complicated sentences for improved readability, while keeping their intended meaning intact.
7. **Retain core findings:** Make sure the rewritten content preserves the original's key results, implications, and contributions.
8. **Avoid unnecessary formatting (IMPORTANT):** Do not use markdown formatting (e.g., bold or italic) for emphasis. Do not include **keywords** or **sections not present** in the original {section_name}.

**Formatting Guidelines for Output:**
- Enclose the new rewritten {section_name} between `<{section_name.upper()}>` and `</{section_name.upper()}>` tags.
- Ensure the output is concise, precise, and professionally written.

**Example of Reasoning (if needed):**
```

```
- If a sentence is flagged as contradictory, explain why the contradiction occurs and how the revised sentence resolves it without losing the original meaning.
- If no contradictions are identified, explain how the rewritten content improves clarity, flow, and alignment with the original meaning.

**New rewritten {section_name.title()}:**
```

ارتقای کیفیت نگارش علمی با استفاده از مدل‌های زبانی پیشرفته: ارزیابی و ویرایش خودکار

امیرعلی خوارزمی و حمید حسن‌پور*

دانشکده مهندسی کامپیوتر و فناوری اطلاعات، دانشگاه صنعتی شاهرود، شاهرود، ایران.

ارسال ۲۰۲۴/۱۲/۱۸؛ بازنگری ۲۰۲۵/۰۱/۰۸؛ پذیرش ۲۰۲۵/۰۲/۰۴

چکیده:

پیشرفت‌های هوش مصنوعی منجر به توسعه مدل‌های زبانی قدرتمندی شده است که ارزیابی و ویرایش متون علمی را به‌صورت خودکار تسهیل می‌کنند. استفاده مؤثر از این مدل‌ها وابسته به مهندسی درخواست (Prompt Engineering) است— به معنای تنظیم دقیق درخواست‌ها— که به‌طور مستقیم بر کیفیت خروجی تأثیر می‌گذارد. همان‌طور که مصطلح است: «درست پرسیدن، نصف علم است» که بیانگر اهمیت پرسش‌های دقیق و منا سب ا ست. در این مطالعه، رویکردی نوین معرفی می‌کنیم که از مدل زبانی ساده Gemma-7b-it برای بهبود نگارش علمی بهره می‌برد. با تو صیف ویژگی‌ها و ساختارهای خاص هر بخش از یک مقاله علمی، مدل را برای ارزیابی و ویرایش متن از نظر وضوح، انسجام، و رعایت استانداردهای علمی هدایت می‌کنیم. روش ما شامل سه مرحله ا ست: ارزیابی اولیه، ویرایش مبتنی بر بازخورد، و بهینه سازی تدریجی متن با ا ستفاده از گرادیان متنی. این روش بر روی مجموعه‌ای از ۲۵ مقاله علمی آزمایش شده و ارزیابی‌های کارشناسان تأیید کرده‌اند که بهبودهای قابل‌توجهی در کیفیت چکیده ایجاد شده ا ست. این یافته‌ها نشان می‌دهند که مهندسی دقیق درخواست‌ها می‌تواند به مدل‌های زبانی ساده‌تر کمک کند تا نتایجی قابل‌مقایسه با مدل‌های پیشرفته‌ای مانند GPT-4 تولید کنند و بر نقش حیاتی بهینه‌سازی درخواست‌ها در دستیابی به نگارش علمی باکیفیت تأکید می‌کند.

کلمات کلیدی: هوش مصنوعی، نگارش علمی، پردازش زبان طبیعی، مدل‌های زبانی، بهینه‌سازی درخواست، بهینه‌سازی مبتنی بر گرادیان.