IARJSET ISSN (O) 2393-8021, ISSN (P) 2394-1588

International Advanced Research Journal in Science, Engineering and Technology

Impact Factor 8.066 ⁜ Peer-reviewed / Refereed journal ⁜ Vol. 11, Issue 12, December 2024

DOI: 10.17148/IARJSET.2024.111208

Automated Review of Research Papers with Large Language Models

Shaik Suhail[1], Subapriya V[2]

Student, Department of Artificial Intelligence and Machine Learning, Sathyabama Institute of Science and Technology, Chennai, India[1]

Assistant Professor, Department of Computer Science, Sathyabama Institute of Science and Technology, Chennai, India[2]

Abstract: Large language models (LLMs) like GPT-4 have shown potential in generating scientific feedback on research papers, but their effectiveness is limited by vagueness, a lack of domain-specific insights, and insufficient technical critiques, especially regarding model architecture and design. This study aims to address these limitations and enhance the LLM-based feedback system for research papers. The identified gaps include the need for specific, actionable feedback and domain-specific expertise. Our multi-faceted approach includes fine-tuning LLMs with domain specific datasets, incorporating expert-driven feedback, and focusing on detailed, section-specific comments. We also introduce specificity metrics, hybrid models combining LLM and human reviews, and iterative feedback mechanisms. These strategies aim to enhance the quality and utility of LLM-generated feedback, making it more actionable and aligned with expert human reviews. The proposed improvements could significantly reduce the number of review cycles before publication, providing timely and relevant feedback to authors. This research fills critical gaps in existing feedback systems, offering a robust solution to improve the academic review process.

Index Terms: Large Language Models (LLMs), GPT-4, Scientific Feedback, Domain-Specific Datasets, Expert-Driven Feedback, Section-Specific Comments, Specificity Metrics, Hybrid Models, Academic Review Process.

I. INTRODUCTION

The peer review plays a critical role in maintaining the Caliber and integrity of scientific research, yet it often encounters challenges such as vague feedback, insufficient domain-specific insights, and inadequate technical critiques, especially concerning aspects like model architecture and design. Large Language Models (LLMs) like GPT-4 hold potential to address some of these challenges by generating feedback on research papers. However, their effectiveness is limited by generic and superficial feedback. This study aims to enhance LLM-based feedback systems by tackling these limitations. We propose a multi-faceted approach to improve the specificity and quality of feedback. This involves fine-tuning LLMs with domain-specific datasets to ensure that feedback is more relevant and detailed. Additionally, expert-driven insights are integrated to align the feedback with current best practices and standards in the field. The approach focuses on generating detailed, section-specific comments to provide more actionable and precise feedback for each part of a research paper. To support these improvements, we introduce various innovations such as specificity metrics to evaluate the relevance and detail of feedback, hybrid models that combine LLM-generated responses with human reviews to enhance overall quality, and iterative feedback mechanisms to refine feedback through multiple review cycles. Overall, the goal of this study is to offer a robust solution that reduces the quantity of review cycles required prior to publishing and enhances the academic review process. By providing more actionable, expert-aligned feedback, the system aims to streamline the peer review process, benefiting both authors and reviewers and contributing to a more efficient and effective scientific discourse.

II. TECHNOLOGIES FOR ENHANCING LLM-GENERATED FEEDBACK: AN EXTENSIVE OVERVIEW

A. *Fine-Tuning LLM's*

Large language models (LLMs) such as GPT-4 can be fine-tuned by reducing the Cross-Entropy Loss, which allows the model to perform better on particular tasks. This loss function measures the discrepancy between the predicted probabilities for each token by the model and the actual tokens in the training data. Through the use of task-specific data to refine the model, it becomes more adept at generating relevant and accurate responses, improving its performance on specialized tasks such as scientific feedback generation.

IARJSET ISSN (O) 2393-8021, ISSN (P) 2394-1588

International Advanced Research Journal in Science, Engineering and Technology

Impact Factor 8.066 ⁜ Peer-reviewed / Refereed journal ⁜ Vol. 11, Issue 12, December 2024

DOI: 10.17148/IARJSET.2024.111208

$$L_{CE} = -\sum_{i=1}^{N}\sum_{j=1}^{V} y_{ij}\log(\widehat{y_{ij}})$$

B. Semantic Text Matching

Semantic text matching uses cosine similarity to evaluate how similar two text embeddings are in terms of meaning. By converting texts into vector representations and calculating the angle's cosine between these vectors, cosine similarity quantifies their semantic closeness. This method is crucial for tasks such as information retrieval and text classification, ensuring that generated feedback aligns well with the intended context and meaning of the input text.

$$CosineSimilarity(A,B) = \frac{A \cdot B}{|A||B|}$$

C. Text Summarization

Text summaries are judged on their quality by comparing their ROUGE scores to reference summaries. ROUGE determines how many n-grams overlap between the reference and generated summaries to determine how well the summary conveys the most important information. High ROUGE scores signify that the summary effectively covers key points, making it a valuable metric for evaluating the accuracy and relevance of summarization in generating concise and comprehensive feedback.

$$ROUGE - N = \frac{\sum_{\text{gram}\in\text{RefSums}} \text{Count}_{\text{match}}(\text{gram})}{\sum_{\text{gram}\in\text{RefSums}} \text{Count}(\text{gram})}$$

D. Specificity Metrics

The F1 score, recall, and precision are important metrics for evaluating feedback specificity. While precision quantifies the accuracy of the model's positive predictions, recall assesses the model's ability to identify all relevant events. The F1 score provides a fair assessment of performance by integrating both metrics. These metrics ensure that the feedback generated is both accurate and complete, enhancing the model's ability to provide detailed and relevant responses.

$$Precision = \frac{TP}{TP + FP}$$

$$Recall = \frac{TP}{TP + FN}$$

$$F1 = \frac{2 \cdot \text{Precision} \cdot \text{Recall}}{\text{Precision} + \text{Recall}}$$

E. Hybrid Models

Hybrid models combine feedback from LLMs and human experts using weighted averages to enhance feedback quality. By integrating LLM-generated feedback with human insights, these models leverage the strengths of both sources, balancing efficiency with nuanced understanding. The weighted combination (where $\alpha+\beta=1$) ensures that the final feedback benefits from the automated model's scalability and the human expert's contextual expertise, resulting in more reliable and contextually accurate responses.

$$\text{HybridFeedback} = \alpha \cdot \text{LLM Feedback} + \beta \cdot \text{Human Feedback}$$

F. Transfer Learning

Transfer learning adapts pre-trained models to new tasks by fine-tuning them with task-specific data. This approach reuses knowledge from broad training datasets and applies it to specialized tasks, maintaining the same loss function used in the initial training. Transfer learning enhances model performance on specific tasks with less data, making it an efficient method for generating tailored feedback based on previously acquired knowledge.

$$L_{transfer} = -\sum_{i=1}^{N}\sum_{j=1}^{V} y_{ij}\log(\widehat{y_{ij}})$$

IARJSET ISSN (O) 2393-8021, ISSN (P) 2394-1588

International Advanced Research Journal in Science, Engineering and Technology

Impact Factor 8.066 ⁜ Peer-reviewed / Refereed journal ⁜ Vol. 11, Issue 12, December 2024

DOI: 10.17148/IARJSET.2024.111208

III. STUDY OF RELATED WORK

The peer review process is crucial for maintaining the integrity and quality of scientific research, yet it faces several challenges that hinder its effectiveness. The existing literature provides valuable insights into these challenges and proposes various solutions. Liang et al. [1] offers a comprehensive survey on the peer review process, highlighting common issues such as vague feedback and inefficiencies. While this paper provides a broad overview, it lacks specific solutions for improving feedback precision and actionability, which are central to your proposed system. Liu and Shah [2] discuss the application of AI to enhance feedback quality in peer reviews. It introduces the idea of using machine learning to improve feedback but does not address the need for domain-specific models or hybrid approaches. Your research extends this work by integrating domain-specific datasets and expert-driven insights, thereby enhancing the relevance and detail of the feedback generated.

Robertson [3] delves into fine-tuning large language models (LLMs) for domain-specific applications, providing foundational methods for improving model performance in specific fields. This paper directly supports your approach to fine-tuning LLMs with domain-specific datasets but does not cover the use of hybrid models or iterative feedback mechanisms, which are crucial aspects of your proposed system. Peter et al. [4] focuses on evaluating feedback specificity in academic review systems, which aligns with your goal of improving feedback detail. It emphasizes the importance of specificity metrics, a concept that you build upon by developing new metrics and implementing them within a hybrid feedback model.

Zhou et al. [5] explores hybrid models that combine AI and human input to enhance review processes. This paper provides a theoretical basis for your hybrid model approach, but your research advances this idea by incorporating iterative feedback mechanisms and focusing on domain-specific fine-tuning. Finally, Aczel et al. [6] examines strategies for reducing review cycles through automated feedback systems. It complements your objective of accelerating publication by providing timely feedback. However, your proposed system goes further by combining automated feedback with expert reviews and introducing iterative feedback mechanisms to further streamline the review process.

IV. ADDITIONAL STUDIES AND TECHNIQUES

To advance LLM-based feedback systems, several techniques offer valuable enhancements. Transfer learning enables the adaptation of pre-trained models to specialized domains, improving their performance and relevance for domain-specific feedback. Active learning further refines feedback quality by focusing on ambiguous cases and incorporating expert input, which enhances the iterative feedback process. Model interpretability techniques make AI models more transparent, thus improving the clarity and reliability of feedback, making it actionable and understandable. Human-AI collaboration emphasizes the integration of human expertise with AI capabilities, resulting in more nuanced and contextually relevant feedback. Iterative refinement involves continuously updating models based on multiple review cycles, which enhances feedback quality and specificity. Lastly, multi-modal feedback systems suggest that integrating text, audio, and visual data can enrich the feedback process, improving user interaction and accessibility. These techniques collectively contribute to a more effective and insightful LLM-based feedback system.

TABLE I
COMPREHENSIVE LITERATURE REVIEW OF RELATED WORK

Reference	Title	Technique	Remarks
[1]	Can large language models provide useful feedback on research papers? A large-scale empirical analysis	Empirical analysis of LLM-generated feedback	Demonstrates potential of LLMs in providing feedback but highlights limitations in specificity and domain-specific insights.
[2]	Reviewergpt? An exploratory study on using large language models for paper reviewing.	Exploratory study of LLMs for paper reviewing	Explores the feasibility of LLMs in the review process but notes the need for improvement in actionable and detailed feedback.

IARJSET ISSN (O) 2393-8021, ISSN (P) 2394-1588

International Advanced Research Journal in Science, Engineering and Technology

Impact Factor 8.066 ⁜ Peer-reviewed / Refereed journal ⁜ Vol. 11, Issue 12, December 2024

DOI: 10.17148/IARJSET.2024.111208

[3]	GPT4 is Slightly Helpful for Peer-Review Assistance: A Pilot Study	Pilot study on using GPT-4 for peer review assistance	GPT-4 can assist in reviews but requires fine-tuning and expert feedback for better results.
[4]	Blockchain-based paper review system.	Implementation of a blockchain-based review system	Focuses on using blockchain to enhance transparency and security in the review process but does not address improving feedback quality and specificity.
[5]	Blockchain-based file-sharing system for academic paper review.	Blockchain-based file-sharing system for paper review	Discusses secure file sharing in peer review using blockchain, focusing on security and integrity over feedback quality.
[6]	A billion-dollar donation: estimating the cost of researchers' time spent on peer review	Estimation of time and cost of peer review process	Highlights the time and cost of peer review, suggesting a need for efficiency but lacks specific feedback improvement techniques.
[7]	A large annotated corpus for learning natural language inference	Development of a large-scale NLI dataset	Introduces the SNLI corpus, advancing NLI research with a large, high-quality dataset, but highlights coreference indeterminacy and instruction complexity challenges.
[8]	Fighting reviewer fatigue or amplifying bias? Considerations and recommendations for use of ChatGPT and other large language models in scholarly peer review	Exploration of LLMs in peer review	Evaluates the potential and challenges of using LLMs in peer review, noting benefits in efficiency but raising concerns about biases and data confidentiality.

V. PROPOSED SYSTEM AND RESULTS

Our proposed system seeks to enhance the effectiveness of large language models (LLMs) in generating scientific feedback for research papers by addressing key limitations identified in existing studies. The focus is on delivering precise, actionable, and domain-specific feedback, which is critical for improving the quality and relevance of the review process.

The process begins with the raw pdf stage, where users upload their scientific papers. This raw document is then processed through the Parse module. Here, key elements such as the title, abstract, introduction, figures, tables, and captions are extracted. This extraction ensures that all relevant sections of the paper are identified and organized, forming the parsed pdf. This structured data is crucial for creating a coherent and well-defined prompt.

In the Prompt Generation phase, the extracted information from the parsed pdf is used to craft a detailed prompt for the GPT-4 model. This prompt includes specific instructions and content from the paper, designed to elicit meaningful and targeted feedback. The goal is to bridge the gap between raw content and the LLM, ensuring that the feedback is both relevant and comprehensive.

IARJSET ISSN (O) 2393-8021, ISSN (P) 2394-1588

International Advanced Research Journal in Science, Engineering and Technology

Impact Factor 8.066 ⁜ Peer-reviewed / Refereed journal ⁜ Vol. 11, Issue 12, December 2024

DOI: 10.17148/IARJSET.2024.111208

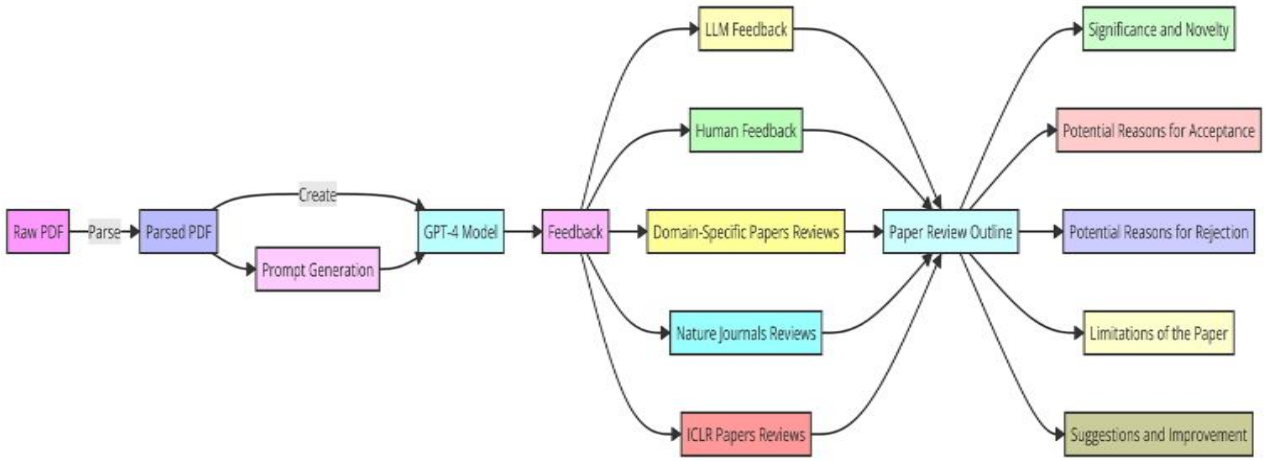

Fig. 1 Proposed system architecture

The GPT-4 Model then processes the generated prompt to produce detailed feedback. This feedback is enhanced by incorporating domain-specific insights from existing reviews and expert knowledge, which helps in addressing technical aspects such as model architecture and design. The system generates feedback that is detailed, including sections on the research's uniqueness and importance, possible justifications for approval or rejection, and practical recommendations for enhancement.

To ensure the feedback's relevance and quality, the system performs a comparison with human reviews. This comparison helps validate the feedback provided by the LLM and ensures that it aligns with expert human evaluations.

The final feedback is organized into a structured outline, the paper review outline, which includes comprehensive sections covering key aspects such as significance, reasons for acceptance or rejection, limitations, and suggestions for improvement. This clear and organized format helps authors interpret and act upon the feedback more effectively.

The system also introduces specificity metrics to measure the relevance and depth of the feedback, ensuring it meets the authors' needs for paper improvement. By combining LLM-generated feedback with human reviews, our hybrid model leverages the strengths of both approaches, resulting in more thorough and reliable feedback.

Iterative Feedback Mechanisms are incorporated to refine the LLM's outputs over multiple review cycles. This iterative process allows for continuous enhancement of feedback quality and relevance.

Overall, our proposed system aims to significantly reduce the number of review cycles required for publication by providing timely, actionable, and domain-specific feedback. By enhancing the quality and relevance of the review process, our approach offers a robust solution that benefits both authors and reviewers, facilitating a more efficient and effective academic review process.

VI. DIFFICULTIES AND FUTURE GOALS

Despite advancements in ML-based review automation, several challenges remain. Addressing these challenges can improve the accuracy, relevance, and adaptability of automated feedback systems in scientific review processes.

A. Model Interpretability and Transparency:

Interpretable models are essential for effective scientific feedback. Complex ML models, like transformers, can lack transparency. Implementing interpretability methods, such as feature importance and output explanation techniques, will enhance trust in automated review feedback.

B. Real-Time Processing and Scalability:

Handling high volumes of review data while maintaining quick processing speeds is essential for an efficient review cycle. Leveraging scalable cloud platforms or optimized models can support real-time feedback delivery, especially under heavy loads.

IARJSET ISSN (O) 2393-8021, ISSN (P) 2394-1588

International Advanced Research Journal in Science, Engineering and Technology

Impact Factor 8.066 ⁜ Peer-reviewed / Refereed journal ⁜ Vol. 11, Issue 12, December 2024

DOI: 10.17148/IARJSET.2024.111208

C. Integration with Existing Review Platforms:

Integrating ML-based feedback systems with existing platforms (e.g., academic journals) can be complex. Standardized APIs and data compatibility solutions can streamline the integration process, enhancing the overall review experience.

D. Ensuring Quality and Specificity of Feedback:

Generating precise, actionable feedback is crucial. Future goals include improving feedback specificity by fine-tuning models on domain-specific datasets and employing quality-check mechanisms to uphold feedback standards.

E. Adaptability to Evolving Research Trends:

As scientific research evolves, models must adapt to new terminologies and review criteria. Continuous learning and regular retraining on updated datasets will improve model relevance and responsiveness to current scientific advancements.

VII. CONCLUSION

This research presents a novel system designed to significantly enhance the scientific paper review process through advanced natural language processing techniques. By leveraging PDF parsing, text summarization, semantic text matching, and fine-tuning large language models (LLMs), our system generates detailed and actionable feedback that mirrors expert human reviews. The integration of domain-specific insights ensures that the feedback is both comprehensive and tailored to various scientific fields.

Our approach addresses key limitations in existing review systems by improving feedback quality, relevance, and specificity. It also reduces the time and effort required for multiple review cycles through iterative feedback mechanisms and hybrid models that combine LLM-generated insights with human expertise.

Overall, the proposed system represents a robust solution for streamlining the academic review process, offering significant benefits to both authors and reviewers by facilitating more efficient and effective evaluations.

REFERENCES

[1]. Liang, W., Zhang, Y., Cao, H., Wang, B., Ding, D., Yang, X., Vodrahalli, K., He, S., Smith, D., Yin, Y., McFarland, D., & Zou, J. (2023). Large language models provide useful feedback on research papers: A large-scale empirical analysis.

[2]. Liu, R., & Shah, N. B. (2023). Reviewergpt? An exploratory study on using large language models for paper reviewing.

[3]. Robertson, Z. (2023). GPT-4 is slightly helpful for peer-review assistance: A pilot study.

[4]. Peter, A. X., Hassan, S., Krishnan, K. S., Jose, T., & Rakhee, M. (2023). Blockchain-based paper review system.

[5]. Zhou, I., Makhdoom, I., Abolhasan, M., Lipman, J., & Shariati, N. (2019). A blockchain-based file-sharing system for academic paper review.

[6]. Aczel, B., Szaszi, B., & Holcombe, A. O. (2021). A billion-dollar donation: Estimating the cost of researchers' time spent on peer review. Research Integrity and Peer Review.

[7]. Bowman, S. R., Angeli, G., Potts, C., & Manning, C. D. (2015). A large annotated corpus for learning natural language inference.

[8]. Hosseini, M., & Horbach, S. P. (2023). Fighting reviewer fatigue or amplifying bias? Considerations and recommendations for use of ChatGPT and other large language models in scholarly peer review. Research Integrity and Peer Review.

[9]. Erkan, G., & Radev, D. R. (2004). LexRank: Graph-based lexical centrality as salience in text summarization. Journal of Artificial Intelligence Research.

[10]. Barzilay, R., & Elhadad, M. (1999). Using lexical chains for text summarization. In Mani, I., & Maybury, M. T. (Eds.), Advances in Automatic Text Summarization

