Fine-tuning a Small LLM with LoRA for Automated Paper Review Insights

Camilla Bonomo

University of Trento

2025/2026

1 Introduction

Peer review remains a foundational pillar of the scientific process, ensuring rigor, validity, and quality through independent expert evaluation (Bornmann, 2011). However, as the number of submissions to journals and conferences continues to grow, the peer review system faces increasing strain, leading to delayed editorial decisions, inconsistent feedback, and reviewer fatigue (Aczel et al., 2025). In parallel, recent advancements in Large Language Models (LLMs) have opened new opportunities for assisting or partially automating aspects of the review process. Current approaches vary in cost, flexibility, and performance, and include: (1) prompt-based zero-shot generation, (2) full supervised fine-tuning on human review data, and (3) parameter-efficient fine-tuning such as Low-Rank Adaptation (LoRA) (see Table 1). Prompt engineering involves using general-purpose LLMs (e.g., GPT-4) in a zero-shot setting to generate review-like text given paper abstracts and task-specific instructions (Gilardi et al., 2023). This approach is simple and requires no additional training, but its outputs are highly sensitive to prompt design and lack domain adaptation. In contrast, full supervised fine-tuning retrains all parameters of an LLM on large corpora of human-written reviews, producing highly specialized models (Zhou et al., 2022). While this method can achieve strong task-specific performance, it demands significant computational resources and large-scale annotated datasets. Parameter-efficient fine-tuning techniques, such as LoRA (Hu et al., 2021), offer a compromise by updating only a small number of low-rank parameters within the model, enabling task adaptation even with limited data and compute.

This project investigates the feasibility of using LoRA to fine-tune a small open-source LLM (e.g., LLaMA 3.2B) on real-world peer review data. Specifically, we aim to construct a lightweight model capable of generating structured outputs—highlighting the strengths and weaknesses of a scientific paper—while also producing numeric predictions (rating and confidence) aligned with human annotations. Our dataset was derived from the OpenReview platform, which provides open-access peer reviews linked to scientific submissions. The experimental design compares zero-shot predictions of the base model against LoRA fine-tuned outputs, evaluated through regression metrics (MAE, RMSE, R^2, Pearson's r, Spearman's ρ) and semantic similarity measures (BERTScore). By contrasting these approaches, we assess whether parameter-efficient fine-tuning can improve alignment with human evaluators while remaining computationally affordable. Unlike studies that rely on large proprietary systems (Tyser et al., 2024), this work emphasizes accessibility, reproducibility, and efficiency, contributing to the emerging field of Automated Scholarly Paper Review (ASPR).

[0]Project repository on GitHub: `https://github.com/camillabonomo02/ML_project.git`

Method	Description	Advantages	Limitations
Prompt Engineering (1)	Use general-purpose LLMs with hand-crafted prompts to simulate reviews.	No training required; fast deployment; flexible across domains.	Output quality depends heavily on prompt; lacks task-specific adaptation.
Full Supervised Fine-Tuning (2)	Train the entire LLM on a dataset of human-written reviews.	High accuracy; task-specific specialization.	Very compute-intensive; requires large annotated datasets.
LoRA (PEFT Fine-Tuning) (3)	Train a small number of task-specific parameters using Low-Rank Adaptation.	Efficient; resource-light; adaptable to smaller LLMs.	May underperform full fine-tuning on highly complex tasks.

Table 1: *Comparison of LLM-based methods for automated paper review.*

2 Methods

This project employed a multi-stage methodology designed to explore whether a lightweight large language model (Meta's LLaMA 3.2B Instruct model) fine-tuned with Low-Rank Adaptation (LoRA) can generate structured peer reviews that approximate human assessment. The pipeline was structured to ensure reproducibility and to provide a rigorous framework for both data preparation and model evaluation.

1. Data collection and preprocessing

The dataset was derived from the OpenReview repository of conference submissions (`tp_2017conference.xlsx`). Each entry contained a paper title, abstract, reviewer comments, rating, confidence score, and final decision. The raw Excel file was first cleaned by removing rows with missing values in essential fields (title, abstract, review) and by standardizing text through regular expressions to eliminate non-printable characters, HTML tags, and formatting artifacts. Duplicate entries based on paper titles were consolidated by concatenating multiple reviews into a single entry. The dataset was then split into training (72%), validation (8%), and test (20%) sets using `train_test_split` from `scikit-learn`. For subsequent regression analysis, the original `rate` and `confidence` fields, which appeared as annotated strings (e.g., "Rating:###7: Good paper, accept"), were normalized into numeric columns (`rating_num`, `confidence_num`). The `decision` field was removed to prevent label leakage, ensuring that the model did not simply learn to justify acceptance or rejection outcomes.

2. Zero-Shot baseline

Before fine-tuning, a zero-shot baseline was established using an instruction-tuned LLM (LLaMA-3.2-3B-Instruct). To guarantee comparability, a rigid prompt template was employed requiring the model to generate outputs in a fixed structure:

```
Strengths:
1. ...
2. ...

Weaknesses:
1. ...
2. ...
```

```
Rating: <number>
Confidence: <number>
```

This format enforced consistency across generated reviews and ensured the simultaneous production of structured text and numeric predictions. The zero-shot model outputs served two purposes: (i) they acted as pseudo-labels to bootstrap fine-tuning on limited human-annotated data, and (ii) they provided a baseline for performance evaluation.

3. Fine-Tuning with LoRA

LoRA was employed to fine-tune the base model efficiently. By introducing low-rank adaptation matrices into attention layers, LoRA enables training with significantly fewer parameters while preserving the frozen base model weights. The fine-tuning dataset paired paper titles and abstracts with human **and** zero-shot generated reviews. Inputs were tokenized with a maximum length of 300 tokens, and labels were aligned with target review outputs. Training employed the Hugging Face `Trainer` API with 4-bit quantization for memory efficiency, a batch size of 2, and three epochs. The structured format of the outputs allowed the model to learn not only to generate strengths and weaknesses but also to provide numerical ratings aligned with human annotations.

4. Evaluation

Model performance was evaluated on the held-out test set using two complementary approaches. First, regression metrics (MAE, RMSE, R^2, Pearson's r, and Spearman's ρ) were computed by comparing predicted `rating` and `confidence` values against ground-truth human scores. This ensured a quantitative assessment of alignment with human evaluation standards. Second, semantic similarity between generated and human-written reviews was measured using BERTScore, providing an auxiliary perspective on textual quality. The zero-shot baseline served as a point of comparison, allowing us to assess whether fine-tuning improved alignment with human assessment or merely replicated generic review patterns.

5. Challenges and limitations

Several challenges were encountered during the process. The human review dataset was relatively small and noisy, with heterogeneous writing styles and variable annotation quality. While zero-shot pseudo-labeling helped expand the dataset, it introduced the risk of propagating biases and generic formulations present in the base LLM. Furthermore, extracting structured numeric predictions reliably required carefully engineered prompts, as unconstrained models often produced incomplete or inconsistent outputs. Finally, computational constraints necessitated the use of quantization and LoRA rather than full fine-tuning, which limited the exploration of larger architectures.

3 Results

This section presents the outcomes of the evaluation of both the zero-shot and fine-tuned models. The focus is on three main aspects: (1) predictive accuracy of numerical ratings and confidence scores, (2) correlation between predicted and ground-truth assessments, and (3) semantic alignment of generated reviews with human-written reviews.

Quantitative Evaluation

Table 2 reports the performance of the models across regression-based metrics, using human-provided ratings and confidence scores as ground truth. Metrics include Mean Absolute Error (MAE), Root

Mean Squared Error (RMSE), Coefficient of Determination (R^2), and Pearson correlation.

Task	MAE	RMSE	R^2	Pearson
Zero-shot Rating	1.24	1.58	-0.05	0.45
Fine-tuned Rating	1.18	1.47	-0.03	0.43
Zero-shot Confidence	0.89	1.12	-1.26	0.05
Fine-tuned Confidence	0.88	1.10	-0.90	0.17

Table 2: *Regression metrics for predicted ratings and confidence scores. Lower MAE/RMSE and higher correlation value indicate better alignment with human annotations.*

For textual similarity, BERTScore was employed to measure the semantic overlap between machine-generated reviews and the original human-written reviews. The fine-tuned model achieved an average F1 score of 0.838, indicating high semantic consistency in content and phrasing.

Qualitative Observations

Manual inspection of outputs confirms that both zero-shot and fine-tuned models reliably produce structured peer reviews with explicit *Strengths* and *Weaknesses*. The fine-tuned model shows improved adherence to the expected template and tends to generate more concise and coherent arguments. However, confidence predictions remain unstable, with weak correlations to ground truth values.

4 Final Discussion

The results suggest that fine-tuning through LoRA slightly improves predictive accuracy for ratings compared to zero-shot prompting, as evidenced by lower MAE and RMSE. Although the R^2 values remain negative—indicating limited explanatory power—the presence of moderate Pearson and Spearman correlations (0.42–0.50) highlights that both models capture meaningful, if noisy, patterns in human assessment. For confidence prediction, both models underperform, with near-random correlation coefficients. This highlights a limitation in using text-based supervision for modeling reviewer certainty, which may require richer contextual features or direct supervision from reviewer metadata. The semantic evaluation with BERTScore demonstrates that fine-tuning improves alignment between generated and human-written reviews, achieving a strong F1 score of 0.838. This suggests that, while numerical predictions remain challenging, the models are capable of generating text that reflects the style and substance of expert reviews.

Limitations and Future Work

Several limitations remain. First, the relatively small dataset constrains both the statistical robustness of the regression models and the generalizability of results. Second, reliance on string-based extraction of ratings and confidence introduces brittleness into evaluation. Third, the models occasionally produce verbose or redundant arguments despite structural improvements.

Future work should explore: (1) larger and more diverse datasets to enhance model robustness, (2) multi-task learning approaches that jointly optimize textual and numerical outputs, and (3) prompt engineering or reinforcement learning strategies to stabilize confidence predictions. Incorporating external reviewer metadata or citation context may also improve alignment with human judgment.

Overall, these findings demonstrate the feasibility of lightweight fine-tuning for automated peer review generation, while also highlighting areas where methodological refinements are necessary for reliable deployment.

6 References

Aczel, B., Barwich, A., Diekman, A., Fishbach, A., Goldstone, R., Gomez, P., Gundersen, O.E.and von Hippel, P., Holcombe, A., Lewandowsky, S., Nozari, N., Pestilli, F., and Ioannidis, J. (2025). The present and future of peer review: Ideas, interventions, and evidence. *Proc. Natl. Acad. Sci. U.S.A.*, 122.

Bornmann, L. (2011). Scientific peer review. *Annual Review of Information Science and Technology*, 45.

Gilardi, F., Alizadeh, M., and Kubli, M. (2023). Chatgpt outperforms crowd-workers for text-annotation tasks.

Hu, E., Shen, Y., Wallis, P., Allen-Zhu, Z., Li, Y., Wang, S., and Chen, W. (2021). Lora: Low-rank adaptation of large language models.

Tyser, K., Segev, B., Longhitano, G., Zhang, X.-Y., Meeks, Z., Lee, J., Garg, U., Belsten, N., Shporer, A., Udell, M., Te'eni, D., and Drori, I. (2024). Ai-driven review systems: Evaluating llms in scalable and bias-aware academic reviews.

Zhou, Y., Muresanu, A., Han, Z., Paster, K., Pitis, S., Chan, H., and Ba, J. (2022). Large language models are human-level prompt engineers.