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

Samuele Carrubba

August 22 2025

Abstract

The rapid growth of scientific publications has put unprecedented strain on the traditional peer review system, motivating the exploration of automated solutions which can assist or replace some aspects of the process. This study investigates whether a lightweight model (Qwen3-1.7B), fine tuned with QLoRa, is able to generate reviews which are similar to human ones, highlighting strengths and weaknesses of a paper. Using the OpenReview 2018 dataset and a set of metrics (ROUGE, BERTScore, Perplexity), the fine tuned model has been compared against both a zero shot approach and the ground truth. QLoRa outperforms the zero shot, while challenges remain in review consistency. **keywords: QLoRa - automated peer reviews - LLM - review generation**

1 Introduction

The peer review process serves as the pillar of scientific publications, ensuring the quality, validity, and rigor of papers through the anonymous evaluation by domain experts.
Beyond these criteria, the feedback from experts can be valuable in the overall improvement of the submitted manuscripts. However the traditional system of manual reviews today faces numerous challenges due to the growth in the number of submissions and the lack of highly specialized reviewers. These challenges resulted in longer review times, increased burden for reviewers and concerns about the sustainability of manual peer review.
Recent advances in large language models (LLMs) have opened new possibilities for addressing these scalability challenges. The demonstrated capabilities of LLMs in understanding complex text, generating coherent responses, and performing domain specific reasoning have the potential of partially or fully automating aspects of the review process.
The most recent works addressed the need of automating this process by exploring resource efficient approaches to automated review assistance.
The code can be found here: https://github.com/SamC-dev/QLoRA-for-Automated-Paper-Review-Insights

2 State of the art

2.1 literature review

The exponential growth of scientific publications has created unprecedented challenges for traditional peer review systems. As Yuan et al. (2021) note: "the rapid development of science and technology has been accompanied by an exponential growth in peer-reviewed scientific publications... providing high-quality reviews of this growing number of papers is a significant challenge". This volume growth, combined with the inherent "poor efficiency and low reproducibility" of traditional peer review (Lin et al., 2023), has motivated the development of automated solutions.
Foundational work in this area began with the recognition that the peer review system is not only bur-

dened by volume, but that the entire peer reviewing system is flawed, peer reviewing lacks reproducibility because of a subjective component during the evaluation, which is furthermore characterized by the knowledge limitations of the reviewer, particularly accentuated in highly specialized subfield.
It's evident the need of delegating the whole reviewing process to a machine. While initial works found out that machines could have definitely been used as assistants during the process, soon studies experimented on making the machine more independent.
Yuan et al. (2021) gave the foundations which allowed to go from a simple assistant to a more independent agent. they proposed a NLP model capable of generating reviews for scientific papers. The system was able to generate more detailed reviews, offering a starting point in this direction.
Lin et al. (2023) formalized the concept of Automated Scholarly Paper Review (ASPR), developing a comprehensive "pipeline... for achieving a full-scale computerized reviewing process." They outlied the biases brought by human reviewers while also highlighting the challenges of ASPRs: imperfect document parsing and representation, defective human–computer interaction, and flawed deep logical reasoning.
Recent advances demonstrate remarkable progress in addressing these limitations. Wu et al. (2025) used a LLM to match or exceed human quality. Their four-module architecture includes literature search, topic formulation, knowledge extraction, and review composition, supported by multi-layered quality control that reduced hallucination risks.
The recent survey by Lin et al. (2025) provides a holistic view of ASPR in the era of LLMs, examining which models are being used for automated review and identifying new methods, new datasets, new source code, and new online systems that come with LLMs for ASPR.
The evolution from AI assisted tools to fully automated ones is clear. The progression of this research establishes a clear foundation for fine-tuning approaches.

2.2 Research question

This study investigates whether fine-tuning a lightweight LLM with QLoRA enables the generation of reliable reviews for scientific papers, with a focus on the ability to identify both strengths and weaknesses and to assess its performance in comparison with a zero shot setting and the ground truth.

3 Research methodology

3.1 Data

This study utilizes the OpenReview 2018 dataset, comprising around 4700 reviews for a total of around 1500 papers, representing a valid dataset which has been split into train, validation and test sets (4:3:3). To each review is associated a rating from 1 (very poor) to 10 (excellent).

3.2 Model

The study employs Qwen 3-1.7B as baseline architecture, and in order to improve its performance it has been fine tuned with QLoRa (Quantized Low Rank adaptation). LoRa is a technique which simplifies the fine tuning process of a large model, focusing on a small number of parameters, while keeping the others frozen, its quantized version (QLoRa) reduces memory heaviness by compressing the parameters, making the approach more convenient, especially with limited computational resources.

3.3 Formalization of the experiments

The experiments followed a comparative approach, taking into consideration QLoRa fine tuning with respect to ground truth and the model's zero shot capabilities. The evaluation focused on the quality of reviews, assessed using the following metrics, as done in the original implementation:

- ROUGE: Measures the overlap between the generated and reference reviews, it can be seen as a similarity measure between generated and real reviews. Its three variants are: ROUGE1 (unigram overlap), ROUGE2 (bigram overlap) and

ROUGEL (longest common subsequence), capturing different levels of similarity.

- BERTscore (RoBERTa): Computes semantic similarity between generated and reference reviews using contextual embeddings from a pretrained model, providing a more meaning aware assessment than simple n-gram overlap.

- Perplexity: Evaluates how well the language model predicts the next word in the reference text. Values are in the range 1-100, with lower ones indicating a more fluent and consistent text.

3.4 Prompt structure

Different trials have been made in order to set an efficient prompt. The final one asks to generate a review starting from the abstract and, based on this, to write 2-4 strengths and weaknesses (or areas of improvement), to summarize in 2-3 sentences and finally to assign a rating to the paper. This way the model prioritizes the review first, while the rate is presented as the logical conclusion.

```
def make_prompt(abstract):
    """
    Constructs the prompt string for the LLM to generate peer reviews.
    """
    return (
        "You are an expert academic peer reviewer. Your task is to provide a constructive, "
        "detailed review of a research paper based on its abstract.\n\n"

        f"Paper Abstract:\n{abstract}\n\n"

        "Instructions:\n"
        "- Analyze the research contribution, methodology, and potential impact\n"
        "- Identify specific strengths and areas for improvement\n"
        "- Provide constructive feedback that would help the authors\n"
        "- Rate the paper objectively on a 1-10 scale (1=reject, 10=accept)\n\n"

        "Required Format:\n"
        "**Strengths:**\n"
        "- [List 2-4 specific strengths]\n\n"

        "**Weaknesses:**\n"
        "- [List 2-4 specific weaknesses or areas for improvement]\n\n"

        "**Overall Assessment:**\n"
        "[2-3 sentences summarizing your overall opinion]\n\n"

        "**Rating:** [X/10]\n\n"
        "Review:"
    )
```

Figure 1: Prompt structure

3.5 Training

Training has been done using the following setting:

- Batch size has been set to 4, with gradient accumulation. Trials with larger batch sizes have been proven unsuccessful due to memory constraints.

- ADAM as the optimizer of choice, it's the one used in the original paper, it represents the state of the art in terms of optimization algorithms and it typically performs better than other conventional optimizers such as SGD.

- The number of epoch has been set to 5 in order to find a compromise between the length of training and optimal results.

- The hyperparameters (learning rate and weight decay) are set in accordance to the original implementation

- an early stop criterion, based on steps and with patience equal to 2, has been implemented in order to stop training if the validation loss doesn't drop further.

4 Results and Discussion

By visually comparing a sample of generated reviews with their ground truth, it's evident that the model still requires refinement. Nonetheless the reviews capture the strengths and weaknesses of their papers quite effectively. The following table summarizes the Rouge and BERT scores of both zero shot and QLoRa:

	Metric	F1	Precision	Recall
0	ROUGE-1	0.255967	0.493851	0.189396
1	ROUGE-2	0.050840	0.101081	0.037182
2	ROUGE-L	0.140696	0.274915	0.103880
3	BERTScore	0.818107	0.831292	0.805806

	Metric	F1	Precision	Recall
0	ROUGE-1	0.261688	0.588397	0.196824
1	ROUGE-2	0.057218	0.137130	0.042427
2	ROUGE-L	0.144081	0.329971	0.108929
3	BERTScore	0.829340	0.851585	0.808447

Figure 2: Comparison of results: zero shot on the left and QLoRa on the right

QLoRa outperforms zero-shot generation across

nearly all metrics, its perplexity (10.01) is around half of the one registered with zero shot (19.34). Good improvements can be observed mainly in precision. To be observed how the generated rating is fare from being close to the real ones, with a MAE of approximately 6. This discrepancy may be due to: the model being unable to assign consistent ratings, the absence of a rating in some of the reviews or a issue occuring during the evaluation procedure.

5 Conclusion

This study explored the potential of fine-tuning a lightweight LLM using QLoRA to assist in automated scholarly paper review.
The results showed an improvement compared to a zero shot setting and a good similarity with human reviews. The fine tuned model effectively captures strengths and weaknesses. These findings are in line with the promising results obtained in this field with LLM based approaches. Despite these encouraging results, several limitations remain:

- First, the generated reviews are not perfect, and inconsistencies in content and rating assignments were observed.
- The study relies on the OpenReview 2018 dataset, which, while substantial, may not fully capture the diversity of scientific domains or review styles, this is a common problem in this field literature.
- The training process was limited due to resource constraints.

Future work should address these limitations.

6 References

1. Yuan et al., Can We Automate Scientific Reviewing?, 2021
2. Lin et al., Automated Scholarly Paper Review: Concepts, Technologies, and Challenges, 2023
3. Wu et al., Automated Review Generation Method Based on Large Language Models, 2024
4. Lin et al., Large language models for automated scholarly paper review: A survey, 2025
5. Mastering QLoRa : A Deep Dive into 4-Bit Quantization and LoRa Parameter Efficient Fine-Tuning, https://manalelaidouni.github.io/4Bit-Quantization-Models-QLoRa.html, 08/06/2024
6. Hu et al., LoRA: Low-Rank Adaptation of Large Language Models, 2021
7. Dettmers et al., QLORA: Efficient Finetuning of Quantized LLMs, 2023
8. ICLR dataset,

   https://github.com/Seafoodair/Openreview
9. source code, https://github.com/SamC-dev/QLoRA-for-Automated-Paper-Review-Insights