University of Trento
Department of Sociology and Social Research
Introduction to Machine Learning
LORENZO D'AMICO 238684

"Automated Paper Reviewing with a Compact LLM: LoRA Fine-Tuning"

Abstract

This paper assesses whether a compact ~3 B-parameter language model, adapted with QLoRA, can support peer review by generating structured feedback and an explicit Accept/Reject decision. Training uses OpenReview ICLR 2017–2019 data, with editorial decisions mapped to a binary label and a token-level cross-entropy optimised over review text plus decision. Evaluation is strictly out of distribution on ICLR 2020. A two-prompt zero-shot baseline (Llama-3.2-3B-Instruct) is fluent but severely biased, accepting 2 199 of 2 203 submissions and achieving only 30.9 % accuracy. Fine-tuning with QLoRA updates about 1.33 % of the parameters and shifts the decision distribution to ≈ 1 650 Accept / 546 Reject (N ≈ 2 196), raising accuracy to 41.7 %. Beyond accuracy, the tuned model improves lexical overlap with human reviews (e.g. token-Jaccard mean 0.130→0.146, Weaknesses ROUGE-1 0.149→0.169, Strengths chrF 0.205→0.217) while reducing validation loss and perplexity (≈ 2.27 and ≈ 9.7 vs. ≈ 3.0 and 20 for the backbone). The zero-shot baseline retains a small advantage on semantic similarity metrics (Sentence-BERT recision/recall/F1) and on Strengths-ROUGE-L. A KL-regularisation experiment, aimed at inheriting the baseline's fluency and semantic alignment, over-constrained adaptation and degraded loss and perplexity without improving accuracy. Overall, lightweight QLoRA adaptation moves a compact model from "fluency first" toward "fluency with selectivity," offering a practical, reproducible improvement for early-stage peer-review support.

Introduction

Peer review is the primary mechanism for quality control in scholarly publishing, yet it is expensive, slow and often inconsistent. Instruction-tuned language models can generate fluent prose on demand; however, fluency alone is insufficient for this task—an automated reviewer must also render a binary verdict. In this study I explore whether a compact ~ 3 B-parameter LLM, adapted with LoRA/QLoRA, can deliver a disciplined first pass on submissions by producing a concise, structured critique and an explicit Accept/Reject decision. To test genuine generalisation, I train only on OpenReview dumps from ICLR 2017–2019 and reserve ICLR 2020 as a strictly out-of-distribution test. Each training record (title, abstract, review, decision) is normalised and cleaned; we construct prompt–completion pairs where the prompt concatenates the paper's title and abstract, and the completion concatenates the review text and a final decision line (`Decision: Accept`/`Reject`). Decisions are binarised using a simple heuristic ("accept"→Accept, else Reject); the decision label is never included in the prompt to prevent leakage. This formulation casts the task as a standard binary classification, aligning the modelling objective with the practical goal of editorial decision-making. As a zero-shot baseline, I use Llama-3.2-3B-Instruct with a two-prompt strategy: one prompt elicits a structured review (Decision, Comment, Strengths, Weaknesses), and a separate prompt forces a single-token verdict. While the baseline always returns a decision on the ICLR 2020 set, it exhibits an extreme acceptance bias, predicting *Accept* for 2 199 out of 2 203 papers and achieving only 30.9 % accuracy. Fine-tuning with QLoRA updates only ~1.3 % of the parameters and uses a token-level cross-entropy loss over the completion tokens; nevertheless, it learns the OpenReview style and decision cues. The QLoRA-tuned model shifts the decision distribution to a much more balanced 1 650 Accept / 546 Reject and improves accuracy to 41.7 %, while incurring a modest degradation in language-modelling fluency. Evaluation includes not only accuracy and class distribution but also lexical (Jaccard, ROUGE, chrF) and semantic (SBERT, BERTScore) similarity metrics, revealing that the tuned model produces reviews that are more lexically aligned with human feedback, whereas the baseline retains a slight edge on semantic similarity. Overall, this compact, reproducible pipeline highlights the promise—and the trade-offs—of lightweight adaptation for early-stage peer-review support.

Data and Preprocessing

We build our dataset from OpenReview's "absolute_data" dumps. Conferences from 2017–2019 are used for training/validation, while ICLR 2020 is held out for out-of-distribution testing. Each raw record includes a paper title, abstract, one or more review texts and an editorial decision. We first normalise column names and remove rows with missing title, abstract, review or decision. For the training split, we then construct prompt–completion pairs suitable for language-model fine-tuning:

– The prompt is formed by concatenating the paper's title and its cleaned abstract, separated by two newlines (`title + "\n\n" + abstract`). Prefixes such as "Abstract:###" are stripped during cleaning.

– The completion consists of the full review text followed by a binary decision line, e.g. `Decision: Accept` or `Decision: Reject`. We binarise the original editorial decision using a simple heuristic: any string containing "accept" (case-insensitive) maps to Accept, and all others map to Reject. This step also strips prefixes like "Recommendation:###" or "Decision:###". After assembling these pairs, we remove exact duplicates, assign a `year` field, and save the resulting dataframe as a Parquet file (`train_17_18_19.parquet`). For the 2020 test set, we perform only the cleaning steps (normalise column names, strip whitespace, drop missing values and duplicates) and save the cleaned table to `test_20.parquet`; ; we do not append the decision to the review, because the true label is used solely for evaluation and is never included in the model's prompt to avoid leakage. This pipeline ensures that the model trains on structured prompts closely matched to the format it will generate at inference time

while keeping the out-of-distribution test data strictly separate.

Metrics

Our task is a binary classification of the editorial decision (Accept vs. Reject) coupled with a generative review. The primary metric is accuracy, which measures the share of correct binary decisions on the out-of-distribution ICLR-2020 set; we also inspect the predicted class distribution to detect degenerate behaviours (e.g., "always accept"). To evaluate review quality, we use lexical overlap metrics—Jaccard, ROUGE-1/2/L, and chrF—that quantify token/character n-gram overlap with human reviews, checking adherence to the requested structure and coverage of strengths/weaknesses. Because surface overlap can miss paraphrases, we report semantic similarity—SBERT (cosine; precision/recall/F1) and BERTScore (F1)—to assess meaning-level alignment. Finally, we track cross-entropy and perplexity as language-model (regression-style) diagnostics of predictive uncertainty and fluency; they are not target metrics but help interpret the trade-off between fluency and the more balanced selectivity achieved by fine-tuning.

Baseline: Zero-Shot with Structured Prompting

In the fine-tuning phase I adapted the same ~3 B-parameter backbone using QLoRA, which combines 4-bit `nf4` quantization with lightweight low-rank adapters. Only 24 313 856 parameters were updated out of 1 827 777 536 total, i.e. roughly 1.33 % of the model. The optimization objective is a token-level cross-entropy computed only over the completion tokens—the structured review text plus the final `Decision:` line—while masking out the prompt tokens (title and abstract). This design forces the model to learn to generate a review and commit to a decision, rather than simply echoing the input. Editorial decisions are mapped directly to a binary target (*Accept* or *Reject*), aligning training with the evaluation goal. Training dynamics were fast: the best checkpoint without any extra regularisation appeared around step 400 (see **Fig. 1**). Measured on a held-out validation split, this model achieved a validation loss of ≈2.27, corresponding to a perplexity of ≈9.7, whereas the unfine-tuned backbone started at loss ≈3.0 and perplexity ≈20. In other words, adjusting just 1.33 % of the parameters reduces predictive uncertainty by more than a third and roughly halves perplexity. These low-rank adapters appear sufficient to internalise the OpenReview writing style and the decision cues specific to ICLR, letting the model move beyond paraphrasing abstracts to actually making a judgement. For inference I truncated title–abstract pairs to fit within the token budget and used separate decoding strategies for the two outputs: greedy decoding for the single-token decision (to guarantee reproducibility) and light sampling (temperature 0.7, top-p 0.9) for the longer review, which keeps the review natural and less repetitive. Because only about one per cent of the parameters are trainable and the rest are held in 4-bit quantized form, the approach remains memory-efficient and practical on commodity GPUs while still delivering a substantial shift from "fluency first" to "fluency with selectivity".

Figure 1. Validation loss vs. training steps. QLoRA peaks at step 400; KL underperforms.

Evaluation

All evaluation is performed on the held-out ICLR 2020 set, which is strictly out of distribution relative to the training corpus (ICLR 2017–2019). To avoid leakage, I reserve 8 % of the 2017–2019 data as a validation split and deduplicate items by *(title, abstract)*; during evaluation I match items using the same unique identifier. Editorial outcomes are mapped directly to a binary Accept/Reject label, and the model never sees the true decision during generation. The primary metric is overall binary accuracy on the 2020 set. As a sanity check against degenerate behaviour (e.g., "always accept"), I also monitor the distribution of predicted decisions (percentage of Accept vs. Reject). To assess the quality of the generated reviews beyond the final verdict, I compute a suite of text-similarity diagnostics between model outputs and human reviews. These include token-level Jaccard overlap, ROUGE-1/2/L and chrF scores on the strengths and weaknesses sections, as well as SBERT precision/recall/F1 and BERTScore (F1) to capture semantic alignment. These metrics provide complementary lexical and semantic perspectives on how closely the model's reviews mirror human feedback. Finally, I monitor validation loss and perplexity as regression-style signals of language modelling quality; although they are not used for selection, they help interpret the trade-off between fluency and selectivity observed in the results.

Results

On the ICLR 2020 test set (N≈2 203), the zero-shot baseline is strongly imbalanced: it predicts *Accept* for 2 199 papers and *Reject* for just 4, yielding an overall binary accuracy of 0.309. By contrast, the QLoRA-tuned model outputs a decision for 2 196 papers with a much more balanced distribution—around 1 650 Accept and 546 Reject—and improves accuracy to 0.417. The shift in predicted distribution is almost as telling as the accuracy gain: QLoRA is less prone to the trivial "always accept" behaviour and more willing to flag incremental novelty or weak empirical evidence (see Fig. 2–3). To assess review quality, I compared generated texts against human reviews. On lexical overlap, QLoRA shows modest gains: the mean

token-Jaccard similarity rises from 0.1300 (baseline) to 0.1457, and the maximum token-Jaccard increases from 0.1480 to 0.1633. ROUGE-1/2/L and chrF scores also improve in many cases (e.g. Weaknesses ROUGE-1 0.149→0.169; Strengths chrF 0.205→0.217; Jaccard on Weaknesses 0.071→0.081). However, QLoRA underperforms the baseline on a few metrics—notably Strengths ROUGE-L (0.124→0.110) and, more importantly, on semantic similarity: Sentence-BERT precision/recall/F1 scores are higher for the zero-shot model (e.g. mean cosine 0.6326 vs. 0.5854; F1 on soft-matching strengths 0.056 vs. 0.030). On language-modelling signals, the fine-tuned model clearly reduces predictive uncertainty: its best checkpoint achieves a validation loss of ≈ 2.27 and a perplexity of ≈ 9.7, compared with ≈ 3.0 and 20 for the unfine-tuned backbone. Thus, the baseline does *not* outperform QLoRA on loss or perplexity; its edge lies solely in smoother fluency and higher semantic similarity. This semantic gap motivated a KL-regularisation test: the aim was to combine the baseline's SBERT-based alignment with QLoRA's balanced selectivity. In practice, anchoring the adapters to the frozen backbone acted as an overly strong prior: the KL-anchored model exhibited higher validation loss (≈ 2.50 vs. 2.27) and worse perplexity (≈ 10.1 vs. 9.7) without improving decision accurac

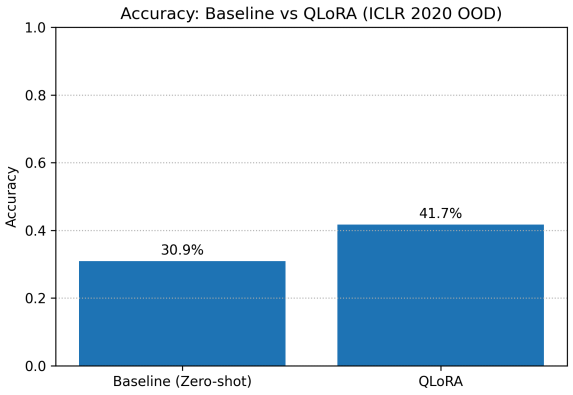

Figure 2. Accuracy on ICLR 2020 (out-of-distribution): zero-shot baseline vs. QLoRA.

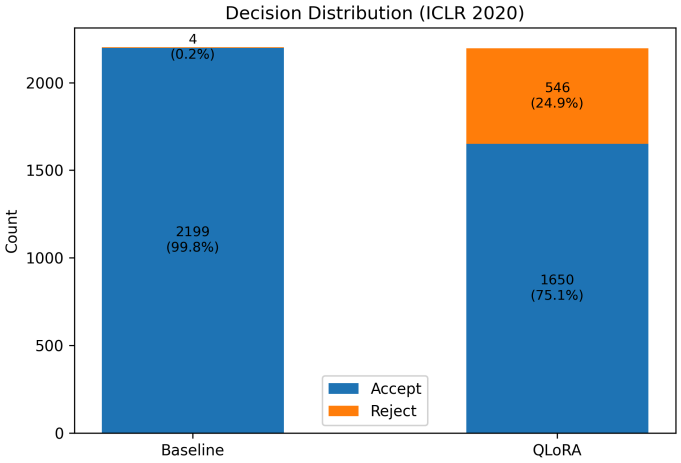

Figure 3. Decision distribution (counts). Note: Baseline N=2203, QLoRA N=2196.

KL Regularization: a good idea, but…

The results above show a clear trade-off: QLoRA dramatically rebalances the Accept/Reject distribution and improves lexical overlap with human reviews, while the zero-shot baseline remains superior only on **semantic similarity** metrics such as Sentence-BERT precision/recall/F1 and, to a lesser extent, on a few lexical scores like Strengths ROUGE-L. This residual semantic gap – not any advantage in loss or perplexity, which QLoRA markedly reduces – motivated an experiment with **Kullback–Leibler (KL) regularization**. The idea was to encourage the fine-tuned model to stay closer to the backbone's probability distribution during training, hoping to inherit its smoother, semantically aligned language while retaining QLoRA's balanced decision-making. In practice, this strategy backfired. The teacher distribution was itself strongly skewed toward *Accept*, and the labelled dataset was small and noisy, so the KL penalty pulled the adapters back toward zero-shot behaviour. The resulting models exhibited **higher validation loss (≈ 2.50 vs. 2.27) and worse perplexity (≈ 10.1 vs. 9.7)** without closing the gap on SBERT-based metrics. In other words, regularizing toward the baseline preserved the very aspects we wanted to improve (fluency and acceptance bias) while suppressing the domain-specific signals that made QLoRA more selective. A more balanced teacher distribution or an annealed KL schedule might work better, but in the present setting the penalty hindered progress precisely when the model needed freedom to diverge from the backbone's semantics.

Discussion

On the ICLR 2020 test set (N≈2 203), the zero-shot baseline (Llama-3.2-3B-Instruct) is extremely imbalanced: it predicts *Accept* for 2 199 papers and *Reject* for only 4, achieving an overall binary accuracy of 0.309. By contrast, the QLoRA-tuned model outputs a decision for 2 196 papers with a much more balanced distribution—around 1 650 Accept and 546 Reject—and increases accuracy to 0.417 (see Fig. 3). This shift in predicted distribution is as telling as the accuracy gain. Beyond the verdict, QLoRA improves several lexical overlap metrics. The mean token-level Jaccard similarity between generated and human reviews rises from roughly 0.130 to 0.146, and the maximum Jaccard scores improve similarly. ROUGE and chrF scores on the *Weaknesses* and *Strengths* sections also increase in most cases (e.g., *Weaknesses* ROUGE-1 goes from 0.149 to 0.169; *Strengths* chrF from 0.205 to 0.217). These gains indicate that the fine-tuned model captures more non-trivial issues such as missing ablations, benchmark coverage and novelty. However, the baseline retains an advantage on semantic similarity: Sentence-BERT precision/recall/F1 scores remain higher for the zero-shot model, and on some lexical metrics—like *Strengths* ROUGE-L—the baseline performs slightly better. In short, QLoRA makes the model more selective and lexically aligned with human reviews, but at the cost of reduced semantic closeness. Regarding language-modeling signals, QLoRA reduces predictive uncertainty. The best fine-tuned checkpoint reaches a validation loss of about 2.27 and a perplexity around

9.7, versus roughly 3.0 and 20 for the unfine-tuned backbone. Thus, the zero-shot baseline does not outperform QLoRA on perplexity; its strength lies instead in smoother fluency and higher SBERT scores. The overall trade-off is between fluency/semantic similarity and selectivity/lexical relevance: the tuned model sacrifices some semantic similarity in exchange for a substantial gain in balanced decision-making and task-specific lexical coverage, while remaining efficient (≈ 1.33 % of parameters trained) and memory-friendly.

Threats to Validity and Limitations

The underlying data remain noisy and biased: OpenReview texts are heterogeneous, and editorial decisions may reflect venue-specific norms rather than universally shared criteria. Holding out the 2020 cycle introduces a realistic distribution shift, but it also changes the topical mix of submissions, making it harder to disentangle methodological improvements from topic effects. Our expanded evaluation highlights another subtlety: different similarity metrics tell different stories. QLoRA improves lexical overlap (Jaccard, ROUGE, chrF) yet the zero-shot backbone still fares better on semantic similarity (SBERT precision/recall/F1), and these automatic metrics are imperfect proxies for human judgment. Moreover, by generating the decision separately from the review, the model never has to justify its verdict within the narrative; linking the verdict to a structured rubric and scoring both the decision and its rationale would provide a stronger test of review quality. External validity is limited: results are reported only on ICLR-style submissions, using a compact 3 B-parameter backbone; larger models and cross-venue evaluation would be needed to generalize to other domains. Finally, the small, skewed dataset and limited GPU budget restrict the breadth of hyperparameter searches and may underestimate variance; more data and computation would allow for a more thorough exploration of the trade-off between fluency and selectivity.

Reproducibility Notes

The core pipeline remains intentionally simple, but it now includes a few additional steps to ensure a faithful replication of the evaluation. After preprocessing the data (with editorial decisions included only in the target) and splitting into train/validation/test partitions (ICLR 2017–2019 for training/validation vs. 2020 for OOD testing), the process runs a two-step zero-shot baseline (structured review followed by a single-token Accept/Reject decision). For fine-tuning, QLoRA adapters are trained on the binary labels, and the best checkpoint is selected using validation loss. At evaluation time, each test paper's title and abstract are matched strictly by normalized `(title, abstract)` keys, and the true decision is never exposed in the prompt. In addition to reporting overall accuracy and the predicted class distribution, the updated pipeline computes a suite of text-similarity diagnostics (e.g., token-level Jaccard, ROUGE-1/2/L, chrF, SBERT cosine and BERTScore) between generated reviews and their human counterparts. A qualitative sample of generated reviews can be provided (with identifiers masked) to illustrate strengths, weaknesses and decision rationale.

Conclusion

Our study demonstrates that a compact 3 B-parameter backbone adapted with QLoRA learns not only the "voice" of a reviewer, but—crucially—a more balanced selectivity for accept vs. reject decisions. Compared with a fluent yet biased zero-shot baseline, the QLoRA-tuned model improves out-of-distribution accuracy and yields higher lexical-overlap scores, though the baseline retains an edge on semantic-similarity measures such as SBERT. The KL-regularization experiment clarifies that regularization is context-dependent: with few, noisy labels and a biased teacher distribution, a KL penalty can over-constrain the very changes that matter, raising validation loss and perplexity without closing the semantic-similarity gap. Overall, the approach remains practical and reproducible, offering immediate utility for early screening and author feedback while leaving final judgment to human reviewers.

References

OpenReview.net. (2020). *Conference peer review data (ICLR 2017–2020 absolute_data dumps)*. Retrieved from https://github.com/Seafoodair/Openreview

Hu, E. J., Shen, Y., Wallis, P., Allen-Zhu, Z., Li, Y., Wang, L., ... & Chen, W. (2021). *LoRA: Low-Rank Adaptation of Large Language Models*. arXiv preprint arXiv:2106.09685.

Dettmers, T., Pagnoni, A., Holtzman, A., & Zettlemoyer, L. (2023). *QLoRA: Efficient Finetuning of Quantized LLMs*. arXiv preprint arXiv:2305.14314.

Meta AI. (2024). *The Llama 3 Herd of Models (Model Card)*. Retrieved from https://ai.meta.com/llama/