Distilling Human Critical Feedback for Aspect-Guided Review Generation with LongLLaMA

Esra Şekerci
Department of Information Systems
Middle East Technical University
Ankara, Turkey
esra.sekerci@metu.edu.tr

***Abstract*—This study will investigate the automated generation of structured peer review feedback for scholarly submissions by fine-tuning LongLLaMA models under human supervision. Leveraging the ASAP-Review dataset, which provides full-text papers annotated with human-written reviews, we will implement a knowledge distillation framework where human critiques serve as teacher signals to guide the generation of aspect-grounded outputs. The methodology will address two key objectives: evaluating the effectiveness of human-generated distillation signals for aspect-aware review generation and assessing interpretability improvements through multi-aspect supervision across clarity, motivation, and soundness dimensions. Overall, the proposed approach aims to advance automated review generation while ensuring transparency and alignment with established peer evaluation standards.**

Keywords—Aspect annotation, interpretability, knowledge distillation, LongLLaMA, multi-aspect supervision, review generation.

I. Introduction

Automated scientific review generation aims to replicate the structured critical assessment traditionally conducted by human peer reviewers. The availability of comprehensive datasets such as ASAP-Review, which provide full-text scholarly submissions annotated with aspect-specific human feedback, facilitates the development of models capable of producing fine-grained evaluative outputs. This study will prioritize two objectives: first, it will distill human-written reviews to guide transformer models toward aspect-grounded critique generation; and second, it will apply multi-aspect supervision across dimensions such as clarity, motivation, and soundness to improve interpretability and review quality. To address the challenges posed by long-form input sequences, the methodology will fine-tune LongLLaMA, a transformer architecture optimized for extended context modeling, using human-generated critiques as supervision targets.

II. Literature Review

Processing full scientific papers presents unique challenges for transformer architectures due to sequence length limitations. Longformer [1] introduced sparse attention mechanisms that scale linearly with input length, addressing computational inefficiencies. More recently, LongLLaMA [2] extended OpenLLaMA models with Focused Transformer (FoT) layers, enabling efficient training and inference with sequences up to 256k tokens, making it particularly suitable for full-text review generation tasks.

Efficient fine-tuning of large models on resource-constrained hardware is critical for practical deployment. The Unsloth framework [3] improves the fine-tuning efficiency of transformer models through memory optimization techniques, including quantization (4-bit, 16-bit) and low-rank adaptation (LoRA) [4]. LoRA enables parameter-efficient transfer learning by injecting trainable low-rank matrices into frozen pre-trained models, substantially reducing memory and computational overhead.

Aspect-guided generation strategies have been proposed to enhance the structure and relevance of generated reviews. Li et al. [5] developed an aspect-aware coarse-to-fine decoding framework, demonstrating that guiding generation across aspect-specific stages improves informativeness and coherence. Incorporating explicit aspect annotations aligns model outputs with the evaluative dimensions emphasized during peer review.

Knowledge distillation [6] provides a mechanism for transferring structured feedback from human-written reviews to student models. The Fault-Aware Distillation via Peer-Review (FAIR) framework [7] improves knowledge distillation by simulating peer review among multiple teacher models. Instead of relying solely on single-model rationales, FAIR focuses on identifying and explaining student mistakes, enhancing supervision quality and promoting more structured, interpretable outputs.

Collectively, the integration of long-document transformers, parameter-efficient fine-tuning techniques, aspect-conditioned generation, and human-guided distillation presents a promising framework for enhancing the quality and interpretability of automated scientific reviews.

III. Exploratory Data Analysis

An exploratory data analysis was conducted to characterize the structural properties, label distributions, and annotation richness of the ASAP-Review dataset. The analysis encompassed submission metadata, review attributes, and aspect-level annotations to assess the dataset's suitability for downstream learning tasks.

The corpus consists of 8,877 scholarly submissions and 28,122 peer reviews, aggregated from ICLR (2017–2020) and NIPS (2016–2019). Each submission is associated with an average of 3.17 reviews, closely mirroring standard peer review assignment practices employed by major machine learning conferences. Acceptance outcomes indicate that 5,408 submissions (60.92%) were accepted, while 3,469 submissions (39.08%) were rejected. This imbalanced distribution arises partially from procedural biases, notably the NeurIPS policy of publicly releasing reviews exclusively for accepted papers. As a result, survivorship bias stemming from the exclusion of rejected NeurIPS submissions may introduce distributional shifts, potentially impacting model generalization performance in downstream tasks.

Review verbosity was analyzed through review length distributions. Figure 1 presents the histogram of review lengths across all submissions. Reviews exhibited an average length of approximately 430 words, with observed lengths

ranging from 43 to 812 words. Such variance suggests heterogeneous reviewer engagement levels. A Mann-Whitney U test was conducted to compare review lengths associated with accepted and rejected submissions, yielding a statistically significant difference ($U = 64{,}299{,}425.50$, $p < 0.001$). Descriptive statistics confirmed that accepted submissions exhibited a slightly higher median review length compared to rejected submissions, further supporting the hypothesis that higher-quality submissions tend to elicit more detailed reviewer feedback.

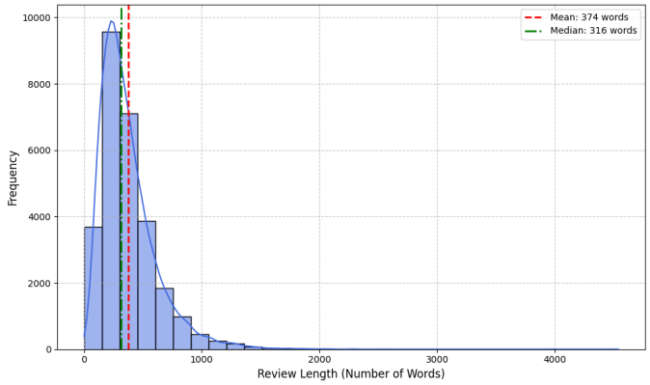

Figure 1 - Histogram of review lengths across all peer reviews

Reviewer confidence scores, reported on an ordinal scale from 1 to 5, exhibited a mean of 3.33 and a median of 4.00. The distribution of reviewer confidence scores is depicted in Figure 2, indicating that most reviewers expressed moderate to high certainty in their evaluations. A Pearson correlation analysis between reviewer confidence and assigned rating scores yielded a weak but statistically significant negative correlation ($r = -0.139$, $p < 0.001$). This finding suggests that higher reviewer confidence is very slightly associated with assigning lower ratings, although the effect size is minimal and unlikely to substantially affect downstream analyses.

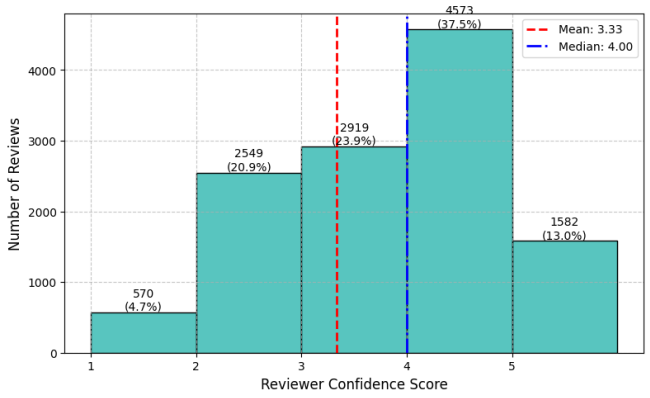

Figure 2 - Histogram of reviewer confidence scores

Aspect-level annotations were analyzed to assess the evaluative focus within peer reviews. Figure 3 shows that clarity and soundness, particularly their negative assessments, dominate reviewer feedback, each annotated over 15,000 times. Positive annotations for clarity, soundness, and motivation follow, indicating attention to core scholarly dimensions. Conversely, aspects such as replicability and meaningful comparison are infrequently annotated, reflecting either lower reviewer emphasis or the difficulty of assessing these attributes in submitted work. This distributional skew suggests that reviewer attention is systematically concentrated on certain dimensions, potentially introducing inductive biases into models trained on aspect annotations. Future modeling efforts must account for this imbalance to avoid disproportionate weighting of dominant aspects during supervised learning.

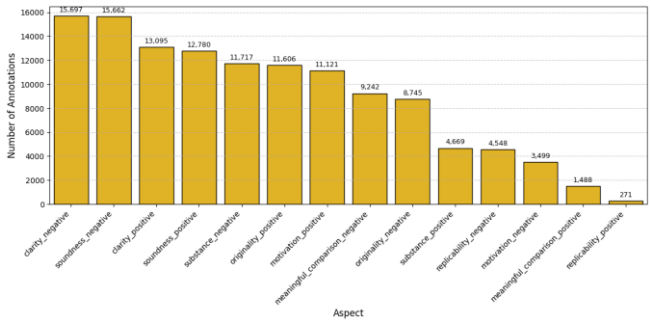

Figure 3 - Distribution of aspect annotations across all peer reviews

Finally, data completeness was evaluated to assess the dataset's suitability for transformer-based deep learning applications. Full-text content was successfully parsed for 8,846 out of 8,877 submissions, yielding a content availability rate of approximately 99.65%. This high availability minimizes concerns regarding data sparsity and ensures robust textual coverage across the corpus. Descriptive analysis of paper contents revealed that the majority of submissions contain between 4,000 and 6,000 words, providing sufficient textual depth for subsequent linguistic and semantic analyses.

In line with the study design, the conference origin (ICLR or NeurIPS) of each submission was not incorporated as an explicit feature, ensuring that subsequent modeling focuses exclusively on intrinsic textual characteristics rather than venue-specific artifacts. Collectively, the ASAP-Review dataset demonstrates high coverage, fine-grained annotations, and minimal missingness, validating its applicability for supervised, semi-supervised, and interpretability-driven modeling methodologies in the context of large-scale transformer-based systems.

IV. Methodology

The goal of this study is to empirically assess the capacity of parameter-efficient transformer fine-tuning for generating aspect-conditioned peer review feedback. The methodology comprises a systematic pipeline for data preprocessing, prompt engineering, model adaptation using LoRA, experimental design, and evaluation—all implemented to maximize reproducibility and alignment with best practices in neural text generation.

A. Data Preparation and Preprocessing

To tailor the dataset for controlled sequence-to-sequence modeling, we first conducted data cleaning—removing any records with missing or null review text—to ensure each example comprised valid input-output pairs.

Each input prompt was programmatically constructed by concatenating three elements:

- A truncated excerpt from the scientific paper (to a maximum length, preserving the most salient context)
- A summary of previous review sentences for that paper

- The aspect to be addressed (e.g., clarity, motivation, soundness)

A fixed, descriptive template string was used to unify the formatting of all prompts, ensuring the model is always cued with the precise task structure expected at inference.

All input prompts and corresponding gold aspect sentences were tokenized using a pre-trained LongLLaMA tokenizer, with strict truncation and padding to fixed sequence lengths (e.g., 512 for inputs, 64 for outputs). The dataset was then partitioned into training, validation, and test splits, guaranteeing that model selection and hyperparameter tuning were performed without test set leakage.

B. Prompt and Label Construction for Causal LM Fine-Tuning

Because causal language models (like LongLLaMA) learn to generate output by continuing input sequences, each training instance was constructed by concatenating the input prompt and the target aspect sentence. During training, the model was required to autoregressively predict the next token, but loss was only computed on the target (aspect sentence) portion. This was achieved by masking the prompt portion in the labels with a special index (-100), so the model would not be penalized for outputs related to the prompt but would be optimized solely on the aspect sentence.

This prompt engineering and label masking approach follows standard causal LM fine-tuning practices, ensuring that the model learns to produce focused, aspect-grounded critique sentences when conditioned on both the paper context and prior reviews.

C. Parameter-Efficient Model Adaptation via LoRA

Given computational constraints and the large parameter count of the LongLLaMA architecture, Low-Rank Adaptation (LoRA) was used for fine-tuning. LoRA is a lightweight adaptation technique that introduces small trainable low-rank matrices into the attention projection layers (q_proj, v_proj) of a frozen pre-trained transformer. Only these additional parameters are updated during training, significantly reducing memory and compute requirements while maintaining strong adaptation capacity.

Experiments were conducted using two different LoRA ranks (8 and 16), enabling comparison of adaptation capacity versus efficiency trade-offs.

D. Training Configuration and Experiment Logging

All experiments were conducted using the PyTorch and HuggingFace Transformers ecosystem, leveraging mixed-precision (fp16) training for increased throughput on commodity GPUs. The main hyperparameters were set as follows:

- Model: LongLLaMA 3B Instruct
- LoRA Ranks: 8, 16
- Batch size: 8
- Learning rate: 2e-4
- Number of epochs: 1 (for preliminary benchmarking)
- Max input/output sequence length: 512/64 tokens

Training was performed using standard AdamW optimization. Each batch consisted of prompt-target pairs as described above. The loss was accumulated only over the target aspect sentence positions, as ensured by label masking.

All runs, including metrics, losses, hyperparameters, and checkpoints, were tracked in real time using the Weights & Biases platform, with public accessibility for verification and future analysis.

E. Baseline Comparisons

To establish reference points for model evaluation, two baseline methods were implemented:

- TF-IDF + Logistic Regression: Prompts were vectorized and used to train a logistic regression classifier to select likely aspect sentences.
- Embedding-based Nearest Neighbor Retrieval: The most similar aspect sentence to each prompt was selected based on cosine similarity in embedding space.

F. Evaluation Metrics

Model predictions were assessed using industry-standard natural language generation metrics, specifically:

- BLEU: Capturing n-gram precision between generated and reference aspect sentences.
- ROUGE-1/2/L: Measuring recall and overlap with reference sentences at various granularity levels.

Results were computed for the validation and test splits, and were compared both against baseline methods and across hyperparameter variants.

V. Preliminary Results

The evaluation of baseline (zero-shot) performance on both validation and test splits revealed substantial room for improvement, underscoring the difficulty of the aspect-guided review generation task and the challenge posed by the dataset. The results are summarized in Table 1.

Metric	Validation	Test
BLEU	0.00139	0.00106
ROUGE-1	0.09478	0.0965
ROUGE-2	0.00514	0.00507
ROUGE-L	0.07557	0.07692

Table 1 - Baseline performance on validation and test sets

As indicated by the results, the zero-shot model is essentially unable to generate meaningful aspect sentences for the task at hand, producing scores near chance levels for all evaluation metrics. This validates the necessity for parameter-efficient fine-tuning to enable the model to condition on domain-specific prompts and generate relevant outputs.

IV. Discussion

All experiments were conducted using a single NVIDIA GPU, imposing strict limitations on both memory and runtime. Several methods were employed to optimize the process:

- Batch size and mixed-precision (fp16) training were used to maximize throughput.
- LoRA was selected for its parameter-efficiency, reducing VRAM usage compared to full-model finetuning.
- Sequence lengths for both input and output were capped (512/64 tokens) to stay within hardware constraints.
- One-epoch training was adopted for rapid benchmarking, with the understanding that additional epochs may be required for optimal results.

Despite these optimizations, runtime for inference on the validation and test sets remains significant. This bottleneck is mainly due to model size and the sequential nature of autoregressive generation. Increasing batch size further was not feasible due to memory constraints.

At the time of writing, the LoRA-adapted model is still being evaluated. Nevertheless, the rapid decrease in training loss indicates effective learning. Final generation metrics and a more comprehensive benchmarking (including hyperparameter sweeps and multiple LoRA ranks) will be provided as soon as computational resources permit.

REFERENCES

[1] Beltagy, I., Peters, M. E., & Cohan, A. (2020). *Longformer: The long-document transformer* (arXiv:2004.05150). arXiv. https://doi.org/10.48550/arXiv.2004.05150.

[2] Tworkowski, S., Staniszewski, K., Pacek, M., Wu, Y., Michalewski, H., & Miłoś, P. (2023). *Focused Transformer: Contrastive training for context scaling* (arXiv:2307.03170). arXiv. https://doi.org/10.48550/arXiv.2307.03170.

[3] Unsloth Documentation, "Fine-tuning Guide," 2025. [Online]. Available: https://docs.unsloth.ai/get-started/fine-tuning-guide.

[4] Hu, E. J., Shen, Y., Wallis, P., Allen-Zhu, Z., Li, Y., Wang, S., Wang, L., & Chen, W. (2022). *LoRA: Low-Rank Adaptation of Large Language Models*. International Conference on Learning Representations. https://openreview.net/forum?id=nZeVKeeFYf9.

[5] Li, J., Zhao, W. X., Wen, J.-R., & Song, Y. (2019). *Generating long and informative reviews with aspect-aware coarse-to-fine decoding*. In A. Korhonen, D. Traum, & L. Màrquez (Eds.), *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics* (pp. 1969–1979). Association for Computational Linguistics. https://doi.org/10.18653/v1/P19-1190.

[6] Hinton, G. E., Vinyals, O., & Dean, J. (2015). *Distilling the knowledge in a neural network*. arXiv preprint arXiv:1503.02531. https://arxiv.org/abs/1503.02531.

[7] Li, Z., Ji, Y., Meng, R., & He, D. (2025). *Learning from committee: Reasoning distillation from a mixture of teachers with peer-review*. arXiv preprint arXiv:2410.03663. https://arxiv.org/abs/2410.03663.