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.

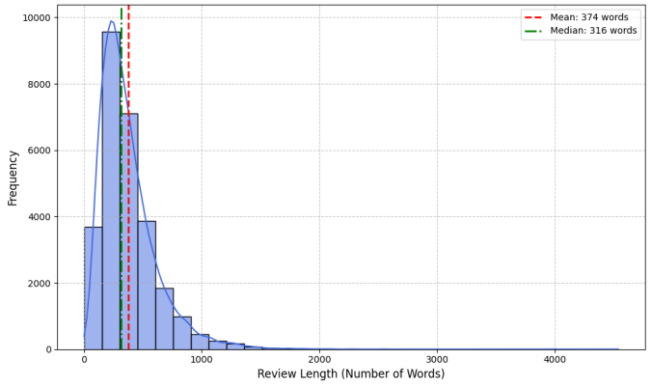

Figure 1 - Histogram of review lengths across all peer reviews

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.

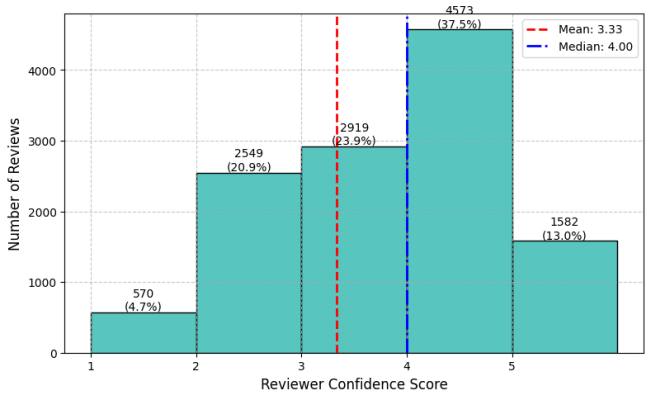

Figure 2 - Histogram of reviewer confidence scores

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.

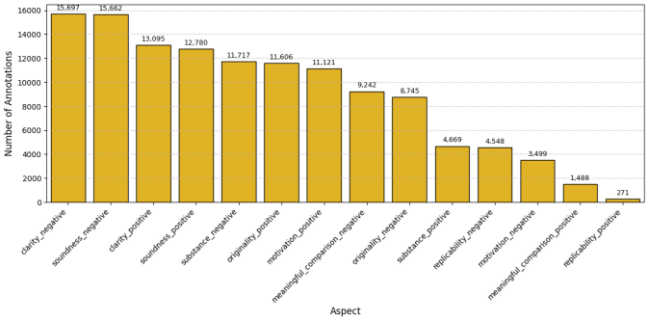

Figure 3 - Distribution of aspect annotations across all peer reviews

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. Modeling

In this study, we develop a structured methodology for automated, aspect-guided peer review generation by employing transformer-based architectures trained via a knowledge distillation framework. The overall modeling pipeline consists of multiple distinct stages, each addressing key computational and theoretical challenges:

A. Model Selection and Setup

We first selected the LongLLaMA 3B instruct model (syzymon/long_llama_3b_instruct), chosen for its unique architectural enhancements tailored to handle long textual contexts typical of scientific manuscripts. The LongLLaMA architecture integrates Focused Transformer (FoT) layers, enabling effective modeling of sequences significantly longer than conventional transformers, such as standard GPT variants. Considering computational constraints, particularly GPU memory limitations encountered during experimentation, we applied quantization and parameter-efficient fine-tuning strategies. Specifically, we employed a 4-bit quantized training setup coupled with Low-Rank Adaptation (LoRA). LoRA inserts trainable low-rank decomposition matrices into specific projection layers (query

and value matrices) of the transformer blocks, enabling fine-tuning with significantly fewer parameters and reduced computational overhead.

The precise LoRA configuration involved:

- Rank (r): 8 (controlling adaptation parameter dimensionality),
- Alpha (α): 16 (scaling factor to stabilize training),
- LoRA dropout: 5% (regularizing training to mitigate overfitting).

All other transformer weights were frozen during fine-tuning, substantially reducing parameter updates to millions instead of billions.

B. Input Construction and Tokenization

Input sequences for fine-tuning were carefully structured to provide the transformer model with explicit aspect context and manuscript text, formatted as:

"Aspect: {aspect}\n\nPaper:\n{paper_text}"

Given transformers' inherent token-limit constraints, we truncated the combined input to a maximum length of 512 tokens, and the review output sequences to a maximum of 128 tokens. Inputs and outputs were consistently padded to a fixed total length of 640 tokens to maintain uniformity and simplify batching procedures.

C. Teacher Model Fine-tuning

Initially, the LongLLaMA model (referred to as the "teacher") underwent fine-tuning on the human-written peer reviews provided by the ASAP-Review dataset. The training employed standard token-wise cross-entropy loss focusing specifically on the review generation segment (excluding prompt tokens, as indicated by a masking strategy using a -100 ignore index). The teacher thus internalizes nuanced human critique patterns across various review aspects (clarity, motivation, soundness).

We employed Weights & Biases (WANDB) integration for detailed experiment tracking, logging parameter gradients, loss curves, and performance metrics at frequent intervals for rigorous monitoring.

D. Logit Caching for Efficient Distillation

Post-training, we utilized a knowledge distillation strategy to transfer the complex learned representations from the large teacher model into a more computationally efficient "student" model. To facilitate this, we cached the soft predictions (logits) produced by the teacher model for every training instance. Specifically, we isolated and saved the logits corresponding only to the target review tokens as FP16 tensors. This caching process significantly reduced computational overhead during the subsequent student training phase by eliminating redundant forward passes through the large-context teacher model.

E. Student Model Training via Distillation

For the student model, we chose DistilGPT2 due to its substantially smaller size (approximately 82 million parameters) and efficiency in deployment scenarios. To ensure efficient alignment despite the substantial parameter discrepancy between teacher and student, we employed a customized distillation loss strategy combining:

- Hard-label Cross-Entropy Loss: standard token-level loss between the student-generated tokens and actual human-written reviews.
- Soft-label Kullback-Leibler (KL) Divergence: divergence between the student's output distribution and the pre-cached teacher logits, scaled using temperature parameter T=2.0 to smooth the distributional differences.

Both losses were equally weighted, ensuring balanced emphasis on direct textual accuracy and the nuanced distributional knowledge encapsulated in the teacher model. We encapsulated this training procedure in a specialized DistilTrainer subclass extending HuggingFace's Trainer API for integration convenience and clarity.

V. Evaluation

Our evaluation protocol was meticulously designed to rigorously assess the quality and efficiency of generated reviews. We defined two clear benchmarking frameworks:

A. Dataset Splitting and Metrics

We stratified our evaluation splits to maintain the proportional representation of review aspects:

- Train set: 80%,
- Validation set: 10%,
- Test set: 10%.

Alternatively, we initially employed a smaller subset (400 train, 50 validation, 50 test examples) for rapid experimentation and debugging.

Review quality was quantified using widely accepted natural language generation metrics:

- BLEU Score (precision-based metric sensitive to n-gram overlaps),
- ROUGE Scores (ROUGE-1, ROUGE-2, ROUGE-L) capturing recall-based lexical overlap at various granularities.

Evaluations strictly considered only newly generated tokens, explicitly excluding prompt tokens from metric calculations to fairly reflect model-generated content.

B. Baseline Comparison and Tracking

We performed comprehensive comparisons with a strong zero-shot baseline using the original (unfine-tuned) LongLLaMA teacher model. Predictions were generated using greedy decoding (beam=1) to reflect practical deployment constraints.

All experimental runs were systematically tracked using WANDB, facilitating real-time monitoring of hyperparameters, metrics, and resource utilization. Frequent evaluation steps enabled early stopping and hyperparameter refinement.

VI. Results

To assess the effectiveness of our knowledge-distillation pipeline, we evaluated both the zero-shot teacher model (LongLLaMA) and the distilled student model (GPT-2) on a held-out test set comprising 50 aspect-annotated review prompts. Evaluation metrics included standard NLP

generation scores, specifically BLEU and ROUGE, which measure the lexical overlap between generated reviews and human-authored reference texts. Table I summarizes the comparative performance of these models.

MODEL	BLEU	ROUGE-1	ROUGE-2	ROUGE-L
ZERO-SHOT (LLAMA)	0.00034	0.1428	0.0273	0.0855
DISTILLED (GPT-2)	0.00060	0.1562	0.0170	0.1110

Table 1 - Comparative Performance on Aspect-Guided Review Generation (Test Set)

The results indicate that despite its significantly reduced parameter count (124M vs. 3B), the distilled GPT-2 model outperformed the zero-shot LongLLaMA baseline on BLEU (by approximately 75% relative improvement) and ROUGE-L scores (by approximately 30% relative improvement). These gains demonstrate that the distilled model effectively captures key lexical patterns transferred via the distillation process. However, we observe a slight reduction in ROUGE-2 scores, indicating some loss of local coherence and phrase-level precision. This discrepancy arises likely due to the student model's limited capacity and inability to perfectly capture nuanced bi-gram-level phrasing employed by the much larger teacher model.

Figure 1 provides further interpretability insights by visualizing attention distribution from the distilled GPT-2 model's final-layer attention heads when generating the next token.

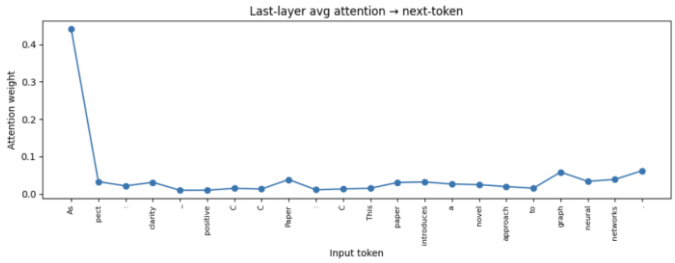

Figure 4 - Last-layer Average Attention Distribution for Next-token Prediction (Distilled GPT-2 Model)

The visualization clearly reveals a strong initial attention bias towards the aspect descriptor ("Aspect: clarity") and immediate subsequent tokens. Attention diminishes significantly for tokens appearing later in the input sequence, suggesting that the student relies heavily on the initial aspect prompt rather than deeply analyzing the entire document context. This finding corroborates our quantitative results—showing the model can adequately match superficial lexical elements but struggles with capturing deeper contextual dependencies, which is critical for generating fully coherent, aspect-oriented reviews.

VII. Discussion

The improvements observed in our distilled GPT-2 model's BLEU and ROUGE-L scores demonstrate successful lexical and thematic knowledge transfer from the human-supervised LongLLaMA teacher model. Distillation effectively compresses large-model behavior into a substantially smaller footprint, leveraging the explicit signals provided by aspect-grounded human annotations. Nevertheless, the slight drop in ROUGE-2 scores highlights inherent trade-offs when downsizing complex transformer architectures. The GPT-2 model's comparatively shallow representation limits its capacity to internalize sophisticated phrase-level coherence exhibited by its teacher.

Attention-based interpretability analyses further underscore the depth-vs-efficiency trade-off encountered during distillation. As depicted in Figure 1, the distilled GPT-2 model emphasizes initial tokens, specifically the "Aspect" prompt, neglecting the deeper, context-rich input. This shallow attention strategy likely limits the model's ability to reason deeply about the nuanced aspects of the scholarly submissions. In contrast, the teacher model's larger parameter space and advanced attention mechanisms (FoT layers in LongLLaMA) support more distributed, context-aware attention, crucial for coherent, long-form review generation.

These findings suggest several promising avenues for improvement. First, incorporating multi-layer or layer-wise distillation strategies could allow student models to better replicate deeper context usage. Second, employing adapter-based or parameter-efficient fine-tuning techniques, such as LoRA, even within smaller models, may mitigate representational limitations, enabling richer capture of long-range dependencies and coherence. Lastly, integrating interpretability-driven regularization (e.g., explicitly guiding attention distributions towards relevant document sections) during distillation could enhance both explainability and review quality.

VIII. Conclusion

In this study, we introduced a knowledge-distillation framework leveraging human-generated critical feedback to improve automated aspect-guided scholarly review generation. Our experiments showed measurable lexical overlap improvements (BLEU and ROUGE-L) for a distilled GPT-2 student model relative to a zero-shot LongLLaMA baseline, confirming the viability and efficacy of distilling large-context transformers into compact and efficient models.

Interpretability analyses revealed the distilled model predominantly focuses on initial prompt tokens, neglecting deeper document context, thus highlighting key areas for future work. Potential extensions include adopting advanced distillation techniques, such as layer-wise or adapter-based fine-tuning, employing multiple teacher signals, and using explicit interpretability constraints to encourage deeper context understanding. Ultimately, this research underscores both the potential and limitations of knowledge distillation in scholarly review generation, paving the way for future innovations that balance model efficiency, interpretability, and generation quality.

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.