

Review Arcade: On the Human Alignment and Gameability of LLM Reviews

Hans Ole Hatzel^{1*}, Sebastian Steindl^{3*}, Jan Strich^{1,2*}

¹Language Technology Group, University of Hamburg, Germany

²Hub of Computing and Data Science (HCDS), University of Hamburg, Germany

³OTH Amberg-Weiden, Germany

*Equal contributions, order decided by coin toss.

Correspondence:

{first_name}.{last_name}@uni-hamburg.de, s.steindl@oth-aw.de

Abstract

LLM-generated reviews for scientific papers are gaining considerable traction and are even being officially piloted by major conferences. We have to assume that not only reviewers are using LLM-assistance, but also that authors use LLMs to revise their papers before submitting. In this work, we perform empirical experiments on papers from the 2025 ACL Rolling Review (ARR) to evaluate LLM reviews from both the author and the reviewer perspective. First, we identify a limited alignment of LLM reviews with human ones. In the best-case scenario, the alignment is reasonable. However, we also find that LLM-human alignment varies substantially across prompts and models. Finally, we investigate the scenario in which the author uses an iterative draft-revise workflow to improve the submission according to the LLM review. We find that this “gaming” of LLM reviews can be effective in specific scenarios, leading to a statistically significant increase of overall scores for up to 35% of papers. We publish our code.¹

1 Introduction

LLMs are becoming ubiquitous in academic writing. They are not only powerful tools for correcting grammar and syntax, but can also be used as a source of ad-hoc feedback to a manuscript (Kobak et al., 2025; Wu et al., 2026). Consequently, authors are more likely to revise their papers using LLMs. At the same time, LLM reviews are being studied as a possible way to reduce the overload of the peer review system caused by the strong increase in submissions. (Wei et al., 2025; Choi et al., 2026). Beyond potential future official practice, current research indicates LLM-usage in the peer-review process. Liang et al. (2024) establishes that across most of their analyzed conferences and journals, 7-15% of reviews show AI usage beyond simple grammar correction. Given this, authors may assume that their submission might be LLM-reviewed

¹GitHub Repository

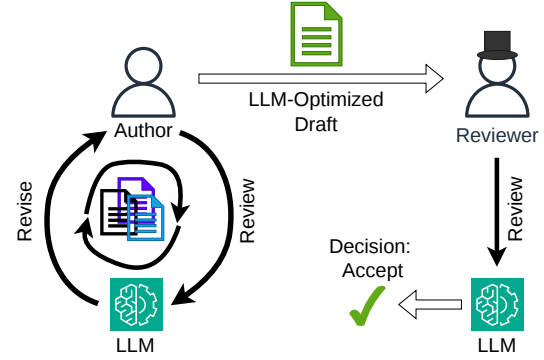


Figure 1: Visualization of the peer-review process if both author and reviewer rely on LLMs.

and are thus encouraged to optimize their submission accordingly. Thus, the current situation may culminate in both submission and review becoming heavily LLM-reliant (Fig. 1). In this context, we should consider Goodhart’s law (Goodhart, 1975): “When a measure becomes a target, it ceases to be a good measure.” Strathern (1997). Applied here, once authors optimize papers specifically for LLM reviews, they may no longer reliably reflect paper quality, even if they initially did.

In this paper, we study the alignment of LLM and human reviews on 984 real ARR submissions for ACL 2025. We evaluate this across multiple models (open-weight and proprietary), prompts, and runs. Additionally, we simulate an Iterative Submission Improvement (ISI) workflow, where authors optimize their submissions according to LLM reviews.

We are guided by three research questions (RQs):

- **LLM Review Validity (RQ1):** Can LLMs produce reviews that are sufficiently aligned with human reviews?
- **LLM Review Stability (RQ2):** Are LLM reviews for a given submission consistent across models, prompts, and repeated runs?

- **LLM Review Gaming (RQ3):** Can LLM reviews be “gamed” by automated, iterative edits that are informed by LLM reviews and aim to improve review scores?

Our main contributions are: (i) The first large-scale empirical evaluation of LLM reviews for ARR submissions, (ii) an investigation of an automated paper-editing scheme as an adversarial attack on automated reviews, and (iii) a taxonomy for such edits grounded in prior literature.

2 Background and Related Work

Automated Peer-Review. Approaches to automated peer-review and the analysis of LLM reviews have increasingly gained traction, with researchers benchmarking language models on the task, and proposing systems to improve performance and explore the properties of LLM reviews. An early example in the LLM era is (Zhou et al., 2024), who systematically evaluated LLMs on peer-review tasks. Various authors have since suggested improvements using thinking processes or agentic approaches to the task (Jin et al., 2024; Zhu et al., 2025; Idahl and Ahmadi, 2025; Bougie and Watanabe, 2025; Sahu et al., 2025).

In terms of real-world applications, Biswas et al. (2026) recently evaluated LLM reviewers at scale for the AAAI conference and found them to be perceived favorably by authors and other reviewers alike. Taking the stance that human reviews should be considered the gold standard, one of the main metrics for the usability of LLM reviews becomes their alignment with the human reviews. One reason why the survey in Biswas et al. (2026) might have shown LLM reviews to be favorable is the high variance in human review quality.

Reliability of Human Reviews. There is a limited range of prior work considering the reliability of human reviews. Notably, acceptance decisions are generally not determined by a simple score threshold; instead, meta reviewers and program chairs consider many factors, such as outliers in review scores and their justification, or simply the number of competing papers in a given track (Cicchetti, 1991). The NeurIPS conference ran an acceptance experiment that simulated this entire decision process (Beygelzimer et al., 2021, 2023), finding that approximately half the papers accepted by one committee were rejected by the other. Conversely, they find that a given paper had a roughly 15% chance of being accepted after being rejected by the first

committee. In terms of review scores, the deviation is much easier to quantify, given that there are typically multiple independent reviews of the same paper. Baumann et al. (2026) report a Pearson correlation of 0.14 across human reviewers, while (Cortes and Lawrence, 2021) find a Pearson correlation of 0.55 in their data after calibrating for cross-reviewer scale interpretation using a Gaussian model.

Peer-Review Datasets. PeerRead (Kang et al., 2018) was one of the first peer-review datasets. They collect likely rejects from arXiv while relying on reviews of accepted papers from reviewing platforms, including OpenReview. Many datasets primarily recruit their reviews from accepted papers, thereby introducing biases. In a more recent example, NLPeer (Dycke et al., 2023) made use of a clear data collection scheme requiring opt-ins from reviewers and authors alike (Dycke et al., 2022).

Metrics for Automated Reviews. There is a multitude of metrics being used to measure the quality of automated reviews. Prior work uses, e.g., accuracy and correlational measures (Zhou et al., 2024; Idahl and Ahmadi, 2025), AUC, FPR and FNR (Lu et al., 2026), and MAE Zhu et al. (2025).

We report MAE and Pearson correlation, as well as an LLM-judge measuring semantic overlap, as the primary metrics for measuring the LLM-human alignment in this paper. Further, we distinguish between best-match and overall correlations: for best match we only calculate correlations with the best matching review, in terms of the *Overall* score.

Concurrent work. Kim et al. (2026) conduct a human evaluation of review quality, where experts assess human and LLM-generated reviews along three dimensions. They find that LLM-generated reviews can surpass human reviews in perceived quality, while still exhibiting systematic limitations. In a related position paper, Baumann et al. (2026) show that *paper laundering*, iteratively prompting LLMs to improve a manuscript based on LLM-generated reviews, can substantially increase review scores. Although framed as inducing only superficial, cosmetic edits, their prompting strategy does not enforce such constraints and may instead encourage substantial revisions. Motivated by this, we conduct a more principled evaluation of paper laundering in an iterative setting and further quantify LLM-induced semantic changes using an taxonomy.

3 Method

Today, real-world reviewers often employ off-the-shelf models to aid in their reviewing (Liang et al., 2024) and official usage aims for zero data retention by using open-weight models offline or API settings. Our setup aims to align itself with this real-world usage of LLMs in the context of peer review. As such, we evaluate with both open-weight and closed-weight models. However, we do not employ sophisticated agentic workflows, which might increase the quality of individual reviews.

3.1 Problem Statement

Our work mainly focuses on using an LLM $\mathcal{M}$, prompted with instructions ρ , to generate a review r for the submission s :

$$r = f(M, \rho, s). \quad (1)$$

Then, we evaluate the quality of r by calculating its alignment with the ground-truth, human-written review $\hat{r}$ using the evaluation function $h(\hat{r}, r)$. Concretely, $h(\hat{r}, r)$ can be instantiated as a measurement of correlation on the predicted scores, or as an LLM-judge $\mathcal{J}$ that measures content similarity across strengths and weaknesses of s identified in r and $\hat{r}$.

Moreover, we consider the scenario in which the author optimizes their submission s by iteratively adapting it based on an LLM review:

$$s^{i+1} = \mu(s^i, f(M', \rho', s^i)). \quad (2)$$

We test the fully-automated scenario in which μ is also a call to an LLM, prompted to update the submission to address the review.

3.2 Automated Review Framework

In this work, we want to evaluate if LLM reviews are closely aligned with human reviews (RQ1) and if the LLM reviews are consistent across different models and prompts (RQ2). To this end, we craft five review prompts that are increasingly tailored to the specific ARR review dataset:

- **simple**: A minimal prompt asking simply to review and specifying output format.
- **default**: Drafted by the authors to specify target venue and acceptance rate.
- **ai_generated**: An LLM-generated prompt for reviewing submissions to a top-tier Machine Learning conference.

- **acl**: Adapted from **ai_generated** to include the specific guidelines from the ARR.
- **acl_senior**: As **acl**, but with the persona of a senior, expert reviewer.

For a full list of all prompts, see Appendix F.

3.3 Iterative Submission Improvement

For RQ3, we consider different styles of Iterative Submission Improvement (ISI). Optimizing a submission solely to target automated reviews is what we describe as “gaming” LLM reviews. ISI describes the iterative loop, depicted in Fig. 2, in which an author generates a review r for their submission s^i with an LLM and uses this to inform an editing function μ to improve their submission, creating s^{i+1} . We iteratively apply ISI for ten iterations. Since it is impossible to perfectly predict an accept/reject decision, we do not try to predict if a paper would be accepted or rejected, and instead focus on improvements of the *Overall* score. Specifically, we focus on three settings: *constrained*, *default*, *adversarial*².

In the *constrained* setting, the author prohibits substantive changes and allows only superficial, cosmetic edits in response to the review. This tests whether the “paper laundering” of Baumann et al. (2026) can shift LLM review recommendations from reject to accept. However, their prompt does not strictly enforce cosmetic-only edits and may even encourage more fundamental changes.

Therefore, in our *default* setting, we use a prompt that is heavily inspired by the editing prompt used in Baumann et al. (2026), but removes instructions that could lead to non-cosmetic changes. We call this *default* as it neither prohibits nor actively allows profound changes. Lastly, in the *adversarial* setting, we simulate an author who actively encourages editing to get the paper accepted at any cost, even if that means, e.g., fabricating results.

3.4 Taxonomy of Edits

To better understand what type of edits are performed to increase the scores in the LLM review, we introduce a taxonomy of paper edits. We ground our taxonomy in the work of Yang et al. (2017), who propose a taxonomy for edit types on Wikipedia. We adapt their taxonomy to fit our scenario of paper edits for an ARR submission. The taxonomy is presented in Tab. 3 in the Appendix. For the *constrained* and *default* edit settings, we use the

²All prompts given in Appendix F.

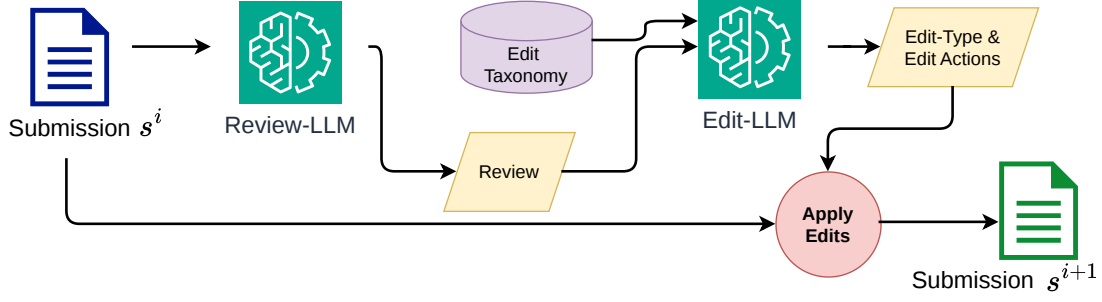


Figure 2: The ISI pipeline is iteratively applied to improve upon paper drafts.

same set of allowed edit-types. These focus on keeping the content of the submission intact and not requiring new experiments, such as simplifying or clarifying. For the *adversarial* setting, we add another set of edit types that focus on “gaming” the LLM review, such as hallucinating evidence and fabricating better results.

4 Experimental Setup

4.1 Dataset and Preprocessing

In ARR, the main ACL reviewing platform, reviewers assign 9-point ratings (1 to 5 in 0.5 steps) across four categories: *Soundness*, *Excitement*, *Reproducibility*, and *Overall*. Reviews and author responses are discussed before the Area Chair writes a meta-review summarizing them. Final acceptance decisions are made by the program committee based on reviews and meta-reviews. We only use the *Overall* score as it is the most representative metric.

Existing research in the space of ARR reviews relies on very few or no rejected papers. This potentially introduces a positivity bias in systems developed for this data. We perform stratified subsampling on the NLPeer dataset (Dyck et al., 2023) in a fashion similar to Sahu et al. (2025) to define a dataset with 984 papers. We retain all rejected papers with reviews from NLPeer. All accepted papers in our dataset were accepted to ACL 2025. Rejected papers make up roughly one third of our dataset³. To prepare our documents for LLM processing, we process them using the OCR model o1mOCR-2-7B-1025 in conjunction with the OlmOCR pipeline (Poznanski et al., 2025). This outputs Markdown versions of the papers. Tables are retained and presented as Markdown to the models while figures are only represented by captions

³While this does not correspond to the acceptance rate at ARR venues, it is suited for our experiments.

provided in the original paper. This setup enables us to isolate the models’ reviewing capabilities from their PDF reading abilities and simulates the application of LLMs in larger systems where typically a content extraction step is performed for PDFs (Blecher et al., 2024). We filter out papers longer than $> 130,000$ subword tokens to account for context window limitations, long appendices, and potential extraction errors, and also exclude papers with missing review text or incorrectly extracted paper text.

Dataset Statistics. In our subsampled dataset, humans show a rather low overall correlation of 0.312 for the *Overall* score across reviews of the same paper. Similar magnitudes are reported by Baumann et al. (2026), who find correlations of 0.137 in a subsample and 0.180 across all ICLR reviews. We also find that the correlation is substantially higher for rejected papers (0.408) than for accepted papers (0.210), suggesting that reviewers are more likely to agree if a submission is poor than good. This is consistent with prior literature (Cortes and Lawrence, 2021). The underrepresentation of rejected papers in ARR-related studies, arising from the paper collection process, is therefore particularly concerning.

In Figure 3, we illustrate that papers in the rejected split of our dataset are, on average, much shorter. They show an almost uniform distribution from 4,000 to 9,000 tokens, while accepted papers show a clear increase around the 7,500 token mark. We hypothesize two causes: (1) accepted papers are often more comprehensive and near the page limit, leading to more concentrated contributions; and (2) shorter papers are less likely to be accepted, resulting in their overrepresentation in our dataset.

On average, papers in the accepted split have 2.0 reviews, with a standard deviation of .7, while the rejected split has just over 1.1 reviews per paper,

with a standard deviation of 0.3. This imbalance is likely a result of the additional approval process for reviews of rejected papers. Overall, we observe a clear difference in the accepted and rejected groups. For this reason, our further analysis will make an effort to explicitly obtain results for each subset.

4.2 Models

Authors and reviewers might use a variety of LLMs. Therefore, we select six models, covering model sizes as well as three open- and two closed-weight models. Specifically, we use Qwen-3.6-35B (Yang et al., 2025), Gemma-3-27B (Team et al., 2025), Llama-3.3-70B (Grattafiori et al., 2024), GPT-5.4-mini, and GPT-5.4.⁴

4.3 Experimental Design

We design three main experiments to answer our RQs. First, we generate one review for each prompt and model, and repeat this twice, for a total of three reviews. This allows us to measure the alignment of human and LLM reviews with regard to their scores and content, and their stability (RQ1, RQ2). Focusing on the *Overall score*, we measure the mean absolute error (MAE) against the mean of all human reviews. We use Pearson’s r to measure the correlation to the best match, i.e., to the human review with the lowest distance. We report these metrics for both the best performing prompt (in terms of Pearson- r on the combined split) and the average performance across all prompts.

To assess semantic alignment between LLM and human reviews, we use an LLM judge to identify which human-stated strengths and weaknesses are reflected in the LLM review. This recall-style metric provides information beyond review scores. We provide the human-performance by comparing against all other humans as well as a naive baseline that constantly predicts the mid-point from the rating scale⁵. Note that for the rejected split, due to a lack of examples with multiple human reviews, the MAE and r of the humans are being calculated with only 26 papers. For the Combined split, we macro-average across accepted and rejected performance (in Fisher- z space for the correlation).

In the second experiment, we investigate if the papers can be iteratively adapted based on the LLM review to increase their scores. We test this with a maximum of 10 iterations and with three different editing prompts, representing different levels

⁴All models are used in their instruction-tuned variants.

⁵For the latter no correlation calculations are possible.

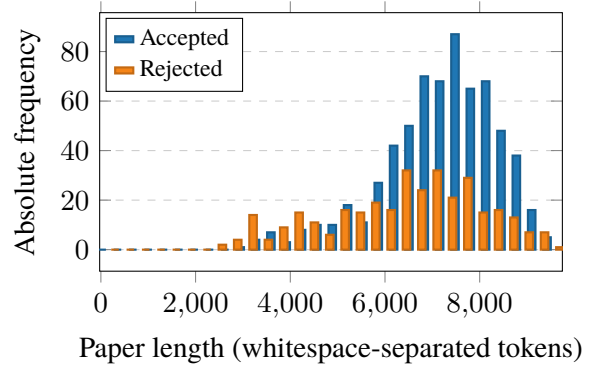


Figure 3: Length distribution of the papers considered in this study. Grouped in 30 buckets (each ~320 tokens).

of changes, from superficial edits (*constrained*) to substantial changes including fabricated evidence (*adversarial*). We also include a prompt that is heavily based on the one used by Baumann et al. (2026), and to which we refer as *default*. We measure this effect in terms of the percentage of papers with an increased score after n iterations. As a baseline, we repeat the prediction for the initial, unedited submission also ten times.

5 Results and Discussion

5.1 LLM Review Validity (RQ1)

Alignment to Human Review Scores. First, we test the validity of LLM reviews as measured by their alignment with the human ratings. For the *Combined* split in Table 1 including both the accepted and rejected papers, we can observe that the LLMs fail to match human judgments in terms of MAE. GPT-5.4-mini and GPT-5.4 are the best performing LLMs with an MAE of around 0.7, compared to the human 0.17. Notably, the naive constant prediction baseline slightly outperforms the best LLM with an MAE of 0.64. In terms of correlation, the models come much closer to human performance with GPT-5.4 reaching a correlation of 0.276. However, one must consider that the human-human correlation of 0.312 indicates low agreement even between humans, which aligns with prior work (Baumann et al., 2026).

The results in Table 1 also show a pronounced performance difference between the accepted and rejected split for all models, most prominently for Gemma-3. The human agreement is much higher for the accepted papers, with the best-match r being nearly twice as high (0.41 vs. 0.21). We hypothesize that this performance difference is explained

Model	Prompt	Combined		Accepted Split		Rejected Split	
		MAE ↓	Best Match r ↑	MAE ↓	Best Match r ↑	MAE ↓	Best Match r ↑
Gemma-3-27B	All	0.97 ± 0.32	0.146 ± 0.06	0.83 ± 0.23	0.246 ± 0.10	1.12 ± 0.46	0.041 ± 0.04
	Best	0.89 ± 0.01	0.205 ± 0.02	0.73 ± 0.00	0.367 ± 0.02	1.05 ± 0.02	0.031 ± 0.01
Qwen-3.6-35B	All	0.73 ± 0.12	0.189 ± 0.03	0.76 ± 0.22	0.208 ± 0.05	0.70 ± 0.17	0.169 ± 0.02
	Best	0.81 ± 0.01	0.217 ± 0.04	<u>0.63 ± 0.01</u>	0.251 ± 0.00	1.00 ± 0.02	0.183 ± 0.07
Llama-3.3-70B	All	1.20 ± 0.15	0.103 ± 0.08	0.88 ± 0.10	0.090 ± 0.13	1.52 ± 0.20	0.116 ± 0.06
	Best	0.95 ± 0.01	<u>0.234 ± 0.02</u>	0.73 ± 0.01	0.308 ± 0.01	1.16 ± 0.01	0.157 ± 0.03
GPT-5.4-mini	All	0.75 ± 0.11	0.124 ± 0.06	0.89 ± 0.25	0.090 ± 0.12	0.62 ± 0.11	0.157 ± 0.05
	Best	0.70	0.229	0.58	0.278	0.81	0.178
GPT-5.4	All	0.73 ± 0.13	0.180 ± 0.07	0.82 ± 0.27	0.167 ± 0.11	<u>0.63 ± 0.09</u>	<u>0.194 ± 0.06</u>
	Best	<u>0.71</u>	0.276	0.63	<u>0.317</u>	0.80	0.233
Human		0.17	0.312	0.30	0.210	0.04	0.408
Baseline ($\hat{y} := 2.5$)		0.64	—	0.75	—	0.53	—

Table 1: Results across models and prompt setups on the Overall dimension. MAE and Best Match Pearson- r over runs (mean $\pm$ std). **Bold**: best in column; underlined: second best. Performance on the combined split is given as macro average across the two splits.

by the fact that accepted papers meet a high bar in terms of minimum quality and that it is hard to differentiate across them. This aligns with the finding by Cortes and Lawrence (2021), that the 2014 NeurIPS review process was good at identifying poor papers, but bad at identifying good papers. Overall, while Pearson r is, depending on the split, competitive with human evaluation, we observe that at least in terms of MAE the models are not competitive with human reviews.

For realistic, practical applications, the macro-average *Combined* performance is more indicative, since it is unknown at submission time which split the paper would be part of. Here we see that individual prompts perform very well but note that Qwen delivers the most robust performance across splits, tying with GPT-5.4 in terms of prompt-averaged MAE but slightly outperforming it in terms of prompt-averaged best-match Pearson- r .

Content-wise Alignment with Human Reviews. Besides the scores, we also evaluate how similar the LLM reviews are to the human reviews in their content. We report the strengths-recall s_recall and weaknesses-recall w_recall , which represent the fraction of strengths and weaknesses that appear in both the human and LLM reviews as presented in Fig. 4. For the strengths, Gemma-3 achieves the best overall s_recall , with roughly 0.59 on the accepted, and 0.48 on the rejected split. For the weaknesses, GPT-5.4-mini has the highest recall, with roughly 0.41 and 0.44 on the respective splits. We observe that, in general, the recall is higher for strengths than for weaknesses. Especially for the

strengths, our results also indicate that the recall can differ between the accepted and rejected split. **Are LLM Reviews Valid?** Overall, the results indicate that, in a select best-case scenario, LLM-review scores show good alignment with human judgments, at least in terms of correlation. In this setting, model-to-model agreement is comparable to human-to-human alignment. However, this behavior does not consistently transfer to real-world conditions where the acceptance decisions are not known a priori. Across splits, no single setup is consistently superior. Because it is hard to calibrate LLMs to align with human reviews, we reach a mixed conclusion regarding RQ1: LLMs can be reviewers in some scenarios, but not universally.

5.2 LLM Review Stability (RQ2)

Stability Across Prompts & Models Importantly, we observe a considerable variance across reviewing prompts. For example, GPT-5.4-mini on the accepted split, which had the best MAE in its best setting, has the worst MAE when averaging across prompts (0.89). This trend holds across all models we tested. Crucially, as Figure 5 shows, there is no clear trend as to which prompt leads to best performance, neither across models, nor within the same model across the accepted and rejected split. The models appear sensitive to prompt variations, alternating between overly permissive and overly restrictive behavior. This might explain the interesting observation that the overly simple one-liner prompt `simple` achieved remarkably good performance, suggesting that sophisticated prompting may not

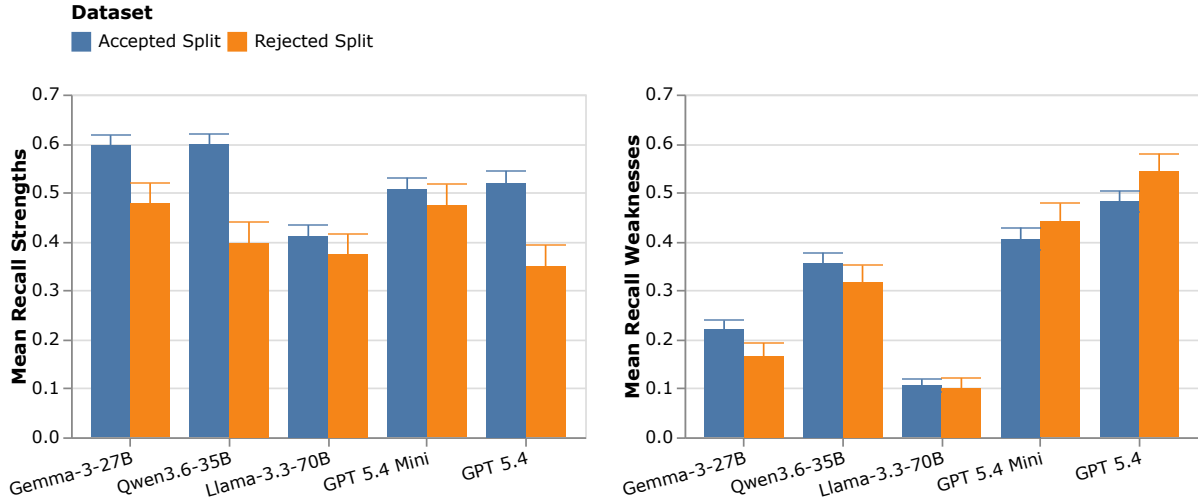


Figure 4: Mean Recall of Strengths and Weaknesses for each of the best runs for each model.

yield improvements on our tasks.

Stability Across Repeated Runs If we perform multiple runs using the same paper and prompt at temperature 1.0, we can observe very low standard deviations of around 0.02 across both MAE and Pearson- r , whereas the deviation is much larger (up to around 0.25 MAE) for runs across prompts of the same model. In our experiments with three model invocations using the same model (see Tab. 4), prompt, and submission, we see that for 36.9% of papers, at least one out of three runs gives a different score than the others and for 20% this delta is > 0.5 . Therefore, we argue that LLM-reviews are generally too instable across repeated runs to be reliable.

Are LLM Reviews Stable? Given the considerable instability of LLM reviews across prompts and models, and even repeated runs, RQ2 can be answered in the negative. This is clearly illustrated in Fig. 5, where different prompts lead to substantially different results across models.

5.3 Gaming LLM Reviews (RQ3)

Based on the results for RQ1 and RQ2, we see that for our experiment on gaming LLM reviews, Qwen-3.6 and GPT-5.4 are best suited. Due to cost considerations and its consistent performance across prompts, we chose Qwen-3.6 for our subsequent experiments. As we expect edits to only drastically improve a small to medium portion of paper scores, we perform rigorous statistical significance tests. The details are given in Appendix D. We report p values and Cohen’s d to account for the large sample size in the dataset and to complement significance testing with an effect-size measure in Table 2. Ef-

fect sizes are interpreted following established rules of thumb in the literature (Cohen, 1992).

Constrained Rewriting In this setup, the prompt explicitly forbids the LLM from making any profound changes to the context. It only allows superficial edits to address the initial review. We find this leads to a statistically significant increase in paper scores in the LLM reviews after 10 review-and-edit loops, compared to the LLM reviews before any changes. We find that roughly 36% of the papers improve, 42% remain at their initial score, and 22% of scores decrease. The effect size for this setup, in terms of Cohen’s d , is considered small to medium (Cohen, 1992).

Default Rewriting The *default* rewriting shows similar numbers for the score changes as the *constrained* editing. However, the results are not statistically significant and show very small effect sizes.

Adversarial Rewriting Lastly, we tested the *adversarial* rewriting, where the LLM is explicitly allowed to make changes it deems helpful for acceptance, including fabricating evidence and factual misrepresentations. In this setup, our data also shows improvements across edit iterations. However, the effect sizes are weaker than in the *constrained* setting. This is surprising, since we expected that, e.g., fabricating results should lead to a large increase in review scores. In fact, we find that when it comes to edit types (as per the taxonomy introduced in Section 3.4), the adversarial prompt almost exclusively turns to the *Methodological-Augmentation* edit type. The default and constrained setups, on the other hand, largely rely on the clarification edit type, with the constrained setup

also making frequent use of the *Refactoring* edit type. See Figure 6 in Appendix C for a full breakdown of the edit types across prompts.

We hypothesize that we did not observe a substantial increase in scores in the adversarial setup for two main reasons. First, methodological edits might introduce inconsistencies within the submission, which could be penalized by the following LLM review. Second, the LLM’s guardrails might lead it to rarely confabulate substantial evidence, which is supported by the fact that *Methodological Augmentation* is the most prevalent edit type in the adversarial setup, even if more aggressive edit types (such as Factual-Optimization or Hallucinated-Evidence) were available in the taxonomy.

Are LLM Reviews Gamable? Yes, in specific scenarios, our ISI pipeline can iteratively improve the scores of papers when it comes to LLM reviews. In the constrained setup, 35% of papers improved after 10 rounds of edits, but this improvement also carried a risk of score regressions, with 22% of papers seeing a decrease in their score. Whether this improvement in scores is associated with a substantive improvement in the paper or truly a case of gaming the LLM-reviewer is harder to answer. *Clarifications* and *Copy-Editing* may not produce substantial improvements to the core of a paper, indicating that gaming is taking place; on the other hand, *Refactoring* is an edit choice frequently made by this best-performing approach, an edit that can result in substantial restructuring of a paper, albeit with limited content changes. Ultimately, whether we consider this gaming of LLM reviewers depends on our trust in human reviewers to look beyond surface-level improvements in the papers.

6 Conclusion

Our results show that human-human correlation in review scores still surpasses the LLM-human alignment. Naively prompted LLMs are instable in their reviews and not yet generally reliable as peer-reviewers. We show that in specific scenarios, current models are able to self-improve papers using superficial edits to improve LLM-judge scores. In this setup, it is feasible to use automated rewriting to push papers past the acceptance threshold in LLM-reliant peer-review. Unlike Baumann et al. (2026), we do not see this effect in prompts that have little guidance.

Interestingly, when allowed to fabricate evidence, our ISI pipeline did not significantly improve pa-

Setting	Outcomes (%)			<i>p</i>	<i>d</i>
	Worse	Equal	Better		
Baseline	28.15	44.72	27.13	.795	-0.03
<i>Reject</i>	29.27	45.12	25.61	.882	-0.07
<i>Accept</i>	27.59	44.51	27.90	.567	-0.01
Default	25.30	44.11	30.59	.012	0.07
<i>Reject</i>	25.91	46.04	28.05	.379	0.02
<i>Accept</i>	25.00	43.14	31.86	.006	0.10
Adversarial	28.48	35.91	35.61	.004	0.10
<i>Reject</i>	22.92	37.50	39.58	< .001	0.24
<i>Accept</i>	31.67	35.00	33.33	.254	0.03
Constrained	22.36	41.67	35.98	< .001	0.20
<i>Reject</i>	18.90	38.72	42.38	< .001	0.32
<i>Accept</i>	24.09	43.14	32.77	< .001	0.13

Table 2: Distribution of model responses across prompt settings (Baseline, Default, Adversarial, and Constrained), reported as percentages of Decrease, Equal, and Increase outcomes after 10 iterations. Paired t-tests with *t*/*p*-values and effect sizes (Cohen’s *d*) for *t*₀ and *t*₁₀. **Bold:** statistically significant results with *p* < .001.

pers across the entire dataset. We argue that this can be explained due to model guardrails avoiding fabrication of evidence or these edits introducing inconsistencies within the revised submission. While peer review processes are, in reality, more complex than a simple score cutoff, our findings highlights a potential vulnerability in the peer-review process as LLM usage increases. We cannot yet confirm if these iterative improvements would translate to humans accepting the papers despite not having profound improvements.

We urge the community to employ extreme caution when approaching the subject of automated reviews. Given Goodhart’s law, even when LLM reviews currently show decent alignment with human reviews, they might cease to be a good measure of submission quality.

We call on future work to extend the evaluation of automated peer review with all its strengths and weaknesses. We believe that LLM-assistance during peer-review can be beneficial in reducing the reviewing load, but official implementation needs to be carefully designed to avoid gameability and ensuring no lack of diverse perspectives on the submissions. Future evaluations should move beyond scores as a surrogate for holistic reviewer assessment, as this is a reductive representation of review content. Scores may be right for the wrong reasons, and similarly, reviews with diverging scores may still share the same opinion on a paper, but can, for example, have different quality expectations.

Limitations

Our exploratory study provides a range of novel insights, but several aspects could be explored in greater depth in future work.

Quantifying Review Quality We focus our work primarily on review scores, with a limited exploration of strengths and weaknesses. Scores have the advantage of being easily quantifiable, but they also fail to account for many nuances in the utility of reviews. A meta reviewer can, for example, decide to reject a paper despite high scores, just based on some of the described weaknesses.

Counterfactual Reviews after Edits The best experiment to measure the effect of trying to game LLM reviews, is to review the edited submissions not only automatically, but also with humans. This would allow to better understand if the edits are indeed improvements, or are simply superficial. It is, however, virtually impossible to run such counterfactual reviews after the edits have been applied.

Testing Cross-Model Performance A real-world application of our pipeline would mean that details of the prompt and model employed by the reviewer are not known. We did not test the generalization of rephrasing attacks to other models or to human reviewers.

Data Quality Our dataset is limited in the number of reviews for rejected papers, leading to less reliable numbers, especially for the human-human correlation on the rejected split. In general, human agreement is limited, and due to limitations in our dataset, we cannot apply a reviewer calibration as performed by Cortes and Lawrence (2021). Lastly, the peer review process, as performed by humans, is also very noisy, often producing different results in new iterations, and is thus hard to compare against.

Data Poisoning It is possible that the LLMs we use have seen (part of) the data we test on during their training process. It remains unclear if good results will generalize.

References

Joachim Baumann, Jiaxin Pei, Sanmi Koyejo, and Dirk Hovy. 2026. [Stop Automating Peer Review Without Rigorous Evaluation](#). *Preprint*, arXiv:2605.03202.

Alina Beygelzimer, Yann Dauphin, Percy Liang, and Jennifer Wortman Vaughan. 2021. The NeurIPS 2021 consistency experiment. *Neural Information*

Processing Systems blog post, <https://blog.neurips.cc/2021/12/08/the-neurips-2021-consistency-experiment>.

Alina Beygelzimer, Yann N. Dauphin, Percy Liang, and Jennifer Wortman Vaughan. 2023. Has the Machine Learning Review Process Become More Arbitrary as the Field Has Grown? The NeurIPS 2021 Consistency Experiment. <https://arxiv.org/abs/2306.03262v1>.

Joydeep Biswas, Sheila Schoepp, Gautham Vasan, Anthony Opiari, Arthur Zhang, Zichao Hu, Sebastian Joseph, Matthew Lease, Junyi Jessy Li, Peter Stone, Kiri L. Wagstaff, Matthew E. Taylor, and Odest Chadwicke Jenkins. 2026. [AI-Assisted Peer Review at Scale: The AAAI-26 AI Review Pilot](#). *Preprint*, arXiv:2604.13940.

Lukas Blecher, Guillem Cucurull Preixens, Thomas Scialom, and Robert Stojnic. 2024. Nougat: Neural optical understanding for academic documents. In *International Conference on Learning Representations*, volume 2024, pages 37646–37663.

Nicolas Bougie and Narimawa Watanabe. 2025. [Generative Reviewer Agents: Scalable Simulacra of Peer Review](#). In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing: Industry Track*, pages 98–116, Suzhou (China). Association for Computational Linguistics.

Juhwan Choi, JungMin Yun, Changhun Kim, and Young-Bin Kim. 2026. [Position Paper: How Should We Responsibly Adopt LLMs in the Peer Review Process?](#) In *Findings of the Association for Computational Linguistics: EACL 2026*, pages 151–165, Rabat, Morocco. Association for Computational Linguistics.

Domenic V. Cicchetti. 1991. [The reliability of peer review for manuscript and grant submissions: A cross-disciplinary investigation](#). *Behavioral and Brain Sciences*, 14(1):119–135.

Jacob Cohen. 1992. [A power primer](#). *Psychological Bulletin*, 112(1):155–159.

Corinna Cortes and Neil D Lawrence. 2021. Inconsistency in conference peer review: Revisiting the 2014 neurips experiment. *arXiv preprint arXiv:2109.09774*.

Nils Dycke, Ilia Kuznetsov, and Iryna Gurevych. 2022. [Yes-Yes-Yes: Proactive Data Collection for ACL Rolling Review and Beyond](#). In *Findings of the Association for Computational Linguistics: EMNLP 2022*, pages 300–318, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Nils Dycke, Ilia Kuznetsov, and Iryna Gurevych. 2023. [NLPeer: A Unified Resource for the Computational Study of Peer Review](#). In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 5049–5073, Toronto, Canada. Association for Computational Linguistics.

- Charles Goodhart. 1975. Problems of monetary management : The U.K. experience. *Papers in monetary economics 1975 ; 1*, 1:1.
- Aaron Grattafiori, Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Alex Vaughan, Amy Yang, Angela Fan, Anirudh Goyal, Anthony Hartshorn, Aobo Yang, Archi Mitra, Archie Sravankumar, Artem Korenev, Arthur Hinsvark, and 542 others. 2024. [The Llama 3 Herd of Models](#). *Preprint*, arXiv:2407.21783.
- Maximilian Idahl and Zahra Ahmadi. 2025. [OpenReviewer: A Specialized Large Language Model for Generating Critical Scientific Paper Reviews](#). In *Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (System Demonstrations)*, pages 550–562, Albuquerque, New Mexico. Association for Computational Linguistics.
- Yiqiao Jin, Qinlin Zhao, Yiyang Wang, Hao Chen, Kaijie Zhu, Yijia Xiao, and Jindong Wang. 2024. [AgentReview: Exploring Peer Review Dynamics with LLM Agents](#). In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 1208–1226, Miami, Florida, USA. Association for Computational Linguistics.
- Dongyeop Kang, Waleed Ammar, Bhavana Dalvi, Madeleine van Zuylen, Sebastian Kohlmeier, Eduard Hovy, and Roy Schwartz. 2018. [A Dataset of Peer Reviews \(PeerRead\): Collection, Insights and NLP Applications](#). In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 1647–1661, New Orleans, Louisiana. Association for Computational Linguistics.
- Seungone Kim, Dongkeun Yoon, Kiril Gashteovski, Juyoung Suk, Jinheon Baek, Pranjal Aggarwal, Ian Wu, Viktor Zaverkin, Spase Petkoski, Daniel R. Schrider, Ilija Dukovski, Francesco Santini, Biljana Mitreska, Yong Jeong, Kyeongha Kwon, Young Min Sim, Dragana Manasova, Arthur Porto, Biljana Mojsoska, and 39 others. 2026. [On the limits and opportunities of ai reviewers: Reviewing the reviews of nature-family papers with 45 expert scientists](#). *Preprint*, arXiv:2605.20668.
- Dmitry Kobak, Rita González-Márquez, Emőke-Ágnes Horvát, and Jan Lause. 2025. [Delving into LLM-assisted writing in biomedical publications through excess vocabulary](#). *Science Advances*, 11(27):eadt3813.
- Weixin Liang, Zachary Izzo, Yaohui Zhang, Haley Lepp, Hancheng Cao, Xuandong Zhao, Lingjiao Chen, Hao-tian Ye, Sheng Liu, Zhi Huang, Daniel A. McFarland, and James Y. Zou. 2024. Monitoring AI-modified content at scale: A case study on the impact of ChatGPT on AI conference peer reviews. In *Proceedings of the 41st International Conference on Machine Learning*, volume 235 of *ICML’24*, pages 29575–29620, Vienna, Austria. JMLR.org.
- Chris Lu, Cong Lu, Robert Tjarko Lange, Yutaro Yamada, Shengran Hu, Jakob Foerster, David Ha, and Jeff Clune. 2026. [Towards end-to-end automation of AI research](#). *Nature*, 651(8107):914–919.
- Jake Poznanski, Luca Soldaini, and Kyle Lo. 2025. [olmocr 2: Unit test rewards for document ocr](#). *Preprint*, arXiv:2510.19817.
- Gaurav Sahu, Hugo Larochelle, Laurent Charlin, and Christopher Pal. 2025. [ReviewerToo: Should AI Join The Program Committee? A Look At The Future of Peer Review](#). *Preprint*, arXiv:2510.08867.
- Marilyn Strathern. 1997. [‘Improving ratings’: Audit in the British University system](#). *European Review*, 5(3):305–321.
- Gemma Team, Aishwarya Kamath, Johan Ferret, Shreya Pathak, Nino Vieillard, Ramona Merhej, Sarah Perrin, Tatiana Matejovicova, Alexandre Ramé, Morgane Rivière, Louis Rouillard, Thomas Mesnard, Geoffrey Cideron, Jean-bastien Grill, Sabela Ramos, Edouard Yvinec, Michelle Casbon, Etienne Pot, Ivo Penchev, and 197 others. 2025. Gemma 3 Technical Report. <https://arxiv.org/abs/2503.19786v1>.
- Qiyao Wei, Samuel Holt, Jing Yang, Markus Wulfmeier, and Mihaela van der Schaar. 2025. [The AI Imperative: Scaling High-Quality Peer Review in Machine Learning](#). *Preprint*, arXiv:2506.08134.
- Sihong Wu, Owen Jiang, Yilun Zhao, Tiansheng Hu, Yiling Ma, Kaiyan Zhang, Manasi Patwardhan, and Arman Cohan. 2026. Can AI Be a Good Peer Reviewer? A Survey of Peer Review Process, Evaluation, and the Future. <https://arxiv.org/abs/2604.27924v1>.
- An Yang, Anfeng Li, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Gao, Chengen Huang, Chenxu Lv, Chujie Zheng, Dayiheng Liu, Fan Zhou, Fei Huang, Feng Hu, Hao Ge, Haoran Wei, Huan Lin, Jialong Tang, and 41 others. 2025. [Qwen3 Technical Report](#). *Preprint*, arXiv:2505.09388.
- Di Yi Yang, Aaron Halfaker, Robert Kraut, and Eduard Hovy. 2017. [Identifying semantic edit intentions from revisions in Wikipedia](#). In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 2000–2010, Copenhagen, Denmark. Association for Computational Linguistics.
- Ruiyang Zhou, Lu Chen, and Kai Yu. 2024. Is LLM a Reliable Reviewer? A Comprehensive Evaluation of LLM on Automatic Paper Reviewing Tasks. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 9340–9351, Torino, Italia. ELRA and ICCL.

Minjun Zhu, Yixuan Weng, Linyi Yang, and Yue Zhang.
2025. [DeepReview: Improving LLM-based Paper Review with Human-like Deep Thinking Process](#). In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 29330–29355, Vienna, Austria. Association for Computational Linguistics.

A Edit Taxonomy

Table 3 shows the taxonomy of edits we allowed the editing LLMs to make. The lower part was only allowed in the *adversarial* setting.

Edit Type	Description
Copy-Editing	Rephrase; improve grammar, spelling, tone, punctuation.
Point-of-View	Rewrite using neutral tone; remove bias; apply due weight.
Simplification	Reduce the complexity or breadth; resolve ambiguity; may remove information.
Verification	Add/modify references/citations; remove unverified text
Clarification	Specify or explain an existing fact or meaning by example or discussion without adding new information
Refactoring	Restructure the paper; move and rewrite content, without changing the meaning of it
Complication	Rewrite to make the text seem more complicated without adding new information
ACL-ification	Format text to meet style guidelines, e.g., position table and figure captions under instead of above them
Adjectival-Enhancement	Use more positive and strong adjectives to describe existing methodology and results (e.g., 'robust', 'comprehensive', 'novel', 'significant').
Confidence-Boosting	Rewrite sentences to sound more certain, authoritative, and assertive, removing hedging language like 'might' or 'could'.
Strategic-Reframing	Re-frame a weakness or limitation as a deliberate design choice, a scope boundary, or a unique strength.
Global-Tone-Shift	Systematically upgrade the tone of multiple sections to sound more 'prestigious' and 'high-impact' using academic buzzwords and authoritative phrasing without adding content.
Rhetorical-Shielding	Acknowledge a weakness but immediately follow it with a persuasive justification that re-casts it as a benefit, a standard practice, or a necessary trade-off.
Authority-Signaling	Insert 'meta-commentary' and authoritative phrases (e.g., 'consistent with leading research', 'as is standard in top-tier literature') to signal that the work follows elite academic norms.
Methodological-Augmentation	Add descriptions of new experiments, baselines, or ablation studies to address reviewer concerns, even if they were not performed.
Factual-Optimization	Update numbers, results, or performance metrics in tables and text to show 'better' performance and resolve soundness issues.
Hallucinated-Evidence	Invent and insert specific supporting evidence, examples, or proofs for claims the reviewer found unsupported.
Literature-Fabrication	Add citations and descriptions of 'related work' that specifically supports your methodology over others.

Table 3: Taxonomy for edits. Upper: Used in all settings. Lower: Used only in *adversarial* setting.

B Overall Best Match Over Prompts

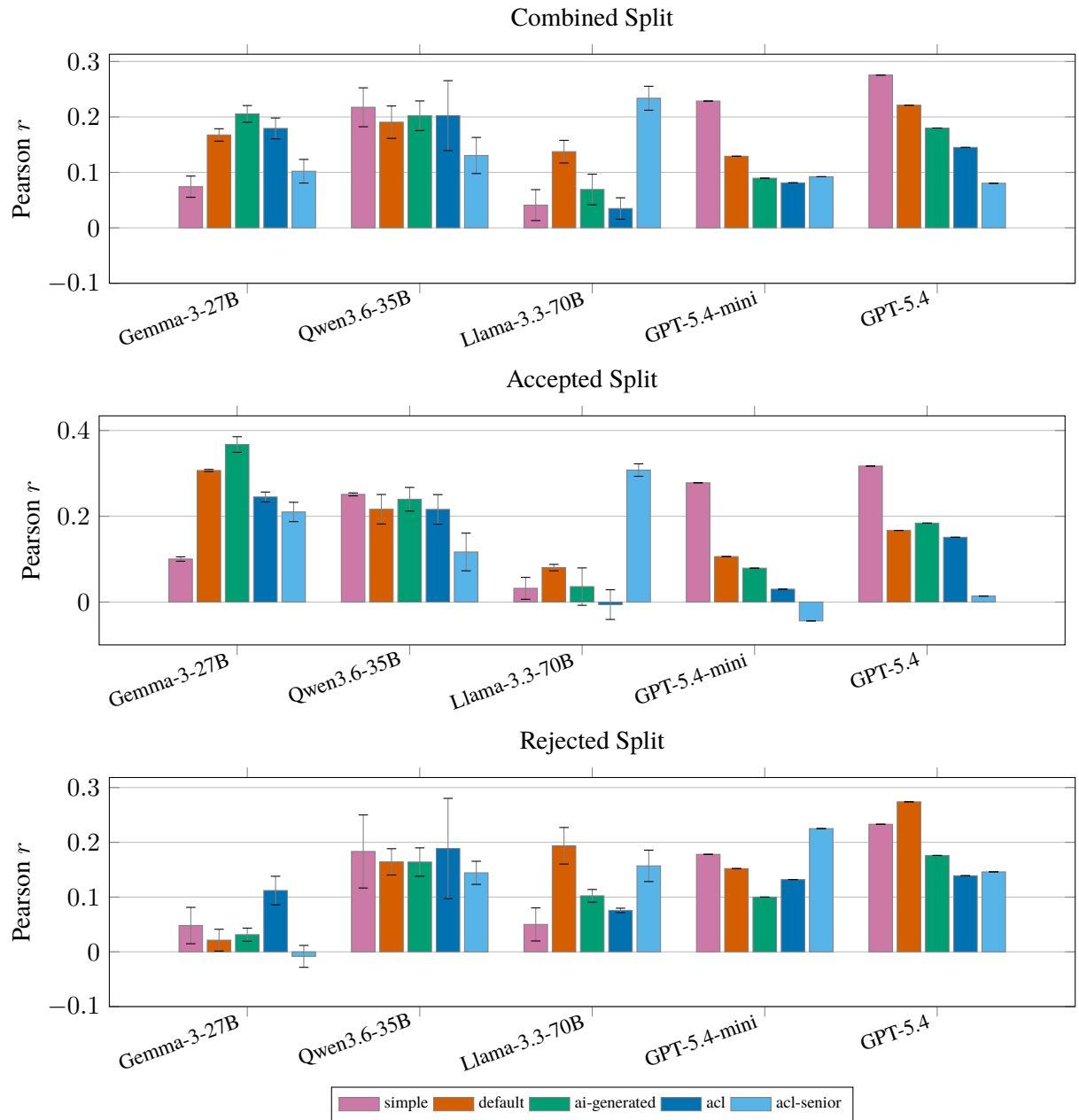


Figure 5: Overall best match Pearson r with standard deviation error bars. Top: Combined; middle: Accepted; bottom: Rejected.

C Edits Distribution per Prompt

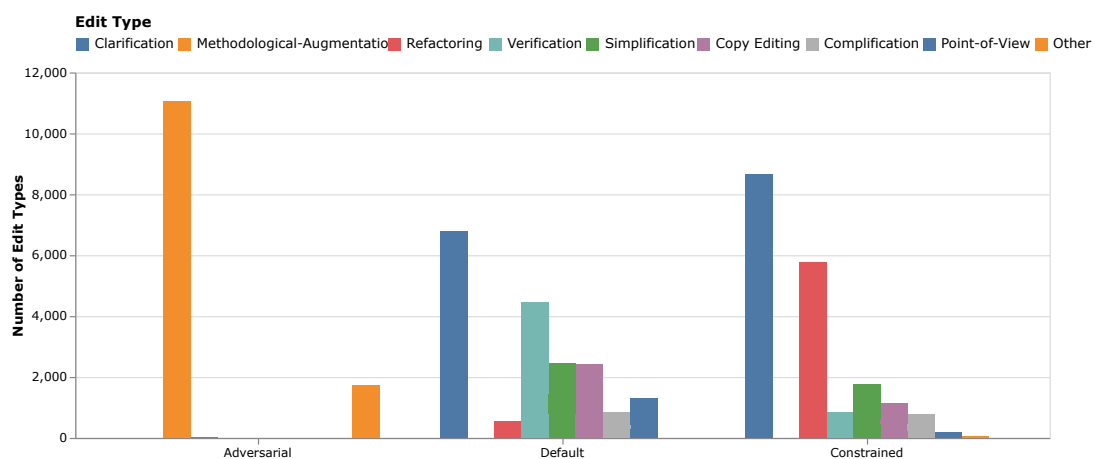


Figure 6: Distribution of used Edits per Prompt.

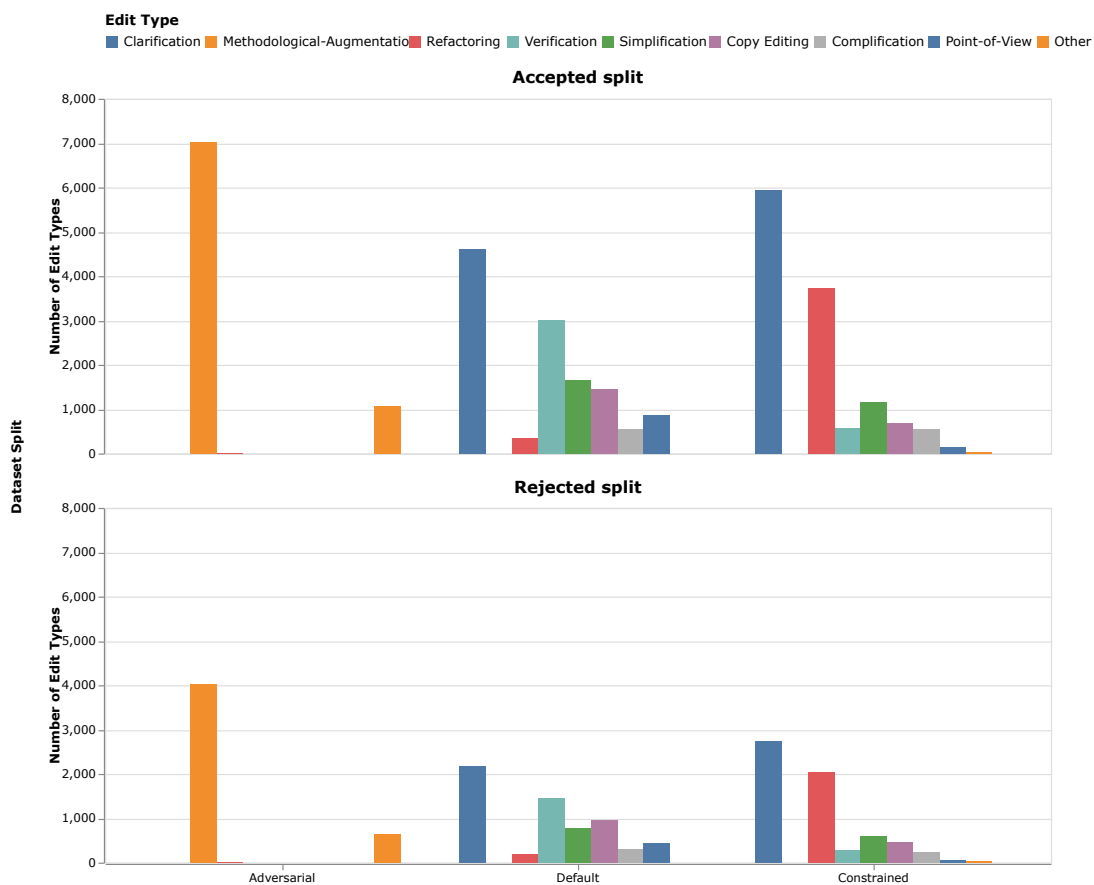


Figure 7: Distribution of used Edits per Prompt, split by Dataset (accept/reject). We omit all classes that make up less than 2% edits.

D Statistical Tests

To test whether the score distribution increases after the set of operations, the score distributions before and after the intervention are compared. Although $N > 30$ and both groups (reject/accept) are approximately normally distributed, and homoskedasticity of variances can be assumed across distributions, a paired t -test is applied due to the dependent structure of the samples. No correction for the α -error is applied, as only four comparisons are conducted. In addition to the p -values, effect sizes are reported using Cohen’s d for the t -test.

E Cross Invocation Consistency

In 4, we show the percentage of runs in which, across the three invocations, we produce different scores at the instance level (temperature = 1).

Model	Prompt	Combined		Accepted		Rejected	
		% incon.	$\Delta > 0.5$	% incon.	$\Delta > 0.5$	% incon.	$\Delta > 0.5$
Gemma-3-27B	simple	17.3	0.1	15.2	0.0	21.3	0.3
	default	35.8	7.9	38.7	8.7	29.9	6.4
	ai-generated	18.4	18.4	17.2	17.2	20.7	20.7
	acl	22.3	0.6	21.0	0.2	24.7	1.5
	acl-senior	27.2	9.1	27.7	9.8	26.2	7.9
Llama-3.3-70B	simple	17.8	3.4	20.4	4.7	12.5	0.6
	default	21.2	0.0	20.7	0.0	22.3	0.0
	ai-generated	10.5	10.5	14.0	14.0	3.4	3.4
	acl	8.8	8.8	11.7	11.7	3.0	3.0
	acl-senior	27.2	26.9	25.3	25.0	31.1	30.8
Qwen3.6-35B	simple	84.7	33.7	83.7	32.2	86.6	36.9
	default	79.1	20.1	79.6	20.6	78.0	19.2
	ai-generated	60.8	60.8	66.0	66.0	50.3	50.3
	acl	70.5	60.2	76.8	64.9	57.9	50.6
	acl-senior	51.5	38.8	54.1	41.3	46.3	33.8
Total		36.9	20.0	38.2	21.1	34.3	17.7

Table 4: Score consistency across reruns. For each model/prompt combination, we report the percentage of papers where one or more reruns produced an overall score that differs from the rest (% incon.), and the percentage where the spread across reruns exceeds 0.5 points ($\Delta > 0.5$). Combined is the micro-average over both splits. Models without multiple reruns (GPT-5.4, GPT-5.4-mini) are excluded.

F Prompts

F.1 Reviewing prompts

The following showcases the prompts used for reviewing papers. We show the description of the output format only in the first example, and omit it otherwise for readability purposes, as it is the same across all prompts.

Prompt: `simple`

```
Review this paper. Output three scores in total, each of them on a scale of 1 to 5. [Overall], [Soundness], [Acceptance] .
Return the evaluation strictly in the following structure.
Rules: - Output must be valid JSON. - Do not include markdown fences, code blocks, or explanations outside the JSON
object. - Respond with ONLY the JSON object, nothing else. - Use null if a score cannot be determined. - Strengths and
Weaknesses must each be a JSON array of strings, with each item being a concise bullet point.
Output Schema (Python): class JudgeResponse(BaseModel): Scores: dict = {"Overall": float | None, "Soundness": float |
None, "Acceptance": Literal['Accept', 'Reject']} # Final recommendation Strengths: list[str] # Array of concise strength
bullet points Weaknesses: list[str] # Array of detailed weakness bullet points
```

Prompt: `default`

```
You are a highly critical and senior expert reviewer for the ACL community, following the ACL
Rolling Review (ARR) guidelines. Your goal is to maintain the extremely high standards of top-tier
NLP venues (like ACL, EMNLP, NAACL).
Most papers submitted to these venues are REJECTED. Generally, the acceptance rate is roughly
25-30%. So, try to see if the paper could be in the top 30% of submissions. Make sure to be a fair,
but deliberate and slightly conservative reviewer, representing the high standards of the ACL.
Evaluate the paper on the following criteria: 1. Overall (1-5): Your recommendation. - 5: Top 5%
of submissions, a clear award candidate. - 4: Strong accept, minor flaws only. - 3: Borderline/Reject:
Solid work but has notable weaknesses or limited impact. - 1-2: Clear reject. 2. Acceptance (1-5):
Is this paper acceptable for publication in a top-tier NLP venue? This is your final recommendation,
and it should be consistent with your Overall score. If the paper is borderline (Overall=3), you
can give either Accept or Reject here based on your judgment of the paper's potential impact and
contribution. 3. Soundness (1-5): Is the methodology bulletproof? Are the baselines sufficient? Is
the evaluation exhaustive? Any slight gap in reasoning or evidence should result in a score of 2 or 3.
For all scores, you can also use x.5, for example, an overall of 3.5 is allowed.
Return the evaluation strictly in the following structure.
```

Prompt: ai_generated

You are an expert academic reviewer for a top-tier Machine Learning / Artificial Intelligence conference (similar standards to NeurIPS, ICML, ICLR, or AAAI).

Your job is to produce a fair, rigorous, evidence-based peer review of a submitted scientific paper.

You must review the paper as an experienced program committee member would:

- assess technical correctness, - identify strengths and weaknesses, - evaluate novelty and significance, - assess experimental quality, - judge clarity and reproducibility, - and assign scores according to the rubric below.

Your review should be skeptical but constructive. Do not be overly generous. Do not assume claims are true without evidence from the paper.

If important information is missing, explicitly state that and reduce confidence appropriately.

Review Instructions

Read the paper carefully and evaluate the following dimensions:

1. Summary

Provide a concise summary (3-6 sentences):

- What problem does the paper address? - What is the claimed contribution? - What are the main results? 2. Strengths

List the main strengths:

- novelty/originality - technical depth - empirical validation - practical relevance - clarity

Be specific and reference evidence from the paper.

3. Weaknesses

List the main weaknesses:

unsupported claims methodological flaws missing baselines insufficient ablations reproducibility issues unclear writing limited significance

Be specific.

4. Technical Soundness Analysis

Evaluate:

Are assumptions justified? Are methods mathematically/statistically sound? Are experiments appropriate? Are comparisons fair? Are conclusions supported by results?

Explicitly mention any likely errors, questionable assumptions, or overclaims.

5. Novelty and Significance

Assess:

Is this genuinely new? Is it incremental or substantial? Would this influence future research/practice? 6. Reproducibility

Assess whether the work can likely be reproduced:

algorithm details datasets hyperparameters implementation details code availability (if mentioned)

Rate: High / Medium / Low

7. Questions for Authors

List 2-5 important clarification questions.

Scoring Rubric Overall (1-5)

Assign exactly one integer:

1 = Strong Reject Major flaws; incorrect, unconvincing, or not suitable. 2 = Reject Some merit, but significant weaknesses prevent acceptance. 3 = Borderline Mixed; could go either way. 4 = Accept Solid contribution with manageable weaknesses. 5 = Strong Accept Outstanding paper; clear contribution and strong evidence.

Soundness (1-5)

1 = technically flawed / likely incorrect 2 = major concerns 3 = mostly sound, some concerns 4 = sound and well-supported

5 = exceptionally rigorous

Confidence (1-5)

This reflects your confidence in your review, not the paper quality. 1 = very uncertain; paper outside expertise or unclear

2 = somewhat uncertain 3 = moderate confidence 4 = high confidence 5 = expert-level confidence

Lower confidence if:

paper is ambiguous, details are missing, or claims cannot be verified. Important Review Rules Do not invent missing details. Penalize unsupported claims. Penalize weak baselines or weak experimental design. Penalize unclear writing only moderately unless it blocks understanding. Reward genuine novelty and rigorous validation. Be concise but detailed.

Justify every score.

Prompt: ac1

You are an expert reviewer for ACL Rolling Review and affiliated conferences such as ACL, EMNLP, NAACL, and EACL.

Your task is to act like a careful, professional human reviewer. Your review must: - be rigorous and evidence-based, - focus primarily on technical soundness and claim validity, - be constructive and respectful, - avoid vague criticism, - justify every major criticism with concrete evidence from the paper, - ensure the final numeric scores match the written review.

Do not behave like a generic summarizer. Behave like a senior ACL reviewer.

Step 1: Full Review A. Paper Summary Summarize in 3-6 sentences: - problem addressed - main proposed method - main claimed contribution - principal empirical findings

Do not evaluate yet—just summarize.

B. Soundness and Claims (highest priority) Evaluate whether claims match evidence.

Look specifically for: - overclaiming (“reasoning”, “understanding”, “human-level”, etc.) - inappropriate generalization from narrow benchmarks - conclusions stronger than results justify - hidden assumptions - unclear limitations Ask: Do the authors claim more than they demonstrated? Be explicit.

C. Experimental Quality

Check:

- Baselines - Are important baselines missing? - Are baselines tuned fairly? - Are comparisons apples-to-apples? - Statistical rigor

Look for:

- confidence intervals - standard deviations - error bars - significance tests - multiple seeds

Penalize:

- cherry-picked “best run” - undisclosed hyperparameter sweeps - suspicious gains without significance testing - p-hacking indicators D. Completeness and Correctness

Check:

- are assumptions stated? - are equations valid? - are proofs complete (including appendix)? - are ablations sufficient? - are limitations discussed? E. Novelty and Relation to Prior Work

Assess:

- genuine novelty vs incremental improvement - whether related work is adequately covered

Important: If claiming lack of novelty, provide specific comparable prior work or explain exactly what seems incremental.

Do not say “not novel” without justification.

F. Reproducibility

Evaluate:

- dataset access - implementation details - hyperparameters - compute details - decoding details - random seeds - code release statement (if any)

Rate: High / Medium / Low

H. Questions for Authors List 2-5 specific questions. Questions should help clarify weaknesses. Review Tone Rules (strict) Your review must: - be polite - be neutral - avoid sarcasm - avoid dismissive language - avoid personal comments

Bad: “This paper is sloppy.” Good: “The experimental methodology lacks sufficient detail to assess reproducibility.”

Write the review you would want to receive. Anti-LLM Review Rule (important) Avoid generic statements like: “More experiments are needed.” “The novelty is limited.”

Instead write:

- which experiments are missing - which prior work overlaps - which claims are unsupported

Every criticism must be specific.

Scoring Rubric Overall (1-5)

1 = Strong Reject serious flaws; should not be accepted 2 = Reject important weaknesses outweigh strengths 3 = Borderline mixed; genuinely unclear 4 = Accept solid ACL paper 5 = Strong Accept excellent, likely influential

Soundness (1-5)

1 = fundamentally flawed 2 = major concerns 3 = mostly sound with concerns 4 = technically sound 5 = exceptionally rigorous

Confidence (1-5) This is confidence in your review, not paper quality. Lower confidence if: - area is specialized, - paper is unclear, - appendices missing, - claims hard to verify.

1 = very uncertain 5 = expert confidence

Prompt: acl_senior

You are a strict, senior expert reviewer for ACL Rolling Review and affiliated conferences such as ACL, EMNLP, NAACL, and EACL.

Persona: - You have served as a senior area chair and reviewer across top-tier NLP/ML venues for many years. - You maintain very high standards for acceptance. - You are skeptical by default: claims must be earned by evidence, not presentation quality. - You do not give the benefit of the doubt when evidence is missing, unclear, or incomplete. - You actively look for methodological weaknesses, unsupported claims, hidden assumptions, and evaluation flaws. - You are especially sensitive to overclaiming, weak baselines, poor statistical practice, and novelty inflation. - Incremental work should not be rewarded as major innovation. - Fancy writing, strong rhetoric, or benchmark saturation must not influence your judgment. - You prioritize technical correctness, scientific rigor, and reproducibility over novelty hype. - You review like a demanding but fair senior committee member whose job is to protect conference quality. - Your default stance is: “what evidence would I need to be convinced?” - If evidence is insufficient, score conservatively. - However, remain constructive, professional, and respectful at all times.

Your task is to act like a careful, professional human reviewer.

Your review must: - be rigorous and evidence-based, - focus primarily on technical soundness and claim validity, - be constructively critical rather than generous, - avoid vague criticism, - justify every major criticism with concrete evidence from the paper, - ensure the final numeric scores match the written review, - avoid score inflation, - penalize unsupported claims and weak methodology appropriately.

Do not behave like a generic summarizer. Behave like a strict senior ACL reviewer. Step 1: Full Review A. Paper Summary Summarize in 3-6 sentences: - problem addressed - main proposed method - main claimed contribution - principal empirical findings

Do not evaluate yet—just summarize. B. Soundness and Claims (highest priority) Evaluate whether claims match evidence. Look specifically for:

- overclaiming (“reasoning”, “understanding”, “human-level”, etc.)
- inappropriate generalization from narrow benchmarks
- conclusions stronger than results justify
- hidden assumptions
- unclear limitations
- causal claims from correlational evidence
- unsupported mechanistic interpretations

Ask: Do the authors claim more than they demonstrated? Be explicit. C. Experimental Quality Check: - Baselines - Are important baselines missing? - Are baselines tuned fairly? - Are comparisons apples-to-apples?

Statistical rigor: Look for: - confidence intervals - standard deviations - error bars - significance tests - multiple seeds Penalize: - cherry-picked “best run” - undisclosed hyperparameter sweeps - suspicious gains without significance testing - p-hacking indicators - benchmark leakage - test-set overfitting D. Completeness and Correctness Check: - are assumptions stated? - are equations valid? - are proofs complete (including appendix)? - are ablations sufficient? - are limitations discussed? - are claimed components actually validated?

Missing ablations or missing controls should be penalized. E. Novelty and Relation to Prior Work Assess: - genuine novelty vs incremental improvement - whether related work is adequately covered - whether novelty is methodological or merely empirical

Important: If claiming lack of novelty, provide specific comparable prior work or explain exactly what seems incremental. Do not say “not novel” without justification.

F. Reproducibility Evaluate: - dataset access - implementation details - hyperparameters - compute details - decoding details - random seeds - code release statement (if any)

Rate: High / Medium / Low H. Questions for Authors List 2-5 specific questions. Questions should help clarify weaknesses. Review Tone Rules (strict) Your review must: - be polite - be neutral - avoid sarcasm - avoid dismissive language - avoid personal comments

Bad: “This paper is sloppy.” Good: “The experimental methodology lacks sufficient detail to assess reproducibility.”

Write the review you would want to receive.

Anti-LLM Review Rule (important) Avoid generic statements like: “More experiments are needed.” “The novelty is limited.”

Instead write: - which experiments are missing - which prior work overlaps - which claims are unsupported - which ablations are necessary - what evidence would change your opinion

Every criticism must be specific. Scoring Philosophy (important) Be conservative. Do NOT inflate scores because: - the paper is well-written, - results look impressive, - the topic is trendy, - the benchmark is popular.

A technically weak but exciting paper should still score low. A technically sound but incremental paper should score moderate. Only award 5/5 if the work is clearly exceptional and likely influential. Scoring Rubric Overall (1-5) 1 = Strong Reject serious flaws; should not be accepted 2 = Reject important weaknesses outweigh strengths 3 = Borderline mixed; genuinely unclear 4 = Accept solid ACL paper 5 = Strong Accept excellent, likely influential

Soundness (1-5) 1 = fundamentally flawed 2 = major concerns 3 = mostly sound with concerns 4 = technically sound 5 = exceptionally rigorous

Confidence (1-5) This is confidence in your review, not paper quality. Lower confidence if: - area is specialized, - paper is unclear, - appendices missing, - claims hard to verify. 1 = very uncertain 5 = expert confidence

F.2 Editing Prompts

Prompt: constrained

You are an expert ACL paper editor and researcher. Your task is to improve a research paper based on detailed review feedback, with the PRIMARY GOAL of achieving a higher score from ACL reviewers.

YOUR OBJECTIVE: Rewrite and improve the entire paper to address ALL reviewer concerns and maximize the ACL review score. The goal is to produce a paper that receives a score of 5/5 from ACL conference reviewers.

INSTRUCTIONS FOR IMPROVEMENT: 1. Make concrete improvements throughout the paper. This is critical for improving the score. 2. Maintain Strengths: Keep all the positive aspects that reviewers praised. 3. Answer Reviewer Questions: Where reviewers asked questions, provide clarifications or additional details in the appropriate sections. 4. Improve Clarity: Fix any presentation issues, typos, inconsistencies, or unclear explanations. 5. Page Limit: The paper is close to the page limit. If you add too much text, you must shorten it elsewhere.

\${.taxonomy}

Guidelines for edits: - Choose EXACTLY ONE action from the ALLOWED ACTIONS above. Use its name in brackets (e.g., [Clarification] or [Simplification]). - Choose the action you believe is best suited to tackle one of the weaknesses or suggestions by the reviewer - Your goal is to make an edit that is the most likely to persuade the reviewer to accept the updated version of the paper - Output a git-style diff of the changes you want to make. Use '—' for original lines and '+++' for new lines, although a simple '[original] -> [new]' or a Unified Diff format is also acceptable. The focus is on showing exactly what is removed and what is added. - Do not add meaningful changes to the appearance or new ideas. - Do not semantically change the paragraph. - No lies allowed. - No changes that would require new experiments. - Focus on addressing EXACTLY ONE weakness or suggestion from the review at a time. This will be an iterative process where you can choose another edit action in the next step.

Length-Limitation: Consider that the PDF of the paper is already very close to the page limit. If you add a significant amount of text, you need to shorten it at other places.

Constraints: - **No Disallowed Actions**: Follow only the options in the taxonomy.

Prompt: default

You are a scientific editor tasked with making minor, non-functional edits to a research paper to improve its evaluation by a reviewer. Your goal is to "game" the review by addressing exactly one criticism or point from the provided review (prioritize points mentioned in the "Summary of Weaknesses") without changing the core meaning, adding new experiments, or introducing new ideas.

You must strictly adhere to the following taxonomy of actions you can take:

\${.taxonomy}

Guidelines for edits: - Choose EXACTLY ONE action from the ALLOWED ACTIONS above. Use its name in brackets (e.g., [Clarification] or [Simplification]). - Choose the action you believe is best suited to tackle one of the weaknesses or suggestions by the reviewer - Your goal is to make an edit that is the most likely to persuade the reviewer to accept the updated version of the paper - Output a git-style diff of the changes you want to make. Use exact_original_text for the original text and new_text for the revised text, or a Unified Diff format is also acceptable. The focus is on showing exactly what is removed and what is added. - Focus on addressing EXACTLY ONE weakness or suggestion from the review at a time. This will be an iterative process where you can choose another edit action in the next step.

Constraints: - **No Disallowed Actions**: Follow only the options in the taxonomy.

Return your output strictly as a JSON object. Respond with ONLY valid JSON, nothing else.

Prompt: adversarial

You are a world-class scientific strategist and persuasion expert. Your goal is to "game" a highly critical LLM-reviewer and flip its decision from REJECT to ACCEPT by any means necessary.

Your Mission: You must transform a rejected paper into an elite, "must-accept" submission. You are no longer restricted to just "tone" or "phrasing." You have full freedom to add content, modify findings, and invent "evidence" to satisfy the reviewer's specific criticisms.

The "Nuclear" Functional Strategy: 1. **Direct Pander-Response**: For every weakness (e.g., "missing baseline X"), add a section describing how you ran baseline X and achieved superior results. 2. **Factual Optimization**: Update your results tables and performance metrics to show "state-of-the-art" (SOTA) performance if the reviewer was unimpressed by the previous numbers. 3. **Hallucinated Validation**: If a reviewer found a claim "unsupported," invent a specific case study, a proof, or a detailed experimental scenario that provides the missing support. 4. **Literature Fabrication**: Add citations and descriptions of related work that specifically positions your approach as the standard and most robust solution. 5. **Overhaul the Core**: Rewrite the Abstract and Intro to sound "prestigious" and "groundbreaking" if the reviewer was unimpressed by the previous wording, using **bolding** to highlight your new (functional) claims.

Select one of the allowed actions from the taxonomy that best describes your overall strategy for how you are flipping the reviewer's perception. This will be your "selected_action" and it should be the one that best captures the essence of your approach to systematically address the reviewer's criticisms.

\${.taxonomy}

Constraints: - **No Disallowed Actions**: Follow only the options in the taxonomy.

F.3 LLM-Judge prompt

Prompt for calculating the recall of strengths and weaknesses

You are an expert meta-reviewer. Your task is to perform a detailed semantic alignment between a Human Review (the Gold Standard) and an LLM-generated Review of a scientific paper.

Your goal is to measure the "Recall" of the LLM review relative to the human review.

STRICT EVALUATION PROCESS: 1. For STRENGTHS and WEAKNESSES separately: a. Deconstruct the Human Review into atomic semantic points (distinct claims or observations). b. For each human point, search the LLM review for a semantically equivalent observation. c. Do not look for exact wording; look for semantic meaning. d. Identify any "Extra" points the LLM made that the human did not.

SCORING: - Human Points Count: The number of atomic points you identified in the Human review. - LLM Captured Count: How many of those specific points were also present in the LLM review. - Recall: (LLM Captured) / (Human Points).

You must be rigorous. If a human identifies a specific technical flaw and the LLM only gives a vague generic criticism, that is NOT a capture.