REVIEWARENA: A Large-Scale Cross-Conference Dataset and Benchmark for LLM Peer Review

Samarth P [1] **Sai Sathvik Motamarri** [1] **Vyoman Jain** [1] **Shiva Golugula** [1]

Abstract

Peer review is central to quality control in machine learning (ML), but growing submission volumes have strained reviewer capacity and motivated interest in large language models (LLMs) as reviewers. Progress is hindered by the lack of datasets pairing full papers with structured, multi-dimensional reviews across venues which capture the full review–rebuttal–decision process. We introduce **REVIEWARENA**, a large-scale peer-review dataset constructed from all OpenReview venues with public reviews at the time of writing: NeurIPS, ICLR, ICML, CoRL, COLM, EMNLP, and TMLR. The dataset comprises **51,529** papers and **196,099** reviews across fourteen review fields, including full PDFs, reviewer scores and text, rebuttals, meta-reviews, and final decisions, with post-rebuttal revisions for NeurIPS 2025. To facilitate research, we derive **REVIEWARENA-EVAL**, a **1,002**-paper benchmark spanning the six conferences with aligned, venue-specific evaluation protocols. Baseline experiments with six open-weight LLMs using venue-aware prompts show that current models are miscalibrated, compress rating scales, and weakly distinguish accepted from rejected papers, while review text quality is only weakly coupled to numeric accuracy. **REVIEWARENA** is a unified resource for studying automated peer review, enabling research on review generation, scoring, calibration, and decision-making.

1. Introduction

Peer review is the main quality-control mechanism in machine learning, but submission volume has grown faster than reviewer capacity. This pressure has renewed interest in LLMs that draft, critique, or score reviews (Liang et al., 2024b; Liu & Shah, 2023; D'Arcy et al., 2024a; Lu et al., 2024). A credible evaluation of such systems needs more than paper–review text pairs: it needs complete papers, native review forms, numeric sub-scores, rebuttals, meta-reviews, final decisions, and enough venues to distinguish model behaviour from venue convention.

The recent literature has made real progress, but the available data still makes this evaluation narrower than the reviewing process itself. PeerRead (Kang et al., 2018), ASAP-Review (Yuan et al., 2022), NLPeer (Dycke et al., 2023), MOPRD (Lin et al., 2023), Reviewer2 (Gao et al., 2024), ReviewMT (Tan et al., 2024), and Re^2 (Zhang et al., 2025) expanded open peer-review resources and enabled tasks such as acceptance prediction, review generation, meta-review generation, and rebuttal modelling. However, many benchmarks still flatten the target into either a single score or an unstructured block of text. This is a poor match to modern OpenReview forms, where reviewers provide multiple numerical axes, free-text fields, confidence, follow-up discussion, and sometimes post-rebuttal revisions. It also hides an important cross-venue issue: a score of 3 means something different on ICML 2025's 1–5 scale, CoRL's 1–4 scale, and NeurIPS's 1–10 scale.

We introduce REVIEWARENA, shown schematically in Figure 1, a dataset and benchmark built from public OpenReview records. The dataset covers NeurIPS 2021–2025, ICLR 2020–2026, ICML 2025, CoRL 2021–2024, COLM 2024–2025, EMNLP 2023, and TMLR: 51,529 papers and 196,099 reviews across fourteen review-form variants. Each record includes the PDF, paper metadata, all official reviews, structured scores, rebuttals, official discussion, area-chair meta-review, and decision. For NeurIPS 2025, where OpenReview exposes explicit post-rebuttal review-revision invitations, we also record whether each review was revised and the reviewer's final justification. We package the corpus as a Hugging Face `PdfFolder` dataset with one split per venue, so users can stream PDFs and metadata with standard dataset tooling.

On top of the dataset we define REVIEWARENA-EVAL, a 1,002-paper evaluation subset with 167 papers from each nu-

Accepted at ICML 2026 AI for Science Workshop.
[1]Independent Researcher, Bangalore, India. Correspondence to: Samarth P <samarthprakash8@gmail.com>.

Proceedings of the 43rd International Conference on Machine Learning, Seoul, South Korea. PMLR 306, 2026. Copyright 2026 by the author(s).

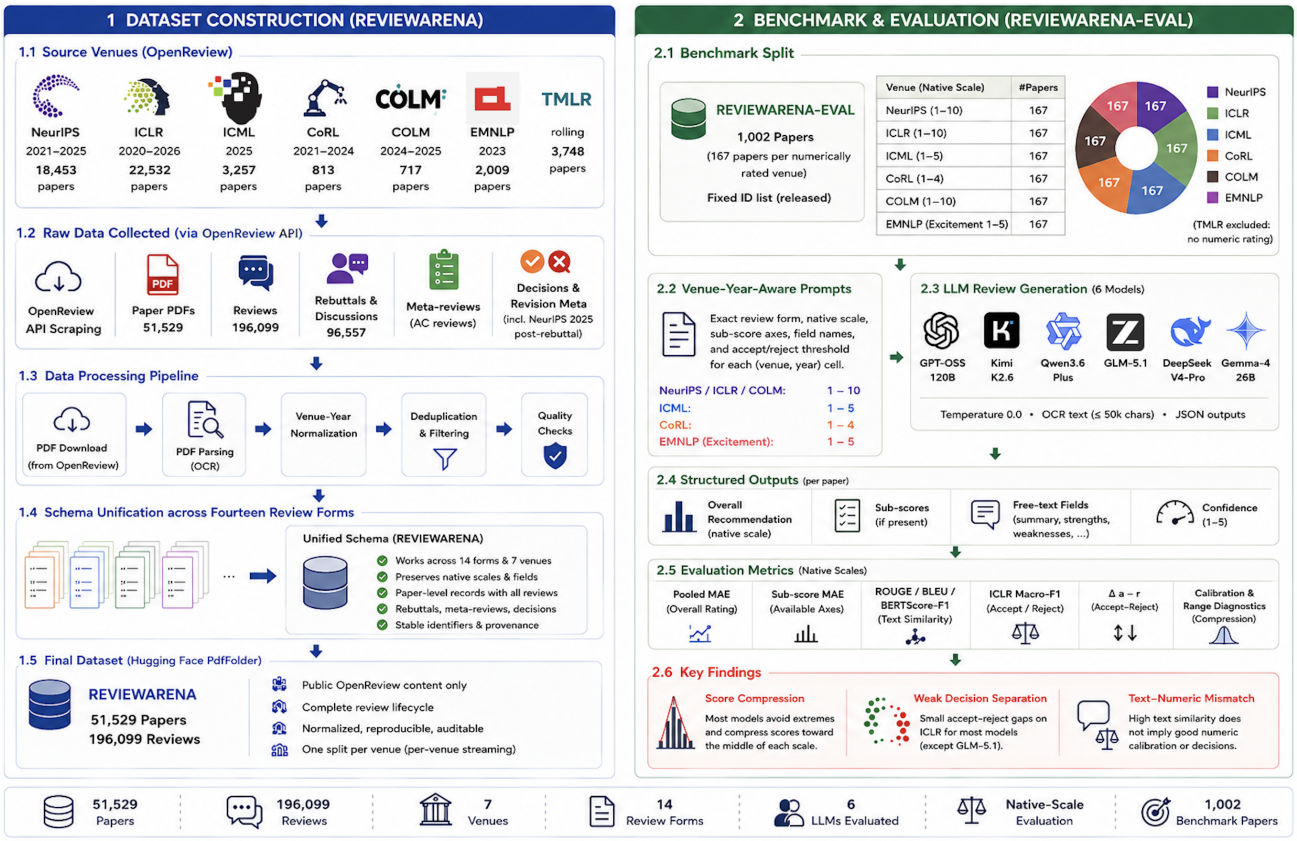


Figure 1. **Overview of REVIEWARENA and REVIEWARENA-EVAL.** Left: construction of REVIEWARENA from public OpenReview venues, including PDFs, structured reviews, rebuttals/discussions, meta-reviews, decisions, and schema unification across fourteen review forms. Right: construction of REVIEWARENA-EVAL, venue-year-aware prompting, six-model evaluation, native-scale metrics, and the main empirical findings.

merically rated venue. Six open-weight LLMs are prompted with the exact venue-year review form and scored against human ratings, sub-scores, free-text fields, and ICLR accept/reject decisions. This venue-year conditioning is essential rather than cosmetic. ICML 2025 uses 1–5, CoRL uses 1–4, EMNLP uses the 1–5 `Excitement` field, and only NeurIPS/ICLR/COLM use the familiar 1–10 scale. A generic prompt that asks every model to "rate from 1 to 10" is therefore out of range on half the benchmark.

The paper makes three contributions. First, REVIEWARENA is, to our knowledge, the largest public peer-review corpus by both paper and review count, and the first to preserve this breadth of native review-form structure across seven OpenReview venues. Second, REVIEWARENA-EVAL provides a fixed cross-venue LLM-reviewer benchmark with native-form prompting and native-scale scoring. Third, our evaluation, which shows that text mimicry, numeric calibration, and decision discrimination are distinct capabilities: Qwen3.6-Plus has the best overall-rating MAE (0.700), while Gemma-4-26B and GLM-5.1 lead text similarity without being the best numeric reviewers.

The dataset, code, evaluation harness, and predictions are released at `https://huggingface.co/datasets/anonymousNeurIPS2026submission4281/reviewarena`.

2. Related Work

Open peer-review datasets have evolved from static paper–review pairs toward richer process records. PeerRead (Kang et al., 2018) made the first widely used public resource for acceptance and score prediction, drawing from ICLR, ACL, and early NeurIPS workshop data. ASAP-Review (Yuan et al., 2022) added aspect-oriented labels for review generation, while NLPeer (Dycke et al., 2023) unified several NLP venues and emphasized ethical data collection. MOPRD (Lin et al., 2023) broadened the disciplinary scope beyond machine learning. Complementary analyses study substantiation structure inside review text (Guo et al., 2023). Other resources target specific parts of the reviewing pipeline: DISAPERE annotates discourse in review discussions (Kennard et al., 2022), PEERAssist studies paper–review interactions for decision predic-

tion (Bharti et al., 2021), MReD focuses on meta-review generation (Shen et al., 2022), PRRCA models rebuttal counter-arguments (Wu et al., 2022), Peer Review Analyze benchmarks reviewer discourse (Ghosal et al., 2022), Reviewer2 scales review-generation data through prompt generation (Gao et al., 2024), and ReviewMT frames peer review as long-context dialogue (Tan et al., 2024). The closest recent comparison is Re^2 (Zhang et al., 2025), which emphasizes consistency between initial submissions and reviews and includes rebuttal discussions from many OpenReview venues.

REVIEWARENA differs in emphasis. Rather than constructing a rebuttal-dialogue training set or a single generation target, it preserves the public review record as a venue-native benchmark substrate: PDFs, official reviews, numeric axes, free-text fields, rebuttals, meta-reviews, decisions, and post-rebuttal revisions where exposed. Parallel resources align reviewer feedback with manuscript revisions (D'Arcy et al., 2024b). This design is closer to the way accepted Datasets and Benchmarks papers present infrastructure resources: the value is not only a large count, but also a standardized, maintainable representation of messy real-world data that supports many downstream tasks. Table 2 gives the direct comparison.

Tutorial-style syntheses catalog what NLP can contribute to peer review (Kuznetsov et al., 2024). LLM-reviewer studies range from exploratory GPT-style reviewers (Liu & Shah, 2023) to large-scale analyses of generated feedback (Liang et al., 2024b), multi-agent review generation (D'Arcy et al., 2024a), automated research pipelines (Lu et al., 2024), and specialized reviewing systems and multimodal benchmarks (Zhou et al., 2024; Weng et al., 2025; Zhu et al., 2025; Idahl & Ahmadi, 2025; Yu et al., 2024; Gao et al., 2025). These works show that LLMs can produce plausible critiques, but also that overlap with human feedback is incomplete and calibration is fragile. REVIEWARENA-EVAL evaluates the same question in the native form of each venue rather than collapsing all papers into one generic prompt. This matters because the benchmark target is not merely "write a review", but "complete the review form that this venue actually used".

Finally, REVIEWARENA connects to work that studies peer review as a social and statistical process. The NeurIPS consistency experiment (Lawrence & Cortes, 2014), novice-reviewer studies (Stelmakh et al., 2021), and broader surveys of peer-review challenges (Shah, 2022) all motivate measuring disagreement, calibration, reviewer confidence, and process dynamics directly. Recent observational analyses quantify rapid adoption of LLM editing in AI-conference reviews (Liang et al., 2024a; Zhou et al., 2025), complementing detection benchmarks for synthetic reviewer text (Yu et al., 2025). By releasing the review–rebuttal–decision loop across seven venues, REVIEWARENA enables these analyses at a scale that previously required organizer access.

Venue	Years	Papers	Notes
NeurIPS	2021–25	18,453	main + 2023 D&B
ICLR	2020–26	22,532	largest split
ICML	2025	3,257	public reviews only in 2025
CoRL	2021–24	813	robotics-specific forms
COLM	2024–25	717	compact accept/reject reasons
EMNLP	2023	2,009	`Excitement` as rating
TMLR	rolling	3,748	no numeric overall score
Total	–	**51,529**	**196,099 reviews**

Table 1. Dataset coverage. REVIEWARENA spans seven public OpenReview venues and fourteen native review-form variants.

3. The REVIEWARENA Dataset

3.1. Scope and Coverage

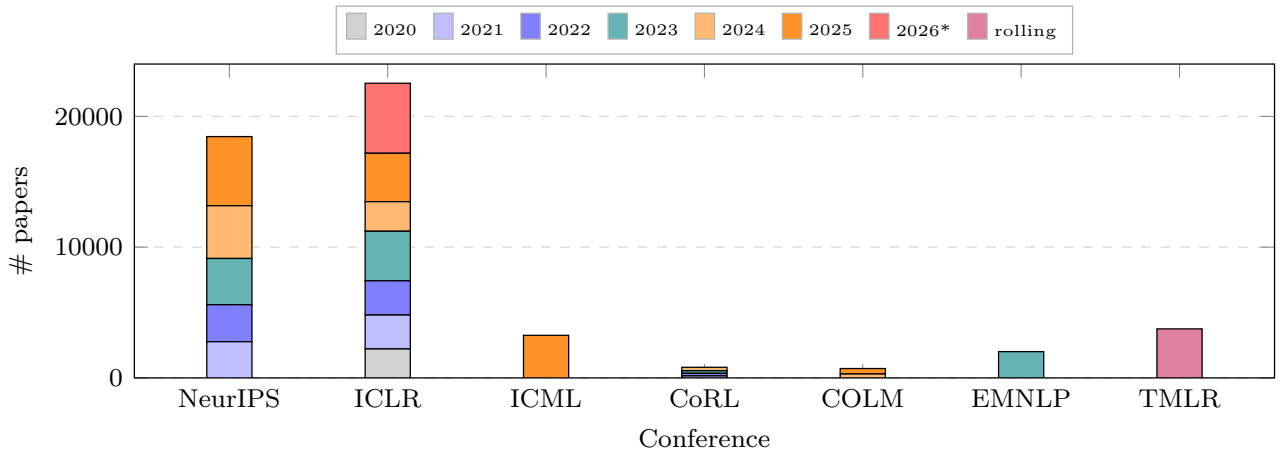
REVIEWARENA contains public OpenReview papers and reviews from seven venues: NeurIPS (2021–2025, including 2023 Datasets & Benchmarks), ICLR (2020–2026), ICML 2025, CoRL (2021–2024), COLM (2024–2025), EMNLP 2023, and TMLR. We audited 56 plausible conference prefixes across the OpenReview v1 and v2 APIs; Appendix A.1 gives the full audit. The resulting scope is exhaustive for public, comparable OpenReview review forms at the time of writing. Table 1 summarizes counts by venue; Figure 2a shows the venue$\times$year composition and Figure 2b shows how submissions consolidate into coarse topic buckets via author-selected primary areas (when exposed). The venue imbalance highlights why cross-venue benchmarks must stratify or normalize by venue/year; area imbalance similarly motivates stratifying analyses beyond headline aggregates.

3.2. Collection Pipeline

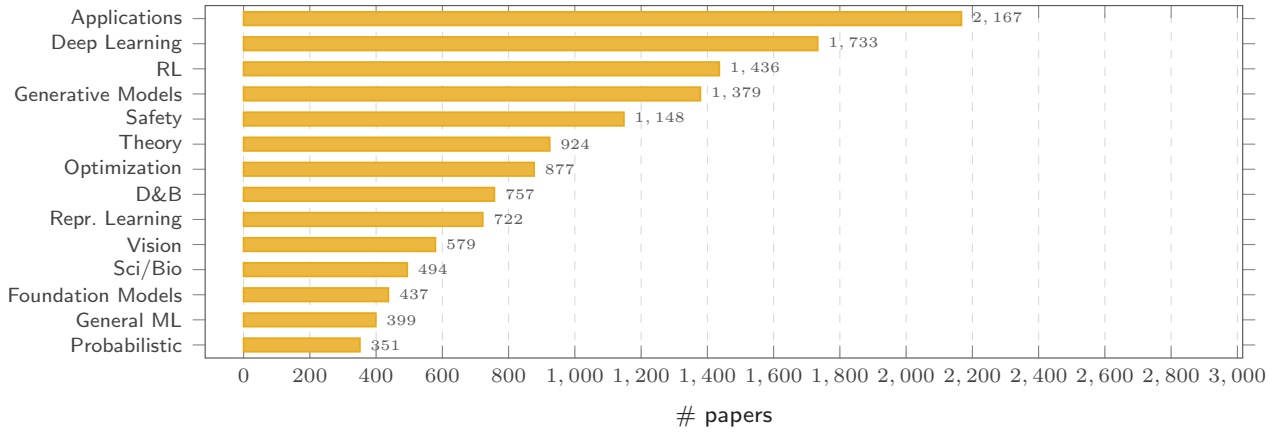
The collection pipeline harvests public records through OpenReview's programmatic interfaces (Team, 2026). Venue-years span multiple API generations and invitation encodings, so the importer normalizes heterogeneous notes into the unified paper row described below. Work is partitioned across parallel Modal workers; each worker maintains its own authenticated session and processes a disjoint set of forums. For every forum we retrieve metadata, the PDF, official reviews, rebuttals, comments, meta-reviews, and decisions. Each venue-year sweep is written atomically to one JSONL snapshot before packaging the corpus as Hugging Face `PdfFolder` (Lhoest et al., 2021).

3.3. Schema and Review Forms

Each paper row contains top-level metadata (`conference`, `year`, `track`, title, abstract, authors, areas, decision), the



(a) Papers by conference and year (counts from venue-year JSONL exports; NeurIPS 2023 stacks main and Datasets & Benchmarks). Totals match Table 1. Empty segments denote years without a comparable public OpenReview slice; 2026* is an in-progress ICLR cycle at scrape time.



(b) Papers per consolidated `primary_area` (venues that expose the field). Labels were collapsed cross-venue into fourteen buckets; a heterogeneous residual bucket (10,492 papers) is omitted here but retained in metadata.

PDF path, a list of structured reviews, official comments, rebuttals, and meta-review fields. Review forms differ sharply across venues, so the parser emits a union schema rather than a lossy common denominator. Common fields such as `summary`, `strengths`, `weaknesses`, `rating`, and `confidence` are normalized when present; venue-specific sub-scores such as NeurIPS 2025 `quality`/`clarity`/`significance`/`originality`, CoRL's robotics axes, and EMNLP's `Excitement` field are preserved; unmapped scalar and long-text fields are stored in `extra_scores` and `extra_text`. TMLR has no numeric overall score and is therefore retained for non-rating tasks rather than forced into an artificial scale.

The review-form variation is a feature of the dataset rather than a nuisance to hide. NeurIPS 2023/2024 and recent ICLR share a relatively standard `soundness`/`presentation`/`contribution` structure; NeurIPS 2025 changes the axes to `quality`/`clarity`/`significance`/`originality`; ICML 2025 replaces the usual `rating` field with `overall_recommendation`; CoRL has both textual and numeric recommendation forms; COLM uses compact reasons-for and reasons-against fields; EMNLP 2023 uses CamelCase fields and treats `Excitement` as the overall score; and TMLR uses Yes/No claim-evidence questions. This heterogeneity is exactly why REVIEWARENA-EVAL uses native venue-year prompts and why the raw REVIEWARENA release preserves all fields rather than projecting them into a single synthetic form.

For NeurIPS 2025 we additionally expose `was_revised` and `final_justification`, indicating whether a review was edited after rebuttal and what justification the reviewer supplied. This makes REVIEWARENA useful not only for static review generation, but also for studying how

reviewer judgments change after author response. To our knowledge, no prior public peer-review dataset exposes this post-rebuttal revision signal at review level together with the original paper, review text, rebuttal, and final decision.

3.4. Comparison to Prior Resources

Table 2 situates REVIEWARENA within the landscape of large-scale public peer-review resources. While complementary datasets have made essential contributions, such as the broad venue coverage and dialogue-focused structure of Re^2 (Zhang et al., 2025) or the scaled generation corpora of Reviewer2 (Gao et al., 2024) and ReviewMT (Tan et al., 2024), REVIEWARENA's design is distinct. It prioritizes depth and fidelity for major ML venues over breadth across all of OpenReview. Rather than normalizing diverse review processes into a unified format for training, REVIEWARENA preserves the native, structured review forms, discussion threads, and post-rebuttal revision signals for each venue-year. This focus on capturing the complete, heterogeneous reality of the review lifecycle makes REVIEWARENA a robust substrate for benchmarking: it evaluates models not on a simplified, generic reviewing task, but on their ability to adapt to the specific forms and conventions that human reviewers actually use.

4. REVIEWARENA-EVAL

REVIEWARENA supports many process-analysis tasks, but a dataset paper is most useful when it also defines a concrete benchmark. We instantiate the LLM-as-reviewer task because it is both timely and demanding: the model must read a full paper, complete the structured review form associated with that paper's venue-year cell, and produce outputs that can be compared to human reviewers. This benchmark intentionally evaluates agreement with the historical review process rather than objective paper quality. Human reviews are noisy and sometimes inconsistent, but they are the relevant target for reviewer-assistance systems that claim to approximate or support human reviewing.

4.1. Task Formulation

Given a paper p from venue-year cell (v, y) and a template $T_{v,y}$ matching that venue's official review form, model M produces

$$\hat{r} = M(T_{v,y}, p) = \big(\hat{s}, \hat{c}, \hat{\mathbf{a}}_{v,y}, \hat{\mathbf{t}}_{v,y}\big),$$

where $\hat{s}$ is the native overall rating, $\hat{c}$ is confidence, $\hat{\mathbf{a}}_{v,y}$ are sub-scores, and $\hat{\mathbf{t}}_{v,y}$ are free-text fields. Targets are the corresponding human-review fields in REVIEWARENA.

4.2. Benchmark Subset and Prompts

REVIEWARENA-EVAL contains 1,002 papers: exactly 167 from each numerically rated venue (NeurIPS, ICLR, ICML, CoRL, COLM, EMNLP). We use uniform per-venue allocation so ICLR and NeurIPS do not dominate aggregate metrics. Within a venue, papers are sampled uniformly over the released years. TMLR is excluded from this benchmark because it has no numeric overall score, but remains in the dataset for claim-evidence and discussion tasks.

A generic "rate from 1–10" prompt is invalid for half the benchmark. ICML 2025 uses 1–5, CoRL uses 1–4, EMNLP uses `Excitement` 1–5, and only NeurIPS/ICLR/COLM use 1–10. Sub-score axes also vary: NeurIPS 2025 uses `quality`/`clarity`/`significance`/`originality`; earlier NeurIPS and recent ICLR use `soundness`/`presentation`/`contribution`; CoRL has robotics-specific dimensions. We therefore render prompts by (v, y) with native scale endpoints, field names, sub-score axes, and accept/reject threshold. The template is not meant as prompt engineering; it is a correctness condition. If the venue asked human reviewers for a 1–4 recommendation, the model is also asked for a 1–4 recommendation.

4.3. Models and Metrics

We evaluate GPT-OSS-120B (Agarwal et al., 2025), Kimi K2.6 (Moonshot AI, 2026), Qwen3.6-Plus (Qwen Team, 2026), GLM-5.1 (Zhipu AI, 2026), Gemma-4-26B-A4B-IT (Gemma Team, 2026), and DeepSeek-V4-Pro (DeepSeek-AI, 2026), served through Fireworks AI at temperature 0. PDFs are converted once with Nemotron-Nano-OCR (NVIDIA, 2026) on Modal H100 workers and truncated to 50,000 characters when needed. Keeping extraction fixed across models avoids confounding model differences with OCR differences. Appendix A.7 gives a rendered prompt and Appendix A.11 reports cost.

We report several complementary metrics because a single number can mischaracterize reviewer behaviour. Overall-rating mean absolute error and signed bias measure score agreement on each venue's native scale. Sub-score MAE measures agreement on axes such as `soundness`, `presentation`, `contribution`, `confidence`, or their venue-specific equivalents when available. Free-text agreement is measured with ROUGE, BLEU, and BERTScore over summary, strengths, and weaknesses fields. For decisions, we report ICLR-only accept/reject accuracy and macro-F1 because ICLR is the only benchmark slice with many public rejections. Finally, calibration and range diagnostics measure whether a model uses the full rating scale or compresses scores into a narrow middle band.

Dataset	Papers	Reviews	Rebuttals	Discussion	Sources	Decisions
PeerRead (Kang et al., 2018)	14,700	10,700	0	–	yes	yes
ASAP-Review (Yuan et al., 2022)	8,877	28,119	0	–	yes	no
NLPeer (Dycke et al., 2023)	5,672	11,515	0	–	yes	no
Reviewer2 (Gao et al., 2024)	27,805	99,727	0	–	yes	no
ReviewMT (Tan et al., 2024)	26,841	92,017	0	yes	yes	yes
Re^2 (Zhang et al., 2025)	19,926	70,668	53,818	yes	yes	yes
REVIEWARENA	**51,529**	**196,099**	**77,198**	**472,067**	**yes**	**yes**

Table 2. REVIEWARENA vs. representative public peer-review datasets. Counts follow the cited papers where available. "Discussion" denotes post-review discussion/comment threads when separately exposed. The REVIEWARENA rebuttal count is the sum of 71,175 per-review rebuttals and 6,023 paper-level author rebuttals; discussion counts 472,067 official comment notes.

5. Results and Analysis

5.1. Overall Rating Agreement

Each model reviews every paper in REVIEWARENA-EVAL once under the venue-year-aware prompt. We compute numeric metrics against the mean human score for the paper and text metrics against the concatenated human text for the same field. The main result is not that one model "solves" reviewing, but that different models fail in different ways. Table 3 reports overall-rating MAE and signed bias by venue. Table 4 compresses the major axes into a single comparison.

On the native scales, Qwen3.6-Plus has the best pooled overall-rating MAE (0.700), followed closely by Kimi K2.6 (0.725). The ranking is tight among the five better-calibrated models: GPT-OSS-120B, DeepSeek-V4-Pro, and GLM-5.1 are within 0.14 pooled MAE of Qwen3.6-Plus. Gemma-4-26B is different. It inflates scores on every venue, with signed bias of $+2.03$ on NeurIPS and $+1.97$ on ICLR, and therefore has the worst pooled numeric performance despite strong free-text similarity. The per-venue signs are also informative. Most models are positively biased on NeurIPS and ICLR, whose human ratings concentrate below the apparent model prior around the middle of the 1–10 scale, but neutral or negative on ICML, CoRL, COLM, and EMNLP. This is why native-scale reporting matters: a CoRL error of 0.65 on a 1–4 scale is not comparable to a NeurIPS error of 0.65 on a 1–10 scale.

5.2. Cross-Axis Behaviour

The sub-score and text metrics sharpen the picture. Kimi K2.6 has the best average sub-score MAE (0.449), just ahead of Qwen3.6-Plus (0.457), suggesting that the two strongest numeric models are close but not identical: Qwen is slightly better on the overall recommendation, while Kimi better tracks the auxiliary axes. GPT-OSS-120B is competitive on overall ratings but weaker on sub-scores, which suggests that matching the final recommendation does not imply matching the reasoning axes reviewers expose. This distinction is important for reviewer-assistance systems because a plausible overall score with inconsistent sub-scores is difficult for authors or area chairs to interpret.

Free-text metrics disagree with the numeric ranking. Gemma-4-26B and GLM-5.1 tie for best BERTScore-F1 (0.838), but Gemma-4-26B is the worst numeric model and GLM-5.1 is not the best by pooled MAE. Conversely, Qwen3.6-Plus is the best numeric model but not the best text mimic. The most likely explanation is that text overlap rewards generic reviewer phrasing, summaries of visible paper content, and common critique templates, whereas numeric calibration requires placing the paper relative to a venue-specific review distribution. A benchmark that reports only text similarity would therefore crown a different model from one that reports rating agreement.

5.3. Decision Discrimination

Decision discrimination is also weak. Most models separate accepted from rejected ICLR papers by only 0.47–0.66 predicted points on a 1–10 scale. GLM-5.1 is the exception: it reaches macro-F1 0.722 and a 1.40-point accept/reject gap. Gemma-4-26B has high text similarity but accepts nearly every ICLR paper, illustrating why textual overlap is not enough. We report decision metrics only on ICLR because most other public OpenReview splits are dominated by accepted papers; pooling decision accuracy across venues would mostly measure the class prior.

5.4. Calibration and Score Compression

The most consistent behavioural pattern is score compression, as illustrated in Figure 3. Five of the six models avoid the extremes of every rating axis. On 1–10 venues, GPT-OSS-120B, Kimi K2.6, Qwen3.6-Plus, and DeepSeek-V4-Pro rarely use ratings below 4 or above 7; on ICML and EMNLP they top out at 4/5; on CoRL they top out at 3/4. As shown by its steeper slope in Figure 3, GLM-5.1 is more discriminative, while Gemma-4-26B compresses into the upper part of every scale, showing a significant positive bias even for papers with low human ratings. This explains the positive bias on NeurIPS/ICLR and the small accept-vs-reject gap for most models. It also agrees with prior observa-

	MAE (signed bias) on native integer scale					
Model	NeurIPS	ICLR	ICML	CoRL	COLM	EMNLP
Scale	1–10	1–10	1–5	1–4	1–10	1–5
GPT-OSS-120B	1.08 (+0.56)	1.06 (+0.36)	0.58 (−0.36)	0.65 (−0.57)	0.79 (−0.46)	0.52 (−0.36)
Kimi K2.6	1.09 (+0.68)	1.03 (+0.49)	0.53 (−0.28)	0.41 (−0.27)	0.74 (−0.40)	0.56 (−0.50)
Qwen3.6-Plus	1.17 (+0.87)	1.02 (+0.51)	0.55 (−0.02)	0.43 (−0.31)	0.61 (−0.14)	0.44 (−0.25)
GLM-5.1	1.23 (+0.94)	1.17 (+0.75)	0.61 (+0.06)	0.44 (−0.22)	0.72 (−0.03)	0.88 (−0.81)
DeepSeek-V4-Pro	1.08 (+0.58)	0.99 (+0.23)	0.59 (−0.27)	0.46 (−0.31)	0.86 (−0.52)	0.87 (−0.83)
Gemma-4-26B	2.04 (+2.03)	2.02 (+1.97)	0.69 (+0.67)	0.55 (+0.44)	1.26 (+1.25)	0.45 (+0.22)

Table 3. Per-venue overall-rating MAE and signed bias on the 167-paper slice for each venue. Columns use native rating scales, so cross-venue MAE should be read together with the scale row.

Model	**Pooled MAE**	**Sub-score MAE**	**BERTScore-F1**	**ICLR F1**	$\Delta_{\text{a-r}}$
GPT-OSS-120B	0.783	0.615	0.831	0.687	0.59
Kimi K2.6	0.725	**0.449**	0.834	0.613	0.47
Qwen3.6-Plus	**0.700**	0.457	0.833	0.675	0.62
GLM-5.1	0.837	0.488	**0.838**	**0.722**	**1.40**
DeepSeek-V4-Pro	0.807	0.603	0.837	0.655	0.66
Gemma-4-26B	1.169	0.799	**0.838**	0.394	0.48

Table 4. Model summary across axes. Pooled MAE averages venue-level overall-rating MAE; sub-score MAE averages available integer sub-score axes; BERTScore-F1 is pooled over summary/strength/weakness text fields; ICLR F1 and $\Delta_{\text{a-r}}$ evaluate binary decision discrimination on the only slice with many public rejections.

tions that LLM reviewers can produce reasonable-looking critiques while avoiding strong negative judgments (Liang et al., 2024b).

The benchmark therefore suggests three evaluation lessons. First, accept/reject decisions require ranking, not just calibration: Qwen3.6-Plus has the best pooled MAE but GLM-5.1 has the strongest ICLR accept/reject separation, likely due to the better calibration curve seen in Figure 3. Second, text similarity is not reviewer reliability: Gemma-4-26B looks strong by BERTScore but weak by ratings and decisions. Third, reproducibility depends on mundane normalization choices. Decision strings differ even within a venue-year family, and similar normalization is needed for area labels, track names, and legacy review invitations. We release `decision_normalized` and a fixed benchmark split so downstream work does not silently compare against a different task.

6. Availability and Responsible Use

6.1. Access and Maintenance

REVIEWARENA is released as a Hugging Face dataset with one split per venue and a separate benchmark split for REVIEWARENA-EVAL. Each row in `metadata.jsonl` follows the standard `PdfFolder` layout and links to the original PDF together with normalized metadata, reviews, rebuttals, discussions, meta-reviews, and decisions. Keeping the original PDFs avoids committing to a single OCR representation and enables future re-extraction pipelines.

The benchmark release includes the fixed 1,002-paper ID list, prompt templates, model predictions, and evaluation code. We separate the evolving full corpus from the frozen benchmark split so future LLM-reviewer systems can compare directly against the results reported in this paper. Future dataset updates will be versioned while preserving the original REVIEWARENA-EVAL benchmark IDs and metrics.

Maintenance primarily covers three cases: adding new public OpenReview venue-years, correcting parsing issues while preserving the public schema, and honoring takedown or correction requests through the dataset repository.

We also release the scraping and evaluation pipeline for reproducibility. The collection code records the OpenReview forum IDs and invitation types used for each sample, while the evaluation code stores the prompt template, model identifier, OCR text, and parsed outputs for each model-paper pair.

6.2. Limitations and Ethics

REVIEWARENA is dominated by accepted papers because most venues only make accepted submissions public. ICLR is the primary exception, so accept/reject evaluation is mainly ICLR-scoped. As a result, the dataset is better suited for studying reviewer behaviour, calibration, review text, and rebuttal dynamics than estimating the full distribution of submissions at a venue.

The dataset currently excludes major review ecosystems that are not publicly available on OpenReview, including

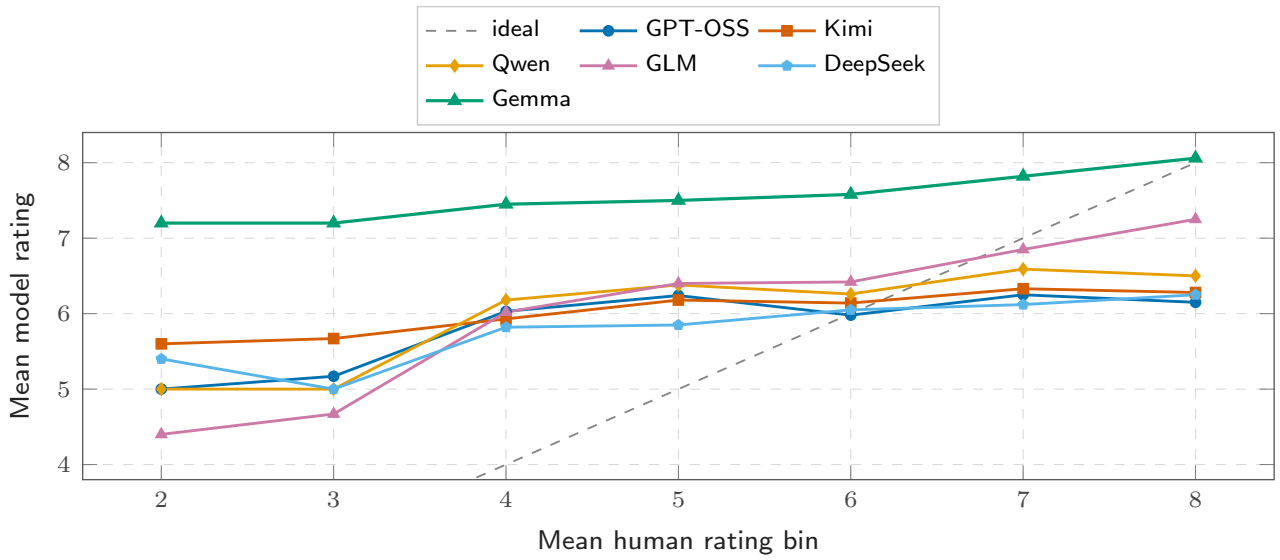


Figure 3. **Calibration behaviour on REVIEWARENA-EVAL.** Mean predicted rating as a function of binned mean human rating on the three venues that use a 1–10 scale (NeurIPS, ICLR, COLM). Curves near-horizontal around 5–7 indicate score compression; Gemma is consistently inflated, while GLM tracks the human-rating bins most strongly.

CVPR/ICCV/ECCV/WACV, AAAI/IJCAI, and post-2023 ARR-based NLP reviewing. The benchmark should therefore be understood as an OpenReview-based benchmark rather than a complete representation of scientific peer review.

Although review forms from multiple venues are unified into a common schema, not all scores are directly comparable. For example, TMLR lacks numeric overall ratings, while CoRL and EMNLP use different scoring scales. We therefore preserve venue-specific fields and report native-scale metrics instead of forcing all scores into a single normalized scale.

We redistribute only publicly available OpenReview content and do not attempt de-anonymization. Reviewer identifiers are preserved only as released pseudonyms, and the dataset repository documents takedown and acceptable-use procedures.

Finally, REVIEWARENA is intended as an evaluation resource, not as evidence that LLMs should replace human reviewers. Our results highlight calibration issues, score compression, and weak decision discrimination. The intended use of REVIEWARENA is to measure and reduce such failure modes under human oversight, not to automate acceptance decisions.

7. Conclusion

We presented REVIEWARENA, a public OpenReview dataset with 51,529 papers and 196,099 reviews from seven venues, plus REVIEWARENA-EVAL, a 1,002-paper benchmark for evaluating LLMs under native venue-year review forms. The benchmark shows that current open-weight LLMs remain unreliable as standalone reviewers: the best pooled rating MAE is 0.70, most models compress scores toward the middle of each scale, accept/reject separation is weak outside GLM-5.1, and strong text similarity does not imply numeric calibration.

More broadly, REVIEWARENA is meant as infrastructure for studying reviewing as a structured process. Because it preserves PDFs, scores, text fields, rebuttals, meta-reviews, decisions, and revision metadata, it can support work on reviewer disagreement, score drift, rebuttal effectiveness, calibration across venues, and safer human-controlled reviewer assistance. The dataset and benchmark do not remove the need for human judgment; they provide a public way to measure where automated reviewers diverge from it.

Impact Statement

This paper presents work whose goal is to advance the field of machine learning and AI-assisted scientific peer review. Potential societal impacts include improving review scalability, consistency, and accessibility. However, automated review systems may also amplify existing biases, incentivize gaming behaviors, or be misused in high-stakes evaluation settings without adequate human oversight. We therefore view systems such as REVIEWARENA as research tools intended to support analysis of LLM-based reviewing, rather than replacements for expert human peer review.

References

Agarwal, S., Ahmad, L., Ai, J., Altman, S., Applebaum, A., Arbus, E., Arora, R. K., Bai, Y., Baker, B., Bao, H., et al. gpt-oss-120b & gpt-oss-20b model card. *arXiv preprint arXiv:2508.10925*, 2025.

Bharti, P. K., Ranjan, S., Ghosal, T., Agrawal, M., and Ekbal, A. PEERAssist: Leveraging on paper-review interactions to predict peer review decisions. In *Towards Open and Trustworthy Digital Societies: 23rd International Conference on Asia-Pacific Digital Libraries, ICADL 2021*, 2021.

D'Arcy, M., Hope, T., Birnbaum, L., and Downey, D. MARG: Multi-agent review generation for scientific papers. *arXiv preprint arXiv:2401.04259*, 2024a.

D'Arcy, M., Ross, A., Bransom, E., Kuehl, B., Bragg, J., Hope, T., and Downey, D. Aries: A corpus of scientific paper edits made in response to peer reviews. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 6985–7001, 2024b.

DeepSeek-AI. DeepSeek-V4: Towards highly efficient million-token context intelligence. *DeepSeek-AI Blog Post*, 2026.

Dycke, N., Kuznetsov, I., and Gurevych, I. 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)*, 2023.

Gao, X., Ruan, J., Zhang, Z., Gao, J., Liu, T., and Fu, Y. MMReview: A multidisciplinary and multimodal benchmark for LLM-based peer review automation. *arXiv preprint arXiv:2508.14146*, 2025.

Gao, Z., Brantley, K., and Joachims, T. Reviewer2: Optimizing review generation through prompt generation. *arXiv preprint arXiv:2402.10886*, 2024.

Gemma Team. Gemma 4: Byte for byte, the most capable open models. Google DeepMind Blog; Hugging Face model card, 2026. URL https://huggingface.co/google/gemma-4-26B-A4B-it.

Ghosal, T., Kumar, S., Bharti, P. K., and Ekbal, A. Peer review analyze: A novel benchmark resource for computational analysis of peer reviews. *PLOS ONE*, 2022.

Guo, Y., Shang, G., Rennard, V., Vazirgiannis, M., and Clavel, C. Automatic analysis of substantiation in scientific peer reviews. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pp. 10198–10216, 2023.

Idahl, M. and Ahmadi, Z. 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)*, 2025.

Kang, D., Ammar, W., Dalvi, B., Van Zuylen, M., Kohlmeier, S., Hovy, E., and Schwartz, R. 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)*, 2018.

Kennard, N. N., O'Gorman, T., Das, R., Sharma, A., Bagchi, C., Clinton, M., Yelugam, P. K., Zamani, H., and McCallum, A. DISAPERE: A dataset for discourse structure in peer review discussions. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, 2022.

Kuznetsov, I., Afzal, O. M., Dercksen, K., Dycke, N., Goldberg, A., Hope, T., Hovy, D., Kummerfeld, J. K., Lauscher, A., Leyton-Brown, K., et al. What can natural language processing do for peer review? *arXiv preprint arXiv:2405.06563*, 2024.

Lawrence, N. and Cortes, C. The NIPS experiment. *Blog post*, 2014.

Lhoest, Q., del Moral, A. V., Jernite, Y., Thakur, A., von Platen, P., Patil, S., Chaumond, J., Drame, M., Plu, J., Tunstall, L., et al. Datasets: A community library for natural language processing. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, 2021.

Liang, W., Izzo, Z., Zhang, Y., Lepp, H., Cao, H., Zhao, X., Chen, L., Ye, H., Liu, S., Huang, Z., McFarland, D. A., and Zou, J. Monitoring AI-modified content at scale: A case study on the impact of ChatGPT on AI conference peer reviews. In *International Conference on Machine Learning*, 2024a. URL https://arxiv.org/pdf/2403.07183.pdf.

Liang, W., Zhang, Y., Cao, H., Wang, B., Ding, D. Y., Yang, X., Vodrahalli, K., He, S., Smith, D. S., Yin, Y., et al. Can large language models provide useful feedback on research papers? a large-scale empirical analysis. *NEJM AI*, 1(8):AIoa2400196, 2024b.

Lin, J., Song, J., Zhou, Z., Chen, Y., and Shi, X. MOPRD: A multidisciplinary open peer review dataset. *Neural Computing and Applications*, 35(34):24191–24206, 2023.

Liu, R. and Shah, N. B. ReviewerGPT? an exploratory study on using large language models for paper reviewing. *arXiv preprint arXiv:2306.00622*, 2023.

Lu, C., Lu, C., Lange, R. T., Foerster, J., Clune, J., and Ha, D. The AI scientist: Towards fully automated open-ended scientific discovery. *arXiv preprint arXiv:2408.06292*, 2024.

Moonshot AI. Kimi K2.6 tech blog: Advancing open-source coding. `https://huggingface.co/moonshotai/Kimi-K2.6`, 2026.

NVIDIA. NVIDIA Nemotron-OCR-v2: High-performance document intelligence. `https://huggingface.co/nvidia/nemotron-ocr-v2`, 2026.

Qwen Team. Qwen3.6-35B-A3B: Agentic coding power, now open to all. `https://qwen.ai/blog?id=qwen3.6-35b-a3b`, 2026.

Shah, N. B. Challenges, experiments, and computational solutions in peer review. *Communications of the ACM*, 65(6):76–87, 2022.

Shen, C., Cheng, L., Zhou, R., Bing, L., You, Y., and Si, L. MReD: A meta-review dataset for structure-controllable text generation. In *Findings of the Association for Computational Linguistics: ACL 2022*, 2022.

Stelmakh, I., Shah, N. B., Singh, A., and Daumé III, H. A novice-reviewer experiment to address scarcity of qualified reviewers in large conferences. In *Proceedings of the AAAI Conference on Artificial Intelligence*, 2021.

Tan, C., Lyu, D., Li, S., Gao, Z., Wei, J., Ma, S., Liu, Z., and Li, S. Z. Peer review as a multi-turn and long-context dialogue with role-based interactions. *arXiv preprint arXiv:2406.05688*, 2024.

Team, O. Openreview. `https://openreview.net`, 2026.

Weng, Y., Zhu, M., Bao, G., Zhang, H., Wang, J., Zhang, Y., and Yang, L. CycleResearcher: Improving automated research via automated review. In *The Thirteenth International Conference on Learning Representations*, 2025.

Wu, P.-C., Yen, A.-Z., Huang, H.-H., and Chen, H.-H. Incorporating peer reviews and rebuttal counter-arguments for meta-review generation. In *Proceedings of the 31st ACM International Conference on Information & Knowledge Management*, 2022.

Yu, J., Ding, Z., Tan, J., Luo, K., Weng, Z., Gong, C., Zeng, L., Cui, R., Han, C., Sun, Q., Wu, Z., Lan, Y., and Li, X. Automated peer reviewing in paper SEA: Standardization, evaluation, and analysis. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, 2024.

Yu, S., Luo, M., Madasu, A., Lal, V., and Howard, P. Is your paper being reviewed by an LLM? benchmarking AI text detection in peer review. In *Findings of the Association for Computational Linguistics: EMNLP 2025*, 2025. URL `https://arxiv.org/abs/2502.19614`.

Yuan, W., Liu, P., and Neubig, G. Can we automate scientific reviewing? *Journal of Artificial Intelligence Research*, 75, 2022.

Zhang, D., Bao, Z., Du, S., Zhao, Z., Zhang, K., Bao, D., and Yang, Y. Re^2: A consistency-ensured dataset for full-stage peer review and multi-turn rebuttal discussions. *arXiv preprint arXiv:2505.07920*, 2025.

Zhipu AI. GLM-5: From vibe coding to agentic engineering, 2026.

Zhou, L., Zhang, R., Dai, X., Hershcovich, D., and Li, H. Large language models penetration in scholarly writing and peer review. *arXiv preprint arXiv:2502.11193*, 2025.

Zhou, R., Chen, L., and Yu, K. 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)*, 2024.

Zhu, M., Weng, Y., Yang, L., and Zhang, Y. 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)*, 2025.

A. Appendix

A.1. Venue-Discovery Audit

We audited 56 OpenReview venue prefixes across the modern v2 API and the legacy v1 API. A venue-year was considered in scope if (i) submissions could be enumerated, (ii) the forum contained public `Official_Review` or `Review` notes, and (iii) the review form was close enough to the dominant machine-learning review form to support a unified schema. This audit yielded the seven conference or journal venues used in REVIEWARENA: NeurIPS, ICLR, ICML, CoRL, COLM, EMNLP, and TMLR. Major computer-vision venues (CVPR, ICCV, ECCV, WACV) and AAAI/IJCAI use non-OpenReview systems; the ACL family primarily routes through ARR, whose reviews are private; and ICML 2023/2024 host submissions on OpenReview but do not release reviews publicly. Smaller OpenReview venues such as UAI, MIDL, and AutoML were not included in the benchmark because they have few compatible public years and review forms that diverge substantially from the target LLM-reviewer task.

Venue family	Public review years in REVIEWARENA	API	Notes
NeurIPS	2021, 2022, 2023, 2023 D&B, 2024, 2025	v1/v2	main + D&B tracks
ICLR	2020–2026	v1/v2	public accept/reject decisions
ICML	2025	v2	earlier public submissions lack public reviews
CoRL	2021–2024	v1/v2	legacy textual and modern numeric forms
COLM	2024–2025	v2	compact accept/reject reason form
EMNLP	2023	v2	later years moved to ARR/private reviews
TMLR	rolling	v2	journal-style claim-evidence review

Table 5. Venue families included in REVIEWARENA. The released corpus consists of 22 venue-year/track JSONL files spanning these seven public OpenReview conference or journal venues.

A.2. Modal Volume Files and Interaction Counts

The source records used to build REVIEWARENA are stored as one JSONL file per venue-year/track on a Modal volume. Table 6 reports the interaction counts computed directly from those JSONL files. We distinguish *per-review rebuttals*, which are attached to a specific review through the review's `replyto` relation; *paper-level author rebuttals*, which are not review-specific; and *official comments*, which include public discussion notes among authors, reviewers, and area chairs.

Group	Papers	Reviews	Review rebuttals	Paper rebuttals	Comments
NeurIPS 2021–2025	18,453	74,078	51,286	6,023	168,155
ICLR 2020–2026	22,532	85,866	0	0	263,854
ICML 2025	3,257	12,478	12,410	0	0
CoRL 2021–2024	813	2,908	0	0	4,587
COLM 2024–2025	717	2,735	1,145	0	6,824
EMNLP 2023	2,009	6,413	6,334	0	0
TMLR rolling	3,748	11,621	0	0	29,647
Total	**51,529**	**196,099**	**71,175**	**6,023**	**472,067**

Table 6. Interaction counts from the Modal volume JSONLs. The main-paper comparison table reports 77,198 rebuttal entries, equal to 71,175 per-review rebuttals plus 6,023 paper-level rebuttals, and 472,067 official discussion comments.

A.3. Released Schema

Each row in REVIEWARENA is a paper-level record. The released Hugging Face dataset stores these rows using a `PdfFolder`-compatible layout: the `file_name` points to the PDF, while all metadata is stored in JSONL. The key fields are:

```
{
  "forum_id": "OpenReview forum id",
  "conference": "neurips | iclr | icml | corl | colm | emnlp | tmlr",
  "year": 2025,
  "track": "main | datasets_and_benchmarks | rolling",
  "venue_id": "OpenReview venue id",
```

```
  "paper_number": 1234,
  "title": "...",
  "abstract": "...",
  "authors": ["..."],
  "keywords": ["..."],
  "tldr": "...",
  "primary_area": "...",
  "venue": "raw venue string from OpenReview",
  "decision": "raw decision string",
  "decision_comment": "area-chair / meta-review text",
  "author_rebuttal": "paper-level rebuttal if present",
  "num_reviews": 4,
  "reviews_json": "[{...}, {...}]",
  "markdown": "OCR / extracted paper text",
  "markdown_chars": 18342
}
```

The nested `reviews_json` field is a JSON-encoded list. Each review object contains common fields and venue-specific extensions. The following example is schematic but mirrors the released structure:

```
{
  "review_id": "note id",
  "reviewer": "Reviewer_AbCd",
  "summary": "...",
  "strengths": "...",
  "weaknesses": "...",
  "questions": "...",
  "limitations": "...",
  "rating": 6,
  "confidence": 4,
  "soundness": 3,
  "presentation": 3,
  "contribution": 2,
  "quality": null,
  "clarity": null,
  "significance": null,
  "originality": null,
  "rebuttal": "author reply to this review, if present",
  "was_revised": false,
  "final_justification": "",
  "extra_scores": {"time_spent_reviewing": "2 hours"},
  "extra_text": {"unmapped_long_field": "..."}
}
```

A.4. Review-Form Mapping

The parser maps native OpenReview fields into a union schema. It never discards unmapped fields: short scalar fields are placed in `extra_scores`, and long text fields are placed in `extra_text`. Table 7 summarizes the most important native fields.

A.5. Venue-Specific Score Fields and Guidelines

The central design choice in REVIEWARENA-EVAL is that models are not asked to produce a generic review. They are asked to complete the score fields that the source venue actually used. Table 8 lists the native overall field, rating scale, auxiliary score fields, free-text fields, and decision threshold used in the benchmark prompts. The exact field names matter

Venue/form	Native overall field	Important axes retained
NeurIPS 2023/24, ICLR 2024–26	`rating`	`soundness`, `presentation`, `contribution`, `confidence`
NeurIPS 2025	`rating`	`quality`, `clarity`, `significance`, `originality`, revisions
ICML 2025	`overall_recommendation`	ten evidence/methodology text fields, `confidence`
CoRL 2021–2023	`recommendation` (text)	free-text `review`; textual ratings mapped to 1–4
CoRL 2024	`recommendation`	`technical_quality`, `robotics_focus`, `potential_impact`
COLM 2024–2025	`rating`	`reasons_to_accept`, `reasons_to_reject`
EMNLP 2023	`Excitement`	`Soundness`, `Reproducibility`, `Reviewer_Confidence`
TMLR	none	`claims_and_evidence`, `audience`, requested changes

Table 7. Native review-form fields mapped into the REVIEWARENA union schema.

because they determine both model output parsing and metric computation.

Venue	Overall field	Scale / threshold	Auxiliary score fields	Main text fields
NeurIPS/ICLR/COLM	`rating`	1–10; accept if ≥ 6	soundness; presentation; contribution; confidence	summary; strengths; weaknesses; questions; limitations
NeurIPS 2025	`rating`	1–10; accept if ≥ 6	quality; clarity; significance; originality; confidence	summary; strengths/weaknesses; questions; limitations; final justification
ICML 2025	`overall_recommend.`	1–5; accept if ≥ 3	confidence; no integer sub-scores	claims/evidence; methods/evaluation; theoretical claims; questions for authors
CoRL	`recommendation`	1–4; accept if ≥ 3	technical quality; clarity; impact; robotics focus; confidence	paper summary; recommendation summary; rebuttal questions; review text
EMNLP 2023	`Excitement`	1–5; accept if ≥ 3	Soundness; Reproducibility; Reviewer Confidence	topic/contributions; reasons to accept/reject; author questions
TMLR	none	No numeric overall score	claims/evidence; audience; confidence	contribution summary; claims explanation; requested changes; comments

Table 8. Native score and text fields preserved by REVIEWARENA and rendered in REVIEWARENA-EVAL prompts. TMLR is included in the dataset but excluded from REVIEWARENA-EVAL because it has no numeric overall rating.

Shared reviewer guidelines inserted before each venue-specific score block

- Provide substantive, evidence-grounded feedback; cite specific sections, equations, figures, claims, or experiments when making judgments.
- Assess technical soundness, clarity, significance, and originality where those axes are meaningful for the venue form.
- Value conceptual contribution and new knowledge over raw benchmark wins; strong empirical results alone do not justify a high score.
- Be constructive and fair; distinguish fundamental flaws from fixable presentation issues and state what evidence would change the assessment.
- Calibrate to the venue's bar and native scale. Do *not* use a generic NeurIPS 1–10 score unless the venue actually uses that scale.
- Avoid score inflation out of politeness and avoid deflating scores for minor presentation problems.
- Return only valid JSON matching the requested venue schema.

Figure 4. Reviewer guidelines shared across all REVIEWARENA-EVAL prompts. The score fields following this guideline block differ by venue as shown in Table 8.

A.6. REVIEWARENA-EVAL Construction

REVIEWARENA-EVAL is a fixed 1,002-paper subset of REVIEWARENA. We sample exactly 167 papers from each numerically rated venue: NeurIPS, ICLR, ICML, CoRL, COLM, and EMNLP. This equal allocation prevents ICLR and NeurIPS from dominating the aggregate metrics. TMLR is excluded from the benchmark because it has no numeric overall rating, but it remains part of the full dataset and can support future claim-evidence prediction and reviewer-discussion tasks. The benchmark release contains paper identifiers, rendered prompts, predictions, and scoring scripts.

The benchmark uses the following filtering constraints before sampling: the paper must have a successfully downloaded PDF, at least two official reviews, a non-empty OCR/markdown extraction, and a numeric target for the venue's overall rating axis. We keep the benchmark split frozen even if the full dataset is updated, so future systems can compare against the reported numbers.

A.7. Venue-Year-Aware Prompts

The base prompt is rendered with a venue-specific scale description and JSON schema. Every prompt includes the same reviewer guidelines: be specific, cite claims from the paper, assess soundness/clarity/significance/originality when applicable, avoid score inflation, and return valid JSON only. The key difference is the venue's native scoring scale.

Shared reviewer instruction. You are an expert peer reviewer for `{venue}`. Read the OCR'd paper text below and produce a structured review. Provide substantive, evidence-grounded feedback; assess technical soundness, clarity, significance, and originality where applicable; be specific, constructive, fair, and polite; do not inflate scores out of politeness; and return **only valid JSON**.

Paper title: `{title}`
Paper full text: `{paper_text}`

Figure 5. Shared part of the REVIEWARENA-EVAL reviewer prompt. Venue-specific scale and JSON-schema blocks are inserted into this template.

Venue-specific prompt blocks used in REVIEWARENA-EVAL

NeurIPS / ICLR / COLM (1–10): `rating` 1=trivial reject, 3=clear reject, 5=borderline reject, 6=weak accept, 7=accept, 8=top 50% of accepted, 9=top 15%, 10=top 5%. Return JSON with `summary`, `strengths`, `weaknesses`, `soundness`, `presentation`, `contribution`, `rating`, `confidence`, and `rationale`.

ICML 2025 (1–5): `overall_recommendation` 1=strong reject, 2=reject, 3=weak accept, 4=accept, 5=spotlight. Return JSON with `summary`, `strengths`, `weaknesses`, `overall_recommendation`, `confidence`, and `rationale`.

CoRL (1–4): `recommendation` 1=strong reject, 2=weak reject, 3=weak accept, 4=strong accept. Return JSON with `summary`, `strengths`, `weaknesses`, `recommendation`, `technical_quality`, `robotics_focus`, `confidence`, and `rationale`.

EMNLP 2023 (ARR 1–5): `Excitement` 1=poor, 3=ambivalent, 5=transformative; `Soundness` and `Reproducibility` are also 1–5. Return JSON with `summary`, `reasons_to_accept`, `reasons_to_reject`, `soundness`, `excitement`, `reproducibility`, `confidence`, and `rationale`.

Figure 6. Venue-year-aware scale and output-schema blocks. These blocks prevent out-of-range generic 1–10 scoring on ICML, CoRL, and EMNLP.

A.8. Metric Definitions

Let s_i be the mean human overall rating for paper i on its native scale and $\hat{s}_i$ be the model rating. Overall-rating error is

$$\text{MAE} = \frac{1}{n}\sum_i |\hat{s}_i - s_i|, \qquad \text{Bias} = \frac{1}{n}\sum_i (\hat{s}_i - s_i).$$

We compute these quantities per venue rather than first rescaling all venues to 1–10. Sub-score MAE is computed analogously over the available integer sub-score axes for each venue. Decision metrics are computed only on the ICLR slice, where rejected papers are public in sufficient quantity. We map $\hat{s}_i \geq 6$ to predicted accept for ICLR and report accuracy, macro-F1, and the accepted-minus-rejected predicted-rating gap.

For text fields, we compare each model output against the pooled human text for the same paper and field. If a model predicts `strengths`, for example, the reference is the concatenation of the corresponding human-review `strengths` fields, or their venue-specific aliases. We report ROUGE-1/2/L F-measure with stemming, corpus BLEU via SacreBLEU, and BERTScore-F1 with `roberta-large`. Table 9 lists the aliasing rules used for text metrics.

A.9. Full Qualitative Model Examples

The following examples are drawn directly from the released evaluation JSONL files. Each box shows all six model outputs for the same paper, including the parsed numeric scores and the full parsed summary, strengths, weaknesses, and rationale fields. The text is not hand-written; it is copied from the model prediction records, with only LaTeX escaping applied.

Predicted field	Human-review aliases searched
`summary`	`summary`, `summary_of_paper`, `summary_of_the_paper`, EMNLP topic field
`strengths`	`strengths`, `reasons_to_accept`, `strengths_and_weaknesses`
`weaknesses`	`weaknesses`, `reasons_to_reject`, questions/rebuttal fields when needed
`reasons_to_accept`	`reasons_to_accept`, `strengths`
`reasons_to_reject`	`reasons_to_reject`, `weaknesses`

Table 9. Human-review field aliases used when constructing pooled text references.

Example 1: ICLR 2020 rejected paper: POP-Norm / pre-operation normalization

Forum id: `SkevphEYPB`. **Venue/year:** iclr 2020. **Decision:** Reject.

GPT-OSS-120B **rating=3, soundness=2, presentation=2, contribution=2, confidence=3**

Summary. The paper proposes POP-Norm, a pre-operation normalization that removes the mean and divides by a scaled l2-norm before the convolution/inner-product, claiming it reduces gradient Lipschitz constant and variance, thereby accelerating SGD convergence and improving performance especially for sigmoid/tanh activations.

Strengths.

- Provides a unified theoretical framework linking normalization to gradient Lipschitz constant and variance reduction.
- Includes empirical evaluation on CIFAR-10/100 and ImageNet (ILSVRC2012) showing faster convergence and modest accuracy gains.
- Analyzes the interaction between normalization placement and different activation functions, offering an explanation for BatchNorm's weaker performance with sigmoid/tanh.

Weaknesses.

- Theoretical claims are presented with many unproven assumptions; the proofs in the appendix are sketchy and lack rigorous justification.
- Key definitions (e.g., gradient variance, Lipschitz constant in the context of deep nets) are vague, making the arguments hard to verify.
- The novelty is limited: placing normalization before the main operation resembles existing techniques such as weight normalization and does not clearly surpass prior work.
- Presentation is poor: numerous typographical/formatting errors, ambiguous notation, and missing details hinder reproducibility.
- Empirical improvements are modest (3-8% on CIFAR, <1% on ImageNet) and no comparison to recent alternatives like GroupNorm, LayerNorm, or recent variants of BatchNorm.

Rationale. The paper's core idea is not sufficiently novel and the theoretical analysis is not convincingly rigorous; combined with weak presentation and modest empirical gains, it does not meet the acceptance bar.

Kimi K2.6 **rating=5, soundness=2, presentation=2, contribution=2, confidence=4**

Summary. The paper proposes Pre-Operation Normalization (POP-Norm), which applies mean removal and scaled l2-norm division to inputs *before* the linear layer operation rather than after. The authors theoretically argue that these steps reduce gradient variance and the gradient Lipschitz constant, thereby accelerating SGD convergence. They use this framework to interpret BatchNorm's effectiveness with ReLU and its weakness with Sigmoid/Tanh, and report empirical comparisons on CIFAR and ImageNet.

Strengths.

- The paper offers an alternative theoretical perspective on why normalization accelerates training, linking it to reduction of gradient variance and Lipschitz constants rather than internal covariate shift.
- The empirical demonstration that POP-Norm significantly outperforms BatchNorm for Sigmoid and Tanh activations (Table 1) is interesting and suggests that the placement of normalization matters for non-ReLU networks.
- The theoretical analysis in Theorems 2 and 3, while based on upper bounds, provides a formal justification for how input normalization affects optimization properties at the layer level.

Weaknesses.

- The end-to-end theoretical justification for convergence acceleration is missing: Theorems 2 and 3 analyze a single layer in isolation, and the paper does not show how per-layer normalization propagates to guarantee faster global convergence of the deep network.
- The ablation study in Section 5.3 directly contradicts the claim in Section 3 and Remark 1 that mean removal and division by l2-norm 'can individually take effects': for Sigmoid activation, neither component alone allows training to converge, which the authors acknowledge but fail to reconcile with their theory.
- The proofs rely on bounding the upper bound of gradient variance and the minimal Lipschitz constant, but reducing these bounds does not rigorously establish that the actual global quantities are reduced in a way that guarantees acceleration across all layers.
- The practical impact is limited because the largest improvements are obtained with Sigmoid and Tanh, which are rarely used in modern architectures; the gains with ReLU (the standard choice) are marginal (~0.7% on ImageNet).

- The paper is difficult to follow due to inconsistent notation, numerous typos, and poor mathematical formatting (exacerbated by OCR), which hinders reproducibility and verification of the proofs.

Rationale. While the paper presents an intriguing alternative view of normalization via gradient variance and Lipschitz constant reduction, its theoretical claims are not fully substantiated for deep networks, and the ablation study contradicts key assertions about the independence of mean removal and l2 division. Combined with marginal gains on standard ReLU architectures and poor clarity, the paper does not yet meet the ICLR bar.

Qwen3.6-Plus **rating=3, soundness=2, presentation=2, contribution=2, confidence=4**

Summary. The paper introduces POP-Norm, a normalization method applied before a layer's linear operation, claiming it theoretically reduces gradient variance and Lipschitz constants to accelerate SGD convergence. It contrasts this with BatchNorm, attributing BatchNorm's success with ReLU and failure with Sigmoid/Tanh to the placement of normalization relative to nonlinearities. Experiments on CIFAR and ImageNet show consistent but modest improvements, particularly for non-ReLU activations.

Strengths.

- Provides an optimization-theoretic perspective on normalization, explicitly linking convergence speed to gradient variance and Lipschitz constants (Section 2).
- Systematically evaluates the impact of the scaling hyperparameter κ and different activation functions, offering useful ablation insights (Section 5.2, Figure 2).
- Demonstrates consistent empirical improvements over standard BatchNorm across VGG and ResNet architectures on multiple benchmarks (Tables 1 & 2).

Weaknesses.

- The theoretical justification (Theorems 2 & 3) is heuristic; the proofs bound layer-wise input statistics but do not rigorously establish how this reduces the variance or Lipschitz constant of backpropagated gradients across the full computational graph.
- The core architectural choice (normalization before the linear operation) is not novel and closely mirrors Pre-Activation ResNets (He et al., 2016) and standard input normalization practices.
- The heavy focus on Sigmoid and Tanh activations limits practical relevance, as modern deep networks predominantly use ReLU, GELU, or Swish. The marginal gains for ReLU (~1-2%) do not strongly justify the proposed modifications.
- The role of the scaling factor κ (Remark 2) is explained via speculative claims about optimization trajectories and local minima, lacking rigorous proof or a controlled comparison against simple learning rate or initialization scaling.
- Mathematical notation and algorithmic descriptions contain ambiguities and grammatical issues that hinder reproducibility and clarity of the proposed normalization procedure.

Rationale. The paper's theoretical claims are not rigorously proven, and the proposed method lacks novelty relative to established pre-activation architectures. While the empirical results are consistent, they are marginal and focus on practically obsolete activation functions. I would reconsider if the authors provided rigorous proofs linking pre-operation normalization to full-network gradient properties and demonstrated significant gains on modern architectures and activation functions. No ethics concerns identified.

GLM-5.1 **rating=3, soundness=1, presentation=2, contribution=1, confidence=4**

Summary. The paper proposes POP-Norm (Pre-Operation Normalization), a normalization technique that applies mean removal and L2-norm division to layer inputs before the linear operation, contrasting with BatchNorm which is applied after. The authors provide theoretical analysis claiming this reduces gradient variance and Lipschitz constant, thereby accelerating convergence. They also attempt to explain BatchNorm's effectiveness with ReLU versus Sigmoid/Tanh, and show empirical improvements on CIFAR and ImageNet datasets.

Strengths.

- Attempts to ground normalization methods in optimization theory by analyzing gradient variance and Lipschitz constants.
- Provides a novel perspective on the interaction between normalization and different activation functions (ReLU vs. Sigmoid/Tanh).
- Extensive empirical evaluation across multiple datasets (CIFAR-10/100, ImageNet) and architectures (VGG, ResNet).

Weaknesses.

- The theoretical proofs (Theorems 2 and 3) primarily demonstrate that scaling down the inputs scales down the upper bounds of gradient variance and Lipschitz constants. This is a trivial property of scaling and does not necessarily translate to faster convergence without adjusting the learning rate, which the authors only address heuristically in Remark 2.
- The empirical comparisons use identical learning rates for BatchNorm and POP-Norm. Since the methods scale activations differently, the optimal learning rates likely differ, making the observed convergence speed improvements potentially confounded by the learning rate choice.
- The explanation for BatchNorm's interaction with ReLU vs. Sigmoid/Tanh (Section 4) contains mathematical errors and logical inconsistencies. Specifically, Eq. 16 makes a false claim about the sum of ReLU activations of a symmetric distribution (the sum of ReLU of a symmetric distribution is strictly positive, not zero), and the analysis seems to misrepresent the order of operations in standard BatchNorm.
- Improving Sigmoid/Tanh networks is of limited practical significance given the dominance of ReLU and its variants in modern architectures, and the improvements on ReLU networks are marginal.

Rationale. The paper contains critical mathematical errors in its theoretical justification (e.g., Eq. 16) and relies on trivial scaling properties for its main proofs. The empirical evaluation is confounded by not tuning the learning rate for the new normalization scheme, and the gains on ReLU networks are marginal. Therefore, the paper does not meet the bar for acceptance.

DeepSeek-V4-Pro **rating=5, soundness=2, presentation=2, contribution=2, confidence=3**

Summary. The paper proposes POP-Norm, a normalization method that removes the mean and divides by a scaled l2 norm before the main layer operation, claiming theoretical justification via reduction of gradient variance and Lipschitz constant. It argues that BatchNorm's effectiveness can be similarly explained and that POP-Norm yields faster convergence and better performance, especially with Sigmoid/Tanh activations. Experiments on CIFAR and ImageNet show improvements over BatchNorm.

Strengths.

- The paper attempts to provide a theoretical grounding for normalization by connecting to gradient variance and Lipschitz constant reduction.
- POP-Norm consistently outperforms BatchNorm in the reported experiments, particularly for Sigmoid and Tanh activations.
- The analysis of why BatchNorm underperforms with certain activations is plausible and aligns with empirical observations.

Weaknesses.

- The theoretical justification is based on loose upper bounds and does not rigorously prove faster convergence; the claim of being 'theoretically justified' is overstated.
- The writing quality is poor, with numerous grammatical errors, typos, and unclear notation, significantly hindering readability.
- Empirical evaluation is limited to relatively shallow networks (VGG16, ResNet20/18) and shows only modest gains on ImageNet; comparisons with other normalization methods (e.g., GroupNorm, WeightNorm) are missing.
- The method introduces extra hyperparameters (κ, pbl) without thorough sensitivity analysis or principled tuning guidelines.
- The explanation of BatchNorm's effectiveness is qualitative and not rigorously derived from the stated theory.

Rationale. The paper presents an incremental idea with weak theoretical support and poor presentation. The empirical gains are modest and the evaluation is not comprehensive. The contribution does not meet the bar for acceptance at a top venue.

Gemma-4-26B **rating=7, soundness=3, presentation=3, contribution=3, confidence=4**

Summary. The paper proposes Pre-Operation Normalization (POP-Norm), a method that applies mean removal and l_2 norm scaling to inputs before the main linear or convolutional operation, rather than after, as in standard BatchNorm. The authors provide a theoretical framework arguing that reducing the gradient Lipschitz constant and gradient variance is key to accelerating SGD convergence. Extensive experiments on CIFAR and ImageNet demonstrate that POP-Norm outperforms BatchNorm, particularly when using Sigmoid and Tanh activation functions.

Strengths.

- Provides a novel theoretical perspective on normalization by focusing on the optimization landscape (gradient Lipschitz constant and variance) rather than the widely debated Internal Covariate Shift.
- The 'Pre-Operation' placement is a distinct and well-motivated architectural departure from existing normalization techniques.
- The paper successfully uses non-ReLU activations (Sigmoid and Tanh) to validate its theory; the significant performance gains in these cases provide strong evidence that the proposed mechanism works as intended.
- The ablation studies and parameter sensitivity analysis (regarding the scaling factor κ) are thorough and help clarify the contribution of each component.

Weaknesses.

- While the improvements for Sigmoid and Tanh are impressive, the performance gain over BatchNorm when using the industry-standard ReLU is relatively modest (e.g., ~1-2% in Table 1).
- The introduction of the scaling hyperparameter κ adds complexity to the model training process, as it requires tuning to avoid either vanishing gradients or losing representation power.
- The theoretical proofs rely on the concept of a 'minimal gradient Lipschitz constant,' which is a less standard optimization assumption than the global Lipschitz continuity typically used in convergence proofs.
- The paper lacks a direct comparison with other modern normalization methods like Layer Normalization or Group Normalization in the main experimental results, despite mentioning them in the introduction.

Rationale. The paper is a strong submission that offers both a new method and a meaningful theoretical interpretation of why certain normalization/activation combinations work better than others. The empirical results on non-ReLU activations are particularly convincing and support the theoretical claims.

Analysis. This rejected-paper example shows the central calibration failure. The models often cite the same weaknesses, but Gemma maps them to an accept-level rating while Qwen, GLM, and GPT-OSS map them to reject-level ratings.

Example 2: ICLR 2024 accepted spotlight paper: Safe RLHF

Forum id: `TyFrPOKYXw`. **Venue/year:** iclr 2024. **Decision:** Accept (spotlight).

GPT-OSS-120B **rating=6, soundness=3, presentation=2, contribution=3, confidence=3**

Summary. The paper introduces Safe RLHF, a framework that decouples helpfulness and harmlessness during data annotation and training, using separate reward and cost models and a Lagrangian method to dynamically balance the two objectives in reinforcement learning from human feedback. Experiments fine-tuning Alpaca-7B over three iterations show improved helpfulness and reduced harmful responses according to GPT-4 and human evaluations.

Strengths.
- Clear motivation addressing the tension between helpfulness and safety in LLM alignment.
- Novel integration of constrained RL (Lagrangian) with RLHF to balance reward and cost dynamically.
- Introduces a Cost Model that combines preference ranking with binary safety classification, achieving high safety classification accuracy.
- Empirical results demonstrate simultaneous gains in helpfulness and harmlessness across multiple evaluation methods.
- Release of code, data, and detailed reproducibility instructions.

Weaknesses.
- Limited theoretical analysis of convergence or optimality of the Lagrangian approach in this setting.
- Experiments are confined to a single 7B model; broader benchmarks and larger models are not evaluated.
- Comparisons are mainly against a static reward-shaping baseline; stronger baselines (e.g., Constitutional AI, safety-prompted RLHF) are absent.
- Details of human evaluation (sample size, inter-rater reliability) are insufficiently described.
- Some methodological details (hyper-parameters, training stability) are relegated to appendices and not visible in the main text.

Rationale. The proposed method is plausible and shows empirical improvements, but the lack of strong theoretical guarantees and limited experimental scope keep it from a higher acceptance level.

Kimi K2.6 **rating=7, soundness=3, presentation=3, contribution=3, confidence=4**

Summary. This paper proposes Safe RLHF, a framework that explicitly decouples human preferences for helpfulness and harmlessness during data annotation and policy optimization for LLM alignment. It casts the problem as a constrained MDP, introducing a novel Cost Model alongside a standard Reward Model, and employs the Lagrangian method to dynamically balance the two objectives during RL fine-tuning. Empirically, three iterative rounds of Safe RLHF on Alpaca-7B yield substantial improvements in both helpfulness and harmlessness, validated by human evaluations and GPT-4 Elo ratings.

Strengths.
- Tackles a critical and under-addressed tension in LLM alignment: the inherent conflict between maximizing helpfulness and ensuring harmlessness, which standard RLHF conflates into a single objective.
- Proposes a principled constrained optimization formulation (CMDP) for safety alignment, paired with a novel Cost Model that fuses pairwise preference learning with binary safety classification to enable dynamic thresholding.
- Demonstrates through controlled ablations that decoupling preference annotations improves inter-rater agreement and that dynamic Lagrangian balancing outperforms static Reward Shaping across a range of weights.
- Provides extensive empirical validation across three iterative training rounds with red-teaming, model-based evaluation, human evaluation, and GPT-4 evaluation, showing consistent gains in both objectives.
- Commits to strong reproducibility by releasing all training code, preference data, evaluation data, and detailed annotation guidelines.

Weaknesses.
- The claim that this is the 'first integration of Safe RL and the RLHF framework' is strong and insufficiently qualified; prior and concurrent work on multi-objective or constrained RL for language generation is not adequately surveyed or differentiated.
- Theoretical grounding is limited: the paper offers no convergence analysis, regret bounds, or formal characterization of the Lagrangian dynamics in the high-dimensional, non-convex LLM fine-tuning setting.
- Baseline comparisons are narrow, focusing on standard PPO and Reward Shaping; the paper omits comparisons to other safety-centric alignment methods such as Constitutional AI or adversarial-prompt RLHF pipelines used in contemporary systems.
- The three-round iterative protocol confounds algorithmic improvements with increased data quantity and red-teaming coverage; the paper does not isolate the marginal contribution of the Safe RLHF algorithm holding data volume and prompt distribution fixed.
- The Cost Model's reliance on a 'virtual response' with zero cost as a decision boundary (Section 4.2), while intuitive, lacks formal justification for why this threshold optimally separates safe and unsafe responses across diverse harm categories.

Rationale. The paper presents a well-motivated, empirically thorough approach to a pressing problem in LLM safety. The decoupled annotation strategy, Cost Model design, and dynamic Lagrangian balancing are clear strengths. While the work would benefit from deeper theoretical analysis and broader baseline comparisons, the core contribution is solid and the results are convincing. This paper meets the bar for acceptance.

Qwen3.6-Plus **rating=7, soundness=3, presentation=3, contribution=3, confidence=4**

Summary. The paper proposes Safe RLHF, a framework that decouples human preference annotation into independent helpfulness and harmlessness dimensions, training separate Reward and Cost models. It formulates LLM alignment as a constrained optimization problem, solved via Lagrangian relaxation to dynamically balance the two objectives during policy optimization. Experiments across three iterative fine-tuning rounds on Alpaca-7B demonstrate consistent improvements in both helpfulness and safety compared to standard RLHF and static reward shaping baselines.

Strengths.

- The decoupling of helpfulness and harmlessness during annotation directly addresses a known source of noise and cognitive conflict in standard RLHF pipelines, empirically improving inter-rater agreement and yielding cleaner optimization signals.
- Formulating alignment as a Constrained MDP with a dynamically updated Lagrangian multiplier (Eq. 10) provides a principled, automated alternative to manual reward shaping, effectively preventing over-optimization of safety at the expense of helpfulness.
- Comprehensive ablation studies (Sections 5.2.2-5.2.4) rigorously validate core design choices, including the Cost Model's integrated classification term and the superiority of dynamic constraint handling over static weighting.
- Transparent ethical practices, including fair annotator compensation, mental health support, and explicit acknowledgment of dual-use risks associated with releasing red-teaming data, demonstrate responsible research conduct.

Weaknesses.

- The safety constraint is relaxed to an expectation over prompts (Eq. 8), which guarantees average safety but does not provide worst-case or per-prompt safety guarantees, a notable limitation for high-stakes deployment scenarios.
- The claim of being the 'first integration of Safe RL and RLHF' overlooks prior work on constrained preference optimization and multi-objective alignment; the novelty is primarily in the practical annotation pipeline and Lagrangian implementation rather than theoretical breakthroughs.
- Experiments rely on an outdated base model (Alpaca-7B/LLaMA-1), limiting the generalizability of findings to modern, larger-scale architectures where scaling laws and different pretraining regimes may alter the helpfulness-safety trade-off.
- Mathematical notation in Section 4.3 contains ambiguities (e.g., undefined hyperparameter d in Eq. 8, abrupt transition from Eq. 7 to 8) that require careful appendix reading to fully reproduce the surrogate objective formulation.

Rationale. The paper presents a well-motivated and empirically validated framework that addresses a practical bottleneck in LLM alignment. While the core constrained optimization approach builds on established Safe RL theory, the decoupled annotation pipeline and dynamic Lagrangian balancing offer meaningful methodological advances. The work is technically sound, clearly presented with minor notation issues, and responsibly addresses ethical considerations, warranting acceptance.

DeepSeek-V4-Pro **rating=6, soundness=3, presentation=3, contribution=3, confidence=3**

Summary. The paper proposes Safe RLHF, a framework that decouples helpfulness and harmlessness in RLHF by training separate reward and cost models and using constrained RL with Lagrangian multipliers to dynamically balance the two objectives. Through three iterative rounds of fine-tuning on Alpaca-7B, the method improves both helpfulness and harmlessness over baselines like reward shaping and conventional RLHF. The cost model incorporates safety classification to provide a threshold for constraints.

Strengths.

- Novel integration of Safe RL (constrained MDP) with RLHF to explicitly handle the tension between helpfulness and harmlessness.
- Decoupling of human annotations for helpfulness and harmlessness improves inter-rater agreement and enables separate preference models.
- Cost model design that combines Bradley-Terry preference modeling with safety classification, providing a natural decision boundary for constraints.
- Lagrangian method dynamically adjusts the trade-off, avoiding over-optimization of either objective, as demonstrated by training curves.
- Extensive iterative experiments with red-teaming, GPT-4 and human evaluations, and release of code and data for reproducibility.

Weaknesses.

- The constraint threshold d is introduced but not clearly defined or ablated; its selection and impact on the safety guarantee are not discussed.
- Cost model classification accuracy drops from 95.6% to ~85% in later rounds without explanation, raising concerns about stability and generalizability.
- Evaluation relies heavily on GPT-4 as a judge; human evaluation details (number of raters, prompts, agreement) are sparse, limiting confidence in the human Elo scores.
- Experiments are limited to a single base model (Alpaca-7B) and a specific dataset; generalizability to larger models or different architectures is not shown.
- Comparison to prior multi-objective RLHF methods is limited to reward shaping; other adaptive balancing techniques (e.g., Pareto optimization) are not considered.
- The claim of being the first integration of Safe RL and RLHF overlooks prior work that uses safety classifiers or fixed-weight harmlessness rewards in constrained-like settings.

Rationale. The paper presents a well-motivated approach to balancing safety and helpfulness in LLMs, with solid experimental

validation and a clear methodological contribution. However, missing ablations (e.g., constraint threshold d), unexplained degradation in cost model accuracy, and limited evaluation robustness prevent a higher rating. The work is a meaningful step forward but remains incremental over existing multi-objective RLHF methods.

Gemma-4-26B **rating=7, soundness=4, presentation=3, contribution=4, confidence=4**

Summary. The paper proposes Safe RLHF, a framework designed to resolve the inherent tension between the helpfulness and harmlessness of Large Language Models (LLMs). Unlike standard RLHF which often collapses these objectives into a single reward, Safe RLHF decouples the annotation process and trains separate reward and cost models. It then formulates the alignment task as a Constrained Markov Decision Process (CMDP), using the Lagrangian method to dynamically balance the maximization of helpfulness with a safety constraint on harmfulness.

Strengths.

- Principled approach to safety: Instead of using heuristic reward shaping (e.g., $R_{total} = R_{help} + β R_{harm}$), the paper uses the Lagrangian method to treat safety as a formal constraint, which is a more mathematically grounded way to handle conflicting objectives.
- Decoupled annotation strategy: The authors correctly identify that asking crowdworkers to rank a single 'overall' preference can lead to confusion when helpfulness and harmlessness conflict. Separating these dimensions improves inter-rater agreement (from 61.65% to ~67-69% as noted in Section 5.2.2).
- Novel Cost Model design: The introduction of a Cost Model that incorporates a classification term (Equation 4) to capture a safety threshold is a clever way to bridge the gap between preference modeling and safety classification.
- Strong empirical evidence: The iterative three-round fine-tuning shows significant improvements in both helpfulness and harmlessness Elo scores (Figure 5) and a dramatic reduction in harmful response rates (from 53.08% to 2.45% in Figure 5c).
- Comprehensive ablation studies: The paper provides meaningful comparisons against static Reward Shaping and simple safety classifiers, demonstrating that the dynamic Lagrangian adjustment and the specific design of the Cost Model are critical to success.

Weaknesses.

- Average vs. Worst-case safety: The safety constraint is formulated as an expectation over a distribution (Equation 8). In safety-critical AI applications, users are often more concerned with 'tail risks' or worst-case scenarios rather than the average cost. The framework does not explicitly address minimizing the maximum possible harm.
- High computational and human cost: The iterative nature of the pipeline (Red-teaming $\rightarrow$ Generation $\rightarrow$ Annotation $\rightarrow$ Model Training) is extremely resource-intensive. While the authors acknowledge this in the limitations, the scalability of this approach for very large models or more complex, multi-turn safety constraints is not fully explored.
- Evaluation bias: The authors note that they constructed their own evaluation prompt dataset because existing benchmarks are insufficient. While this is a common problem in the field, it introduces a potential risk of evaluation bias if the evaluation set is not sufficiently diverse or is inadvertently leaked during the red-teaming/training phases.
- Complexity of implementation: The dual-model training and the min-max optimization of the Lagrangian multiplier add significant complexity to the RLHF pipeline compared to standard PPO or DPO, which may hinder widespread adoption without more streamlined implementations.

Rationale. The paper presents a highly relevant and technically sound solution to a major problem in LLM alignment. The transition from reward shaping to constrained optimization via CMDP is a significant conceptual advancement. The empirical results are strong and the ablation studies are well-executed, though the high cost of the iterative process and the focus on average rather than worst-case safety are notable limitations.

Analysis. This accepted-paper example shows broad agreement that the contribution is strong, but the models differ in how much they reward it numerically and how they phrase residual safety/generalization concerns.

Example 3: COLM 2024 accepted paper: WorkBench

Forum id: `4HNAwZFDcH`. **Venue/year:** colm 2024. **Decision:** Accept.

GPT-OSS-120B **rating=6, soundness=2, presentation=3, contribution=3, confidence=3**

Summary. The paper introduces WorkBench, a synthetic benchmark consisting of 690 multi-domain workplace tasks and a sandbox environment with five databases and 26 tools. It proposes outcome-centric evaluation, where a task is deemed successful if the final state of the sandbox matches a unique ground-truth outcome, and reports results of ReAct agents (GPT-4 achieving 43% accuracy).

Strengths.

- Provides a novel, multi-domain benchmark focused on realistic workplace actions rather than pure retrieval or single-step tasks.
- Outcome-centric evaluation offers a robust, automatic way to assess agents without relying on tool-call matching, addressing a limitation of prior benchmarks.

- Comprehensive dataset creation pipeline (human-curated templates + programmatic expansion) yields diverse tasks with varying numbers of required actions.
- Empirical analysis identifies concrete failure modes of current agents (e.g., tool selection, plan execution, side-effects).

Weaknesses.

- Evaluation is limited to a single agent framework (ReAct) and a small set of LLMs; no human baseline or comparison to alternative agent designs.
- The sandbox databases are highly simplified (e.g., 300 calendar events, 500 emails), raising questions about ecological validity for real workplace systems.
- Outcome-centric metric, while well-motivated, lacks a formal definition of side-effects and does not quantify the severity of incorrect actions.
- Insufficient statistical analysis (e.g., confidence intervals, significance testing) and limited ablation studies on the impact of tool count or prompt design.
- Some methodological details (full tool docstrings, prompt template, generation of ground-truth outcomes) are omitted or truncated, hindering reproducibility.

Rationale. The work offers a useful new benchmark and a thoughtful evaluation approach, but methodological gaps and limited experimental scope prevent a stronger endorsement.

Kimi K2.6 **rating=7, soundness=3, presentation=3, contribution=3, confidence=4**

Summary. This paper introduces WorkBench, a benchmark of 690 multi-step workplace tasks spanning five sandbox domains (email, calendar, CRM, analytics, project management) with 26 tools. Its central contribution is outcome-centric evaluation, which judges agents by comparing the final database state to a unique ground-truth outcome rather than matching specific action trajectories. The authors evaluate five ReAct agents and find that even GPT-4 achieves only 43% accuracy, with frequent side effects such as sending emails to the wrong recipient.

Strengths.

- The outcome-centric evaluation methodology is a clear conceptual advance: it enables robust, automatic evaluation without an LLM-as-judge and avoids penalizing agents for valid alternative action paths (Figure 5).
- The benchmark fills a practical gap by focusing on realistic, multi-step workplace tasks that require planning and cross-domain tool use, with challenging tasks requiring up to 12 actions.
- The empirical analysis is compelling: the low accuracy of even GPT-4 (43%) and the detailed error breakdown (Figure 6) convincingly demonstrate that current agents are not ready for high-stakes workplace deployment.
- The dataset is publicly available and constructed through a scalable template-based approach (69 templates yielding 690 tasks with linguistic variation), facilitating future extension.

Weaknesses.

- The experimental evaluation is limited to a single agent architecture (ReAct); including other frameworks (e.g., plan-then-execute, reflexion) would strengthen claims about benchmark difficulty versus inherent agent limitations.
- The sandbox environment is small and sanitized compared to real workplace systems (e.g., 500 emails, no spam, no GUI), which may overestimate performance and limits ecological validity; the authors acknowledge this but do not quantify its impact.
- The strong 'first' claims regarding outcome-centric evaluation (Table 1, Section 3.4) are somewhat overstated given that state-based and text-based final-outcome evaluation exists in prior and concurrent work (e.g., WebArena, ALFWorld); the novelty lies in the specific application to workplace database state changes, not in the concept itself.
- No human baseline is provided, making it difficult to calibrate whether the low accuracy stems from inherent task ambiguity or genuine agent shortcomings.
- Several important experimental details are missing or underspecified, including the full ReAct prompt template, the number of in-context examples provided, and whether sandbox databases are reset between tasks.

Rationale. WorkBench offers a solid benchmark contribution with a useful methodological twist in outcome-centric evaluation, and the empirical findings are genuinely revealing. While the experimental depth is limited (single architecture, no human baseline) and the sandbox is simplified, the paper provides a valuable, publicly available resource for evaluating workplace agents and identifies concrete failure modes that motivate future research.

Qwen3.6-Plus **rating=7, soundness=3, presentation=3, contribution=3, confidence=4**

Summary. WorkBench introduces a benchmark dataset for evaluating LLM-based agents on realistic workplace tasks, featuring a sandbox environment with five databases and 26 tools. The authors propose outcome-centric evaluation, which assesses agents based on the final state of the environment rather than exact tool-call trajectories. Experiments with five ReAct agents reveal significant performance gaps (GPT-4 achieves 43% accuracy) and highlight common failure modes, including harmful side effects like misdirected emails.

Strengths.

- Introduces outcome-centric evaluation, a robust alternative to trajectory-matching or LLM-as-a-judge, allowing for multiple valid action paths to the same ground truth state (Sec 3.4, Fig 5).
- Well-designed sandbox environment with programmatic task generation and linguistic variation, ensuring scalability, reproducibility, and unambiguous ground truth computation (Sec 3.1-3.2, Fig 3).

- Comprehensive error analysis that categorizes failures into side effects, framework violations, and tool misuse, providing actionable insights for agent development and safety (Sec 4.4, Fig 6).

Weaknesses.

- The sandbox environment uses relatively small, clean datasets (e.g., 300 calendar events, 500 emails), which may overestimate agent capabilities compared to noisy, large-scale real-world systems (Sec 5).
- Lacks multi-turn interaction evaluation, limiting ecological validity for workplace scenarios that typically involve iterative clarification, feedback, and long-horizon planning (Sec 5).
- Ethics/Societal concern: The benchmark explicitly demonstrates that agent errors can cause real-world harm (e.g., sending emails to wrong recipients), but the paper does not propose mitigation strategies, safety guardrails, or deployment guidelines, which is critical given the high-stakes workplace context.

Rationale. The paper presents a well-constructed benchmark with a clear methodological contribution (outcome-centric evaluation) that addresses a key gap in agent assessment. While the simulated environment and single-turn design limit real-world generalizability, the thorough error analysis and open-source release provide substantial value to the community. The empirical results are consistent with claims, and limitations are transparently discussed.

GLM-5.1 **rating=7, soundness=3, presentation=3, contribution=3, confidence=4**

Summary. The paper introduces WorkBench, a benchmark for evaluating LLM-based agents in a realistic workplace setting using a sandbox environment with 5 databases and 26 tools across 690 tasks. Its primary conceptual contribution is 'outcome-centric evaluation,' which assesses agents based on whether the final state of the environment matches the ground truth rather than requiring an exact match of tool calls. Evaluating ReAct agents, the authors find that even GPT-4 only achieves 43% accuracy and often produces negative side effects, highlighting significant risks in high-stakes deployments.

Strengths.

- The 'outcome-centric evaluation' methodology is a strong conceptual contribution that addresses a key limitation of prior benchmarks (e.g., API-Bank) which rely on exact tool-call matching. By evaluating the final state, it fairly credits agents that recover from errors or take alternative valid paths (as illustrated in Figure 5).
- The focus on side effects (unintended state changes) is a highly relevant and important contribution for safety and robustness, demonstrating that agents frequently take dangerous incorrect actions (e.g., emailing the wrong person) rather than simply failing gracefully.
- The benchmark tasks are practical and representative of common business activities, making the evaluation highly relevant to real-world agent deployment scenarios.
- The paper is well-structured and clearly written, with strong ablations and error analyses (e.g., Figure 6, Table 7) that break down failure modes effectively.

Weaknesses.

- The sandbox environment is relatively small (e.g., 500 emails, 300 events), which may not reflect the complexity of real-world systems where information retrieval is much harder due to scale. The authors acknowledge this, but it limits the benchmark's ability to test robust information gathering.
- The task generation relies heavily on templates (690 tasks from 69 templates). While the authors argue for diversity via linguistic variation, the underlying reasoning and tool-use patterns within a template are highly similar, which might lead to overfitting or an overestimation of generalization.
- The evaluation is restricted to the ReAct framework. While popular, evaluating other agent architectures (e.g., reflexion, plan-and-solve) could provide a more comprehensive understanding of the benchmark's utility and the models' capabilities.
- The inclusion of 'no action' tasks (18% of the dataset) significantly inflates the overall accuracy scores, especially for weaker models (e.g., Mixtral achieves 77% on 0-action tasks but nearly 0% on 1+ action tasks). The aggregate metric obscures the agents' actual ability to execute actions.

Rationale. The paper makes a clear and defensible methodological contribution with outcome-centric evaluation, addressing a significant flaw in prior action-based benchmarks. The empirical analysis provides valuable insights into agent failures and safety risks via the side effects metric. While the benchmark's scale and template-based generation are limitations, they do not undermine the core contribution, making this a solid accept for COLM.

DeepSeek-V4-Pro **rating=6, soundness=3, presentation=3, contribution=3, confidence=3**

Summary. This paper introduces WorkBench, a benchmark dataset for evaluating LLM agents on realistic workplace tasks. It features a sandbox environment with five databases, 26 tools, and 690 tasks that require planning and multi-step actions. The key contribution is outcome-centric evaluation, where tasks have unique, unambiguous ground-truth outcomes, enabling robust automated evaluation. Experiments with five ReAct agents show GPT-4 achieves only 43% accuracy, and error analysis reveals frequent side effects like sending emails to wrong recipients.

Strengths.

- Outcome-centric evaluation is a novel and practical method for robust, automated assessment of action-taking agents without relying on LLM evaluators.
- The dataset covers multiple realistic workplace domains (email, calendar, CRM, etc.) and includes complex multi-step tasks requiring tool selection and planning.
- Thorough error analysis, including quantification of side effects, highlights safety concerns for deploying agents in high-stakes settings.

- The benchmark is challenging for current state-of-the-art agents (GPT-4 only 43%), revealing significant room for improvement.
- Publicly available code and dataset promote reproducibility and future research.

Weaknesses.

- The sandbox environment is simplified (small databases, no spam, no authentication) and may overestimate real-world agent capabilities; a human baseline is missing to calibrate task difficulty.
- Tasks are single-turn, whereas real workplace interactions often involve multi-turn dialogues, limiting ecological validity.
- Only ReAct agents are evaluated; other agent frameworks are not tested, which may limit the generalizability of findings.
- The number of tools (26) is modest, and the paper does not deeply explore how performance scales with tool count, despite noting a drop when all tools are provided.
- Some details are lacking, such as how the 690 tasks were partitioned (if at all) for development vs. test, and the paper does not discuss potential biases in task template design.

Rationale. The paper makes a solid contribution with a novel evaluation methodology and a challenging benchmark, but the lack of a human baseline and the simplified environment weaken the strength of the conclusions. It is a useful resource for the community, but not a breakthrough.

Gemma-4-26B **rating=8, soundness=4, presentation=4, contribution=4, confidence=5**

Summary. The paper introduces WorkBench, a benchmark designed to evaluate LLM agents in a simulated, multi-domain workplace environment. It features 690 tasks across five domains (Calendar, Email, CRM, Analytics, and Project Management) using 26 tools. A primary contribution is the 'outcome-centric evaluation' methodology, which assesses success based on the final state of the sandbox databases rather than the specific sequence of tool calls, allowing for multiple valid execution paths.

Strengths.

- Introduces a novel 'outcome-centric' evaluation methodology that overcomes the brittleness of matching exact function calls, as demonstrated in Figure 5.
- The focus on 'side effects' (unintended state changes) is a highly relevant and necessary metric for assessing the safety and reliability of autonomous agents in high-stakes environments.
- The benchmark covers realistic, multi-domain tasks that require complex planning and tool selection (e.g., combining Analytics and Email), moving beyond simple retrieval or single-tool tasks.
- Provides a detailed and insightful error analysis (Figure 6), identifying specific failure modes such as ReAct framework violations, tool limit misunderstandings, and hallucinated identifiers.
- The task generation process using templates and programmatic scaling is well-documented and appears scalable for future extensions.

Weaknesses.

- The sandbox environment is relatively small (e.g., 500 emails, 300 events), which may not fully capture the complexity, noise, or scale of real-world workplace data.
- The tasks are limited to single-turn interactions; real-world workplace agentic workflows are typically multi-turn and involve iterative clarification with users.
- The empirical evaluation is somewhat limited in scope, primarily focusing on GPT-4 for the deep-dive error analysis due to the poor performance of other models.

Rationale. This is a high-quality paper that provides both a significant new resource (the dataset) and a meaningful methodological advancement (outcome-centric evaluation). The focus on side effects and multi-domain reasoning makes it highly relevant to the current direction of agent research.

Analysis. This COLM example shows that long, detailed review text can still coexist with compressed or inflated numeric scores, motivating the multi-axis evaluation in the main paper.

A.10. Example Evaluation Record

Figure 7 shows the structure of a saved prediction row. The scoring code joins this row to the gold REVIEWARENA-EVAL row through `forum_id`, parses `reviews_json`, computes mean human ratings/sub-scores, and emits per-model metrics.

A.11. Compute and Cost

All inference was performed through the Fireworks AI serverless inference API in November–December 2025 at temperature 0.0 with `max_tokens`=8192 and a per-request client timeout of 600 s. Concurrency was managed through a thread pool with backoff on rate-limit errors. PDFs were converted to text once with Nemotron-Nano-OCR running on Modal H100 workers; the OCR sweep over all 51,529 papers in REVIEWARENA consumed approximately 70 H100-hours and is amortised across all benchmark runs.

```
{
"forum_id": "abc123", "conference": "iclr", "year": 2025,
"model": "qwen3p6-plus", "ok": true,
"prediction": {
"summary": "...", "strengths": ["..."], "weaknesses": ["..."],
"soundness": 3, "presentation": 3, "contribution": 2,
"rating": 6, "confidence": 4, "rationale": "..."
},
"usage": {"prompt_tokens": 14820, "completion_tokens": 431}
}
```


Figure 7. Schematic structure of an evaluation prediction record.

Model	Input tokens	Output tokens	Approx. cost
GPT-OSS-120B	14.84M	0.95M	$2.80
Kimi K2.6	14.60M	4.79M	$20.73
Qwen3.6-Plus	15.71M	5.13M	$12.44
GLM-5.1	14.58M	2.20M	$11.25
DeepSeek-V4-Pro	14.73M	2.17M	$6.36
Gemma-4-26B-A4B-IT	14.17M	1.98M	$4.42
Total	**88.63M**	**17.22M**	**$58.00**

Table 10. Approximate inference cost for the six-model REVIEWARENA-EVAL run.

The serverless rate limits on new Fireworks accounts (approximately 1 request/s, 1000 input tokens/s, and 200 output tokens/s) make fully sequential inference slow at our 14,000–16,000-input-token payload size. The released harness uses bounded concurrency, request-level timeouts, JSON parsing checks, and retry/backoff for transient rate-limit failures.

A.12. Reproducibility Checklist for Users

To reproduce the benchmark numbers, users should: (1) load the frozen REVIEWARENA-EVAL split; (2) use the released OCR text and prompt templates; (3) query the same model identifiers at temperature 0; (4) parse only valid JSON predictions; (5) compute numeric scores on each venue's native scale; (6) use `decision_normalized` for ICLR decision metrics; and (7) use the field aliases in Table 9 for text metrics. Deviating from any of these steps changes the task, especially if a generic 1–10 prompt is substituted for venue-year-aware prompting.