FIRSTPASS: A Multi-Domain, Multi-Round Peer Review Dataset Grounded in Real Editorial Outcomes

Prabhjot Singh [1 2] **Somnath Luitel** [3] **Manmeet Singh** [3] **Josh Durkee** [3]

Abstract

Scientific peer review datasets have trained AI systems exclusively on Computer Science and Machine Learning venues, producing models that critique ablation studies yet have never seen a biology reviewer demand contamination controls or a chemist question Nuclear Magnetic Resonance (NMR) spectral assignments. We introduce FIRSTPASS, the first large-scale peer review dataset built on complete multi-round editorial dialogues from a multidisciplinary high-impact journal. Curated from *Nature Communications* mandatory transparent peer review (instituted November 2022), FIRSTPASS comprises 3,668 records spanning five scientific domains (biology, chemistry, neuroscience, physics, and earth science), capturing the full iterative structure of scientific validation: initial referee reports, author point-by-point responses, and updated reviewer assessments. Each record carries an outcome label derived directly from editorial decisions (STANDARD for two-round review; EXTENDED for three or more rounds), providing ground truth absent in all prior corpora. An automated audit confirms 100% content integrity. Expert reviews average 2,155 words, substantially denser than conference venue reviews. All data, parsing pipelines, and evaluation scripts are released to enable reproducible benchmarking of AI scientific judgment across disciplines.

1. Introduction

Every peer review dataset used to train AI systems draws exclusively from Computer Science and Machine Learning venues. PeerRead (Kang et al., 2018) established this precedent with 14.7K drafts from ACL, NIPS, and ICLR; ReviewMT (Tan et al., 2025) added multi-turn structure; MARG (D'Arcy et al., 2024) introduced multi-agent generation. All remain anchored to a single community's norms. A model trained on this data learns to critique ablation studies. It has never seen a biology reviewer demand contamination controls, a chemist question NMR spectral assignments, a neuroscientist challenge motion artifact correction, or an earth scientist dispute a stratigraphic age model. These are not edge cases: they represent the majority of scientific output.

Two deeper failures compound this domain narrowness. First, existing systems treat peer review as a one-shot event. The author-reviewer dialogue (where claims are stress-tested, new data introduced, and positions updated) is the central mechanism of scientific validation and is absent from all prior training corpora. Second, evaluation is circular: systems are graded on whether generated text *resembles* a review, not on whether the critique aligns with what editors actually demanded be changed. A model optimizing for stylistic mimicry is not exercising scientific judgment; it is producing review-shaped text.

We introduce FIRSTPASS: the first peer review dataset that is multidisciplinary by design, iterative by structure, and outcome-grounded by construction, comprising 3,668 multi-round dialogues from *Nature Communications* with outcome labels derived directly from real editorial decisions.

2. Dataset Construction

Source. We build on *Nature Communications* (ISSN 2041-1723) for four reasons: mandatory transparent peer review since November 2022 eliminates opt-in bias; multidisciplinary scope spans biology, chemistry, neuroscience, physics, and earth science; CC BY 4.0 licensing permits derivative use; and reviews averaging 2,155 words, substantially denser than conference venue critiques.

Collection. We query the Springer Nature OpenAccess API for papers published January 2023 to December 2025. Article and peer-review PDFs are parsed via Gemini-3.1-flash-lite-preview (Team et al., 2025) with engineered extraction prompts and JSON recovery (Borrelli, 2024), validated

[1]The University of Texas at Austin, Austin, TX, USA [2]RediMinds Inc., USA [3]Disaster Science Operations Center, Western Kentucky University, Bowling Green, KY, USA. Correspondence to: Prabhjot Singh <prabhjot.singh@utexas.edu>.

Accepted at the AI for Science Workshop at the 43^{rd} International Conference on Machine Learning, Seoul, South Korea, 2026.

against pdfplumber baselines. A record is retained only if all four paper sections (abstract, introduction, methods, results) exceed 20 words each and at least two complete review rounds are present. Discussion and conclusions are deliberately excluded: reviewers primarily critique methodology and results, and including author interpretations would bias models toward echoing conclusions rather than developing independent critical judgment.

Labels. Outcome labels derive from round count: two rounds equals STANDARD; three or more equals EXTENDED. This operationalization reflects editorial assessment of unresolved concerns without dependence on decision letter text, absent in 97.7% of records. The label is consistent across five disciplines (62.8%–71.2% STANDARD), confirming it captures structural patterns of reviewer-author tension rather than single-domain procedural norms.

Integrity and Release. Automated audit of all 3,668 retained records confirms zero hollow files and 100% content integrity (defined as non-empty peer-review text matching source PDFs across all structured fields). The remaining 2.3% of candidate records were excluded for failing the minimum section-length threshold or lacking a second complete review round. Domain distribution: biology (741), chemistry (744), neuroscience (739), physics (727), earth science (717). Paper-level 80/10/10 train/validation/test splits, stratified by domain and label, prevent leakage. All parsing code, extraction prompts, and audit scripts are released to enable independent replication.

3. Statistics and Tasks

Figure 1 illustrates the three-task curriculum; Table 1 reports domain-level statistics. Label balance is consistent across disciplines (62.8% to 71.2% STANDARD), confirming stratified splits maintain representative distributions. At 2,155 words average, FIRSTPASS reviews are more than five times longer than ICLR reviews (∼400 words) and structurally multi-round, providing training signal no prior corpus offers.

Domain	Total	Train	Val	Test	STD (%)	EXT (%)
Biology	741	593	74	74	63.2	36.8
Chemistry	744	595	75	74	62.8	37.2
Neuroscience	739	591	74	74	64.6	35.4
Physics	727	582	73	72	66.7	33.3
Earth Sci.	717	573	72	72	71.2	28.8
Total	**3,668**	**2,934**	**368**	**366**	**65.4**	**34.6**

Table 1. FIRSTPASS dataset statistics by domain. STD = STANDARD (2 rounds), EXT = EXTENDED (3+ rounds).

Three-task structure. Each paper generates three structured training examples targeting distinct scientific judgment capabilities. **Task 1 (Review Generation):** given paper content, generate Round 1 reviewer comments, training expert critique across five domains. **Task 2 (Reviewer Updating):** given paper content, Round 1 reviews, and author response, generate Round 2 reviews, training dialogue understanding: which responses are convincing, which introduce new data, which leave methodological debt unresolved. **Task 3 (Outcome Prediction):** given the full dialogue, predict STANDARD or EXTENDED, the primary evaluation task replacing circular stylistic proxies with real editorial decisions as ground truth.

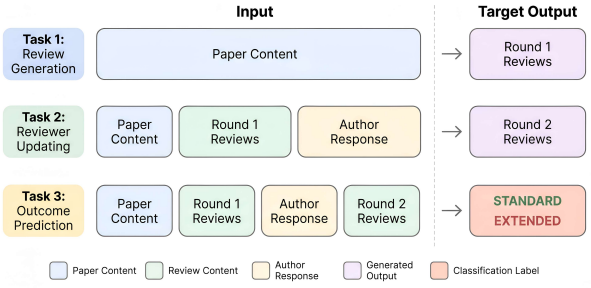

Figure 1. The FIRSTPASS three-task training curriculum.

A fine-tuned Qwen2.5-7B-Instruct (Qwen et al., 2025) achieves 80.5% accuracy and F1-macro 78.2% on Task 3 across all five domains, outperforming Gemini-3.1-flash-lite-preview zero-shot by 10.4 percentage points.

4. Release and Governance

License and access. FIRSTPASS inherits CC BY 4.0 from its source material. The dataset, parsing pipeline, extraction prompts, audit scripts, evaluation suite, and fine-tuned model weights are publicly available via Hugging Face Datasets and GitHub[†].

Schema and sustainability. Records are structured JSON; the schema is versioned (v1.0 at release) with fields covering paper metadata, domain label, four paper sections, separated review rounds, author responses, and outcome label. A unified evaluation script computes accuracy, F1-macro, and F1-EXT, enabling reproducible benchmarking. The pipeline supports annual longitudinal extension as new *Nature Communications* papers publish under mandatory transparent peer review.

Privacy. Reviewer identities are anonymized in the source material (Reviewer 1, 2, etc.) and preserved throughout. No personally identifiable information beyond the published record is included.

Governance signal. FIRSTPASS operationalizes an auditable deployment criterion: a model achieving F1-macro

[†]https://github.com/prabhjotschugh/firstpass-peer-review.

$\geq$ 75% on Task 3 demonstrates alignment with expert editorial judgment grounded in real outcomes, not stylistic proxies. Beyond peer review, this operationalization applies to any domain where AI judgment must be validated against real expert decisions rather than proxy metrics, including clinical triage, grant evaluation, and regulatory review, making FIRSTPASS a methodological template for outcome-grounded AI governance across scientific institutions. FIRSTPASS provides both the benchmark and the methodology to set that bar. Responsible-use risks include bias replication from a single journal source and potential misuse for synthetic review generation; we release an explicit datasheet documenting known limitations, accepted-papers-only scope, and recommended mitigations.

References

Borrelli, S. json-repair: A python library to repair invalid JSON, 2024. URL https://github.com/mangiucugna/json_repair. Accessed April 2026.

D'Arcy, M., Hope, T., Birnbaum, L., and Downey, D. Marg: Multi-agent review generation for scientific papers, 2024. URL https://arxiv.org/abs/2401.04259.

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 Walker, M., Ji, H., and Stent, A. (eds.), *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pp. 1647–1661, New Orleans, Louisiana, June 2018. Association for Computational Linguistics. doi: 10.18653/v1/N18-1149. URL https://aclanthology.org/N18-1149/.

Qwen, :, Yang, A., Yang, B., Zhang, B., Hui, B., Zheng, B., Yu, B., Li, C., Liu, D., Huang, F., Wei, H., Lin, H., Yang, J., Tu, J., Zhang, J., Yang, J., Yang, J., Zhou, J., Lin, J., Dang, K., Lu, K., Bao, K., Yang, K., Yu, L., Li, M., Xue, M., Zhang, P., Zhu, Q., Men, R., Lin, R., Li, T., Tang, T., Xia, T., Ren, X., Ren, X., Fan, Y., Su, Y., Zhang, Y., Wan, Y., Liu, Y., Cui, Z., Zhang, Z., and Qiu, Z. Qwen2.5 technical report, 2025. URL https://arxiv.org/abs/2412.15115.

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: Benchmarking large language models, 2025. URL https://openreview.net/forum?id=uV3Gdoq2ez.

Team, G., Anil, R., Borgeaud, S., Alayrac, J.-B., Yu, J., Soricut, R., Schalkwyk, J., Dai, A. M., Hauth, A., Millican, K., Silver, D., Johnson, M., Antonoglou, I., Schrittwieser, J., Glaese, A., Chen, J., Pitler, E., Lillicrap, T., Lazaridou, A., Firat, O., Molloy, J., Isard, M., Barham, P. R., Hennigan, T., Lee, B., Viola, F., Reynolds, M., Xu, Y., Doherty, R., Collins, E., Meyer, C., Rutherford, E., Moreira, E., Ayoub, K., Goel, M., Krawczyk, J., Du, C., Chi, E., Cheng, H.-T., Ni, E., Shah, P., Kane, P., Chan, B., Faruqui, M., Severyn, A., Lin, H., Li, Y., Cheng, Y., Ittycheriah, A., Mahdieh, M., Chen, M., Sun, P., Tran, D., Bagri, S., et al. Gemini: A family of highly capable multimodal models, 2025. URL https://arxiv.org/abs/2312.11805.