Reviewer Rating Variability and Confidence and Language Model Sentiment Prediction of Machine Learning Conference Papers

Yidan Sun, Ph.D. Student Mayank Kejriwal, Ph.D.

University of Southern California & USC Information Sciences Institute

Objective

OpenReview.net is a public platform for open peer review [1] that posts submitted papers, reviews, reviewer confidence ratings, and final decisions [2]. Using data from the ICLR 2019 Conference Track on OpenReview, we examine:

- **Q1: Confidence & Score Extremity** — Are higher reviewer confidence ratings associated with greater deviation from the rubric midpoint?
- **Q2: Variability & Rejection** — For papers with all reviews in the neutral range (scores 4–7), does greater reviewer disagreement (score variability) predict rejection?
- **Q3: Agreement with Actual Decisions** — Do a BERT-based sentiment model and a score-based ranking identify the same likely acceptances among neutral-scored papers, and what disparities arise when both methods agree on "Accept" but the actual decision is "Reject"?

Data

- **Source:** Source: OpenReview API, ICLR 2019 Conference Track; analyses conducted January 2025.
- **Dataset:** 4,332 reviews from 1,419 submitted papers (deadline: Sep 27, 2018).
- Each review includes:
  - **Score:** integer 1–10 (higher = stronger accept).
  - **Confidence:** integer 1–5 (higher = more confidence).
  - **Final decision:** Accept / Reject from meta-reviewer.
- Neutral papers are defined as those with all review scores between 4 and 7 inclusive (n = 775).

Methods

M1: Confidence vs. extremity

- Extremity = $|\text{score} - 5|$.
- Pearson correlation using all reviews.

M2: Variability in neutral papers

- Compare within-paper SD of scores for accepted vs. rejected papers using a two-sided t-test.

M3: Text sentiment for neutral papers

- Train **BERT** on polarized reviews (≤3 negative, ≥8 positive), matching the neutral definition (4–7 inclusive).
- Accuracy on held-out validation: 90.9%.
- Apply to all reviews of neutral papers; average per paper; rank by sentiment.
- Baseline: rank by mean score.
- Both methods label top N (246) as "Accept," matching the actual accepted neutral papers.

Results

Q1: Reviewer Confidence and Score Extremity

- Reviewers with higher confidence tended to give scores farther from the midpoint of the 10-point rubric.
- **Pearson** $r = 0.15$, 95% CI [0.10, 0.20], $p < 0.001$ (N = 4,332 reviews).

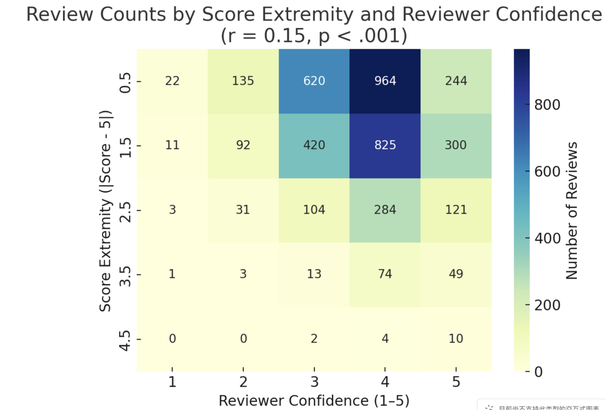


Figure: Relationship between reviewer confidence (1–5) and score extremity (|Score − 5|). Higher confidence is associated with more extreme scores.

Q2: Score Variability and Acceptance (Neutral Papers)

- Among neutral papers, accepted papers had lower within-paper score SD than rejected papers.
- Accepted: **mean SD = 0.59** (n = 246); Rejected: **mean SD = 0.80** (n = 529).
- Difference: **0.21** (95% CI: 0.15–0.28), $p < 0.001$.

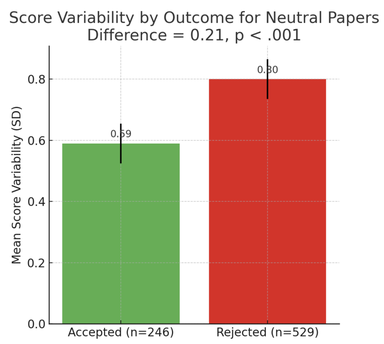


Figure: Mean reviewer score variability (SD) by meta-review outcome for neutral papers. Error bars indicate 95% confidence intervals.

Results

Q3: BERT vs. Score Baseline (Neutral Papers)

- **BERT:** Labeled 137 accepted papers and 109 rejected papers as "Accept."
- **Baseline:** Labeled 203 accepted papers and 43 rejected papers as "Accept."
- **Critical finding:** 20 neutral papers were labeled as "Accept" by both BERT and the score-based method yet were rejected in the actual meta-review, despite consistent positive indications from both independent assessments.
- **Unique cases:**
  - BERT only "Accept": 17 accepted, 89 rejected.
  - Baseline only "Accept": 83 accepted, 23 rejected.

Group	Accepted	Rejected	Total
BERT accept	137	109	246
Baseline accept	203	43	246
Both accept	120	**20**	140
Both reject	18	388	406
Total	246	529	775

Key Takeaways

- **Q1:** More confident reviewers tended to give scores further from the midpoint, indicating more decisive judgments.
- **Q2:** Among neutral papers, those with greater reviewer disagreement (higher score variability) were more likely to be rejected.
- **Q3:** Twenty neutral papers were classified as "Accept" by both independent criteria—a text-based sentiment model and a score-based ranking—yet were rejected in the actual meta-review. These cases suggest that some rejections occurred despite consistent positive indications from both scoring and textual sentiment.

References

[1] OpenReview: an open platform for peer review. Accessed July 15, 2025. `https://openreview.net`

[2] Ford E. Defining and characterizing open peer review: a review of the literature. *J Scholarly Pub*. 2013;44(4):311–326. doi:10.3138/jsp.44-4-00.

[3] Wolf T, Debut L, Sanh V, et al. Transformers: state-of-the-art natural language processing. In: *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*. Association for Computational Linguistics; 2020:38–45.