When Reviews Disagree: Fine-Grained Contradiction Analysis in Scientific Peer Reviews

Sandeep Kumar†, Yash Kamdar†, Abid Hossain†, Bharti Kumari‡, Tanik Saikh‡, Asif Ekbal†
†Department of Computer Science and Engineering, Indian Institute of Technology Patna, India
‡School of Computer Engineering, KIIT Deemed to be University, Bhubaneswar, India
{sandeep_2121cs29, 2201ai45_yash, abid_2311ai22, asif}@iitp.ac.in
{22052975, tanik.saikhfcs}@kiit.ac.in

Abstract

Scientific peer reviews frequently contain conflicting expert judgments, and the increasing scale of conference submissions makes it challenging for Area Chairs and editors to reliably identify and interpret such disagreements. Existing approaches typically frame reviewer disagreement as binary contradiction detection over isolated sentence pairs, abstracting away the review-level context and obscuring differences in the severity of evaluative conflict. In this work, we introduce a fine-grained formulation of reviewer contradiction analysis that operates over full peer reviews by explicitly identifying contradiction evidence spans and assigning graded disagreement intensity scores. To support this task, we present **RevCI**, an expert-annotated benchmark of peer-review pairs with evidence-level contradiction annotations with graded intensity labels. We further propose **IMPACT**, a structured multi-agent framework that integrates aspect-conditioned evidence extraction, deliberative reasoning, and adjudication to model reviewer contradictions and their intensity. To support efficient deployment, we distill IMPACT into **TIDE**, a small language model that predicts contradiction evidence and intensity in a single forward pass. Experimental results show that IMPACT substantially outperforms strong single-agent and generic multi-agent baselines in both evidence identification and intensity agreement, while TIDE achieves competitive performance at significantly lower inference cost. We make our code and dataset publicly available[1].

1 Introduction

Scientific peer review serves as the cornerstone of academic publishing, ensuring the quality and integrity of disseminated research (Bornmann, 2011). Despite its status as the *de facto* standard for validating scholarly research, the traditional peer-review process has come under increasing scrutiny due to perceived systemic fragilities. These concerns range from a lack of transparency (Wicherts, 2016; Parker et al., 2018) and inherent biases (Stelmakh et al., 2021, 2019) to the fundamental arbitrariness (Brezis and Birukou, 2020) and inconsistency of reviewer evaluations (Shah et al., 2018; Langford and Guzdial, 2015). Furthermore, peer review is often critiqued as an ill-defined task (Rogers and Augenstein, 2020), frequently failing to identify high-impact or influential contributions in their early stages (Freyne et al., 2010).

In recent years, the computational linguistics (CL) community has experienced an unprecedented surge in submission volumes, placing the peer review ecosystem under severe strain (Kelly et al., 2014; Gropp et al., 2017). Large-scale venues such as EMNLP 2025 now coordinate review cycles involving over 13,000 reviewers, exacerbating logistical challenges and contributing to declining review quality, including instances of "highly irresponsible" reviewing behavior (Christodoulopoulos et al., 2025). At the same time, the integrity of peer review is increasingly threatened by the unauthorized use of Large Language Models (LLMs) to generate reviews, which often produce superficial or formulaic feedback lacking substantive critical engagement (Liang et al., 2024c; Mishra, 2025). Beyond these operational pressures, the peer review process is inherently subjective, frequently giving rise to reviewer disagreement, where experts offer conflicting assessments of the same manuscript (Rogers and Augenstein, 2020). For Area Chairs and editors, reconciling these contradictions is the most labor-intensive aspect of the decision-making process. While some level of disagreement is expected in cutting-edge research, unresolved or unexplained conflicts can lead to inconsistent outcomes and a perceived lack of fairness in the evaluation system (Wang and Wan, 2018). Prior work has begun to explore computational methods for iden-

[1] https://github.com/sandeep82945/Contradiction-Intensity.git

Findings of the Association for Computational Linguistics: ACL 2026, pages 38267–38290

tifying reviewer disagreement in scientific peer review. Notably, ContraSciView (Kumar et al., 2023) formulates disagreement as a binary contradiction detection task over isolated sentence pairs. While effective for identifying explicit sentence-level conflicts, this formulation abstracts away review-level discourse and collapses graded differences in evaluative judgments into coarse binary labels, limiting its usefulness for Area Chairs who must reason about both the presence and the severity of conflicting reviews.

In this study, we introduce **IMPACT** (Intensity-based Multi-Agent Contradiction estimation), a structured multi-agent framework for fine-grained analysis of reviewer disagreement. Unlike prior approaches that reduce contradictions to binary labels over isolated sentence pairs, **IMPACT** operates over full review contexts to jointly identify aspect-conditioned contradiction evidence spans, generate natural-language explanations, and assign graded disagreement intensity scores. The framework is driven by a deliberative reasoning protocol in which multiple agents explicitly debate conflicting assessments, helping to mitigate individual biases and surface nuanced disagreements. To support this task, we present **RevCI**, an expert-annotated dataset of peer-review pairs with evidence-level contradiction annotations and graded intensity labels. Finally, to support efficient deployment, we introduce **TIDE**, a small language model (SLM) distilled from **IMPACT**'s deliberative reasoning traces. Experimental results show that **IMPACT** substantially outperforms existing baselines in contradiction detection, while **TIDE** achieves performance comparable to much larger models at a fraction of the inference cost.

We summarize our contributions as follows:

- We introduce a new task for analyzing reviewer disagreement that moves beyond binary contradiction detection by jointly identifying explicit contradiction evidence and assigning graded disagreement intensity scores in scientific peer reviews.
- We present **RevCI**, an expert-annotated benchmark of peer reviews that enables fine-grained analysis of reviewer disagreement through evidence-level contradiction annotations, graded intensity labels, and human-written explanation statements.
- We propose **IMPACT**, a structured multi-agent framework that integrates aspect-conditioned evidence extraction, deliberative reasoning, and adjudication to detect contradiction evidence and estimate disagreement intensity from full review contexts, substantially outperforming strong single-agent and generic multi-agent baselines.
- To support efficient deployment, we introduce **TIDE**, a small language model distilled from **IMPACT** that predicts contradiction evidence and intensity in a single forward pass, achieving competitive agreement with human annotations at significantly lower inference cost.

2 Related Work

Reviewer disagreement has been studied in the context of peer review through analyses of score variance (Bornmann and Daniel, 2010), sentiment divergence (Kang et al., 2018), and reviewer bias (Stelmakh et al., 2019; Shah et al., 2018). While these studies characterize disagreement at an aggregate level, they do not explicitly model textual contradictions between reviewer comments. Contradiction detection is commonly studied under the Natural Language Inference (NLI) framework (Bowman et al., 2015; Williams et al., 2018), with transformer-based models achieving strong performance on benchmark datasets (Devlin et al., 2019; Liu et al., 2019). However, prior work has shown that such models struggle with pragmatic and domain-specific contradictions in expert-written text (Nie et al., 2020; Gururangan et al., 2020). Peer reviews pose additional challenges due to hedging, technical assumptions, and discourse-level reasoning, which are difficult to capture using sentence-pair classification alone.

The work most closely related to ours is ContraSciView (Kumar et al., 2023), which frames reviewer disagreement as a contradiction detection problem. ContraSciView formulates the task as binary classification over isolated review sentence pairs using encoder-based models, such as BERT (Devlin et al., 2019). While effective for detecting explicit sentence-level conflicts, this formulation does not capture discourse-level disagreement expressed across full reviews, including hedged, comparative, and assumption-dependent judgments.

3 RevCI Dataset

We introduce RevCI (Review Contradiction Intensity), a dataset of expert-annotated peer-review

pairs with human-written contradiction summaries, contradiction evidence pairs, and graded intensity labels.

3.1 Dataset Collection and Re-Annotation

To annotate RevCI, we reuse the peer-review dataset from ContraSciView, derived from the ASAP-Review corpus (Yuan et al., 2021). The dataset contains reviews from 8,582 papers across ICLR (2017–2020) and NeurIPS (2016–2019). While the underlying data source remains unchanged, we conduct a new round of human annotation to support a more fine-grained analysis of reviewer disagreement. In contrast to prior work that focused on binary contradiction detection, our annotations capture evidence-grounded contradiction statements and graded intensity levels.

3.2 Review Pair Construction and LLM-Based Filtering

We define a *review* as the set of comments authored by a single reviewer. For a paper with n reviews, we construct all unordered pairs of reviews, resulting in $\binom{n}{2}$ review pairs per paper and approximately 28K review pairs across the dataset. Each review pair contains multiple candidate comment-level pairs formed by pairing individual comments from the two reviews.

Explicit contradictions between peer reviews are relatively rare, making uniform sampling inefficient for annotation. To address this, we apply an instruction-following LLM[2] as a screening model prior to human annotation. Given a review pair, the model predicts whether the pair contains a contradiction, and only pairs flagged as potentially contradictory are forwarded for annotation.

Due to space limitations, we discuss the **annotation guidelines, annotation process, and annotator compensation** in detail in Appendix D.

3.3 Final Dataset

The final dataset comprises 800 review pairs[3]. While RevCI is modest in size, its reliance on expert-written peer reviews and evidence-level, graded annotations makes large-scale collection costly, and similar dataset scales are standard for tasks requiring detailed expert judgment, as seen in fine-grained evaluation benchmarks such as SummEval (Fabbri et al., 2020) and Qasper (Dasigi et al., 2021). Appendix, Figures 4 and 5 summarize the annotation statistics for the contradiction-bearing subset.[4]

4 Methodology

4.1 Problem Formulation

Given two full peer reviews r_i and r_j, where i and j index distinct reviews of the same manuscript, our goal is to identify and characterize fine-grained contradictions between them. Unlike prior approaches that operate on isolated sentence pairs or assume pre-segmented review comments, we process full reviews end-to-end and generate a set of contradiction evidence pairs $E = \{(e_1^{(t)}, e_2^{(t)})\}_{t=1}^{m}$. Here, m denotes the total number of contradiction pairs, and each pair index t corresponds to a distinct disagreement instance between the two reviews. Each pair consists of atomic evidence statements grounded in r_i and r_j that express mutually incompatible judgments about the same underlying issue. Each evidence pair is associated with an aspect category $a_t \in \mathcal{A}$ (e.g., novelty, clarity, soundness), a discrete contradiction intensity score $\alpha_t \in \{1, 2, 3\}$, and a natural-language rationale ρ_t explaining the source and severity of the disagreement. The final task output is the collection $\{(e_1^{(t)}, e_2^{(t)}, a_t, \alpha_t, \rho_t)\}_{t=1}^{m}$, providing an aspect-aware, graded representation of reviewer disagreement.

4.2 Overall

Our methodology consists of two complementary components with distinct inference-time trade-offs. **IMPACT** is a standalone multi-agent framework that estimates contradiction intensity at inference time via aspect-conditioned evidence extraction, structured agent disagreement, and adjudication. Although this multi-agent deliberation incurs higher inference latency due to iterative agent interactions, it yields explicit, evidence-grounded reasoning traces. To enable efficient deployment, we further introduce **TIDE**, a Small Language Model (SLM) trained to directly predict contradiction evidence, intensity labels, and associated reasoning distilled from **IMPACT**, resulting in substantially faster inference.

[2]We used GPT-4o mini.

[3]RevCI contains 800 review pairs: 352 contain at least one contradiction, while the remaining 448 contain none and serve as negative instances for contradiction detection and FPR computation.

[4]Figures 4 and 5 are computed over the contradiction-bearing subset (352 pairs).

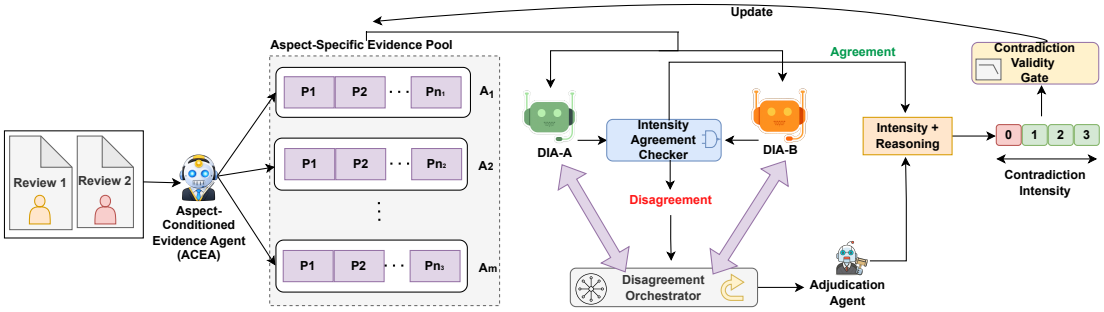

Figure 1: Overview of **IMPACT**, an Intensity-based Multi-Agent Contradiction estimation framework. The framework integrates aspect-conditioned evidence extraction with structured multi-agent disagreement to estimate contradiction intensity.

4.3 IMPACT: Intensity-based Multi-Agent Contradiction Estimation

We propose IMPACT (*Intensity-based Multi-Agent Contradiction estimation*), a multi-agent framework for estimating contradiction intensity by coordinating aspect-conditioned evidence extraction with structured agent disagreement and adjudication. IMPACT consists of an Aspect-Conditioned Evidence Agent (ACEA), two Deliberative Intensity Agents (DIA-A/B) coordinated by an Intensity Agreement Checker, a Disagreement Orchestrator, an Adjudication Agent for resolution, and a final Contradiction Validity Gate.

Aspect-Conditioned Evidence Agent (ACEA): ACEA operates on a single pair of reviews and serves as the high-recall evidence retrieval component of IMPACT. Given a review pair (r_i, r_j) and a predefined set of aspects $\mathcal{A} = \{a_1, \ldots, a_M\}$, ACEA conditions on each aspect a_m to identify candidate pairs that exhibit potential conflicting viewpoints. For each aspect a_m, ACEA produces paired evidence spans:

$$\mathcal{E}_{a_m}^{(i,j)} = f_{\text{ACEA}}(r_i, r_j, a_m) \subseteq \text{Spans}(r_i) \times \text{Spans}(r_j), \tag{1}$$

where f_{ACEA} denotes an LLM-based extraction function configured to retrieve pairs exhibiting potential semantic divergence. This high-recall filtering ensures that the system retains subtle contradictions while discarding clear agreements, providing a unified candidate set for intensity assessment. Across review pairs, extracted evidence is accumulated into an *Aspect-Specific Evidence Pool* $\mathcal{E} = \{\mathcal{E}_{a_1}, \ldots, \mathcal{E}_{a_M}\}$, where

$$\mathcal{E}_{a_m} = \bigcup_{(i,j)} \mathcal{E}_{a_m}^{(i,j)}. \tag{2}$$

Deliberative Intensity Agent (DIA): A Deliberative Intensity Agent (DIA) serves as the core reasoning unit for assigning graded contradiction intensity scores. Given an aspect-aligned evidence pair $(e_1^{(j)}, e_2^{(j)}) \in \mathcal{E}_{a_j}$, the agent functions as a probabilistic mapping that predicts a discrete intensity label $\alpha_j \in \{0, 1, 2, 3\}$ (following the rubric of contradiction intensity [5]) and generates a supporting explanation/reason for the assigned label ρ_j:

$$(\alpha_j, \rho_j) = g_{\text{DIA}}(e_1^{(j)}, e_2^{(j)}, r_i, r_j), \tag{3}$$

where r_i and r_j denote the full review contexts. Conditioning on the full context enables the agent to interpret localized evidence spans within the broader evaluative discourse of each reviewer, distinguishing genuine conflict from rhetorical differences. IMPACT employs two DIAs (DIA-A and DIA-B) which share a functional specification but may be instantiated using diverse underlying LLMs to encourage reasoning variance.

Intensity Agreement Checker: The Intensity Agreement Checker functions as a deterministic control gate. It compares the agents' initial independent predictions, α_j^A and α_j^B, to determine whether they agree (i.e., $\alpha_j^A = \alpha_j^B$). If agreement holds, the shared intensity label is accepted directly and propagated to downstream components without further interaction. Conversely, in the event of disagreement, the deliberation protocol is triggered and managed by the Disagreement Orchestrator.

Disagreement Orchestrator: The Disagreement Orchestrator (DO) manages structured interaction

[5] Here, label 0 denotes "no valid contradiction" (i.e., the candidate pair does not constitute a genuine contradiction and is filtered out by the validity gate). We will discuss this in detail in Section 4.3.

between DIA-A and DIA-B during disagreement. Given fixed predictions $\alpha_j^A \neq \alpha_j^B$, the orchestrator enforces a *score-locking* constraint:

$$\alpha_{j,t}^k = \alpha_{j,0}^k, \quad k \in \{A, B\},\ t \in \{1, \ldots, T\}. \tag{4}$$

This constraint prevents agents from exhibiting *conformity bias* or drifting toward premature "lazy consensus" (Du et al., 2024; Liang et al., 2024a). Instead, it compels each agent to deepen its reasoning and generate maximal conflicting evidence, ensuring the Adjudication Agent receives a comprehensive argumentative trace that exposes the nuances of the disagreement. Deliberation proceeds in alternating turns, where agents receive opponent rationales as feedback and respond by grounding arguments in evidence, clarifying intensity criteria, and addressing counter-arguments. The orchestrator maintains the debate history and terminates interaction after a fixed number of rounds.

Adjudication Agent: Inspired by LLM as judge (Liu et al., 2023; Mao et al., 2024), the Adjudication Agent resolves disagreements by arbitrating over deliberative reasoning rather than re-estimating intensity. Given the deliberation trace, it selects

$$\alpha_j^* \in \{\alpha_j^A, \alpha_j^B\} \tag{5}$$

and produces a consolidated intensity reasoning.

Contradiction Validity Gate (CVG): While contradiction intensity is defined over $\{1, 2, 3\}$, not all extracted aspect-aligned evidence pairs correspond to genuine contradictions. We, therefore, extend the label space with a null label $\alpha = 0$ to jointly model contradiction validity and severity. The Contradiction Validity Gate filters adjudicated outputs based on the final intensity score α_j^*. Evidence pairs with $\alpha_j^* = 0$ are discarded, while those with $\alpha_j^* \geq 1$ are retained. Accordingly, the aspect-specific evidence pool is updated as

$$\mathcal{E}_{a_m}^* = \{(e_1^{(j)}, e_2^{(j)}) \in \mathcal{E}_{a_m} \mid \alpha_j^* \geq 1\}, \tag{6}$$

ensuring that $\mathcal{E}$ contains only adjudicated, valid contradiction evidence with non-zero intensity.

4.4 TIDE: Teacher–Student Distillation for Evidence-Grounded Intensity Reasoning

While IMPACT produces high-quality, evidence-grounded contradiction judgments through deliberation and adjudication, its iterative multi-agent process is computationally expensive, limiting large-scale or real-time deployment. To address this limitation, we adopt a teacher–student paradigm, using IMPACT as a *teacher* to generate a high-fidelity synthetic dataset for training a computationally efficient Small Language Model (SLM) (Taori et al., 2023; Mitra et al., 2024).

Automated Data Generation: Given a pair of peer reviews (r_i, r_j) of the same paper, the teacher framework identifies a set of contradiction instances $\{c_j\}_{j=1}^m$. Each instance $c_j = (e_j, \alpha_j^*, \rho_j)$ consists of an aspect-aligned contradiction evidence pair $e_j = (e_1^{(j)}, e_2^{(j)})$, where $e_1^{(j)}$ and $e_2^{(j)}$ are verbatim sentences extracted from r_i and r_j, respectively, an adjudicated contradiction intensity label $\alpha_j^* \in \{1, 2, 3\}$, and a consolidated intensity reasoning ρ_j produced by IMPACT. Each training example therefore takes the full review pair (r_i, r_j) as input and the corresponding set of contradiction instances as structured supervision. We collect approximately 2,000 review pairs from ICLR 2021–2023 via OpenReview[6]. Since IMPACT-P produces higher-quality outputs than large language models such as GPT-5.2 (OpenAI, 2024), Gemini-3 Flash (Anil et al., 2023), LLaMA-4 Maverick (Meta AI, 2024), and existing multi-agent frameworks, we use it to generate contradiction annotations for these review pairs. Statistics of the resulting dataset are shown in Figures 6 and 7. We split the dataset into 80% and 20% for training and validation, respectively, and use the expert-annotated RevCI dataset for testing.

Student Model Training. The student SLM is trained to directly model the conditional distribution $p_\theta(\{c_j\} \mid r_i, r_j)$ using supervised fine-tuning (SFT) with a next-token prediction objective. To enable parameter-efficient adaptation, we inject Low-Rank Adaptation (LoRA) layers (Hu et al., 2022) into the attention and feed-forward modules of the student model. At inference time, TIDE produces contradiction evidence and the associated intensity score and reasoning in a single forward pass.

5 Experiments

5.1 Implementation Details

In the multi-agent framework, to eliminate stochastic variation, we disable nucleus and top-k sampling ($p = 1.0$, $k = 0$) and set the temperature to 0 for all agents. All experiments are run with a fixed random seed = 42, ensuring determinism across repeated runs. Duplicate contradictions are removed using ROUGE-L similarity with a threshold of 0.9.

[6] https://openreview.net/

We evaluate two variants of our framework: IMPACT (Proprietary), abbreviated as IMPACT-P, and IMPACT (Open-Access), abbreviated as IMPACT-OA. Both variants share the same pipeline structure, agent roles, and aspect-conditioned evidence mechanism, and differ only in the choice of underlying models. Details of the specific models used for each agent, along with additional experimental and training settings, are provided in Appendix A.

For our proposed model TIDE, we fine-tune an LLM (Meta-Llama-3-8B-Instruct[7]) using supervised fine-tuning with LoRA adapters. Training is performed for 5 epochs using the AdamW optimizer (Loshchilov and Hutter, 2019) with a fixed learning rate of 5×10^{-5}, cosine learning rate scheduling, and a warm-up ratio of 0.03. During supervised fine-tuning, only LoRA adapter parameters are updated, while all base model parameters are frozen. LoRA adapters are applied to the attention projection layers (`q_proj`, `k_proj`, `v_proj`, `o_proj`) and the feed-forward network projection layers (`gate_proj`, `up_proj`, `down_proj`).

5.2 Evaluation Metrics

Contradiction evidence extraction yields unordered sets of instances with variable cardinality and variable-length evidence, making count-based evaluation inadequate. We therefore evaluate evidence overlap using ROUGE-L and enforce a one-to-one alignment between the ground-truth set G and predicted set P via maximum-weight matching (Hungarian algorithm). When $|G| \neq |P|$, we discard weak alignments with ROUGE-L below λ_{match} and treat discarded/unmatched instances as FP/FN. We further report agreement metrics for evidence intensity over the matched pairs. Full details are provided in Appendix E.

6 Result and Analysis

6.1 Main Result

We report the results of our proposed multi-agent framework, IMPACT, under two settings: IMPACT-P and IMPACT-OA, along with our proposed SLM, TIDE, in Table 1. We find that the IMPACT framework advances the state of the art on both contradiction evidence detection and intensity estimation. In particular, IMPACT-P and IMPACT-OA achieve the lowest error rates and highest agreement among all baselines, including single-agent models, generic multi-agent frameworks, and the prior task-specific system ContraSciView. Compared to the strongest baseline, **CourtEval**, IMPACT-P reduces the average detection error (avg. FNR and FPR) by **31.2%** and improves the average agreement score (avg. κ, ρ, and τ) by **52.0%**. Even without proprietary models, IMPACT-OA achieves an **8.5%** reduction in average detection error and a **19.4%** improvement in average agreement over CourtEval, indicating that the gains primarily stem from the framework design rather than model scale. Finally, we report results for TIDE. As shown in Table 1, TIDE attains lower average evidence detection error than LLaMA-4 Maverick, with a **6.8%** reduction in avg. FNR/FPR, and achieves higher average agreement with human annotations than GPT-5.2, improving avg. (κ, ρ, τ) by **2.3%**. For fair comparison, all baseline LLMs are evaluated using the same prompting template provided in Appendix G. However, TIDE does not uniformly outperform all baselines across individual metrics; instead, it provides a favorable trade-off between false positive rate and human-alignment metrics. Additionally, TIDE requires only a single forward pass, making it substantially more efficient than multi-agent frameworks such as IMPACT. These results indicate that the structured supervision provided by IMPACT transfers fine-grained intensity reasoning to smaller models. We additionally perform a paired bootstrap test over the RevCI test set and find that the improvement in composite agreement score is statistically significant ($p < 0.05$).

6.2 Effect of Discussion Rounds

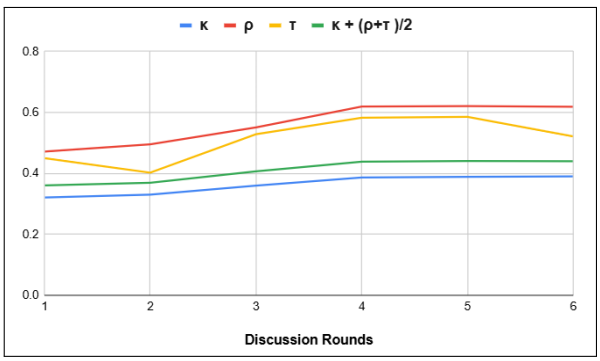

Figure 2: The figure shows the trends of the intensity agreement metrics κ, ρ, τ and their composite score $C = \kappa + (\rho + \tau)/2$ over successive discussion rounds.

To quantify the impact of deliberation depth, we

[7] https://huggingface.co/meta-llama/Meta-Llama-3-8B-Instruct

Category	Method	FNR ↓	FPR ↓	κ ↑	ρ ↑	τ ↑
Single-agent (CoT)	GPT-5.2 (OpenAI, 2024)	0.2935	0.3012	0.2612	0.3679	0.3043
	LLaMA-4 Maverick (Meta AI, 2024)	0.2639	0.4680	0.2348	0.3315	0.2652
	Gemini-3 Flash (Anil et al., 2023)	0.3741	0.2369	0.3041	0.4523	0.3799
	LLaMA-3-8B-Instruct (Grattafiori et al., 2024)	0.3570	0.3948	0.2488	0.3541	0.3364
	Qwen-2.5-7B-Instruct (Yang et al., 2025)	0.4127	0.4534	0.1859	0.2854	0.2283
	ContraSciView (Kumar et al., 2023)	0.3920	0.3360	–	–	–
Generic multi-agent	Self-Refine (Madaan et al., 2023)	0.2750	0.2950	0.2680	0.3810	0.3180
	Debate (Kim et al., 2024)	0.2870	0.2820	0.2550	0.3650	0.3010
	MAD (Liang et al., 2024b)	0.2690	0.2760	0.2720	0.3890	0.3270
	ChatEval (Chan et al., 2023)	0.2580	0.2640	0.2810	0.4020	0.3410
	CourtEval (Kumar et al., 2025)	0.2520	0.2590	0.2860	0.4100	0.3490
Our Proposed	**IMPACT-OA**	0.2390	0.2287	0.3270	0.4783	0.4421
	IMPACT-P	**0.1901**	**0.1613**	**0.3862**	**0.6193**	**0.5826**
	TIDE	0.3771	0.3048	0.2202	0.3793	0.3549

Table 1: Overall comparison of contradiction detection and intensity agreement. We compute FNR and FPR at the *review-pair level* over full test set (positive = contradiction present; negative = no contradiction), where $\mathrm{FPR} = \frac{\mathrm{FP}}{\mathrm{FP+TN}}$ and $\mathrm{FNR} = \frac{\mathrm{FN}}{\mathrm{TP+FN}}$. We compute κ, ρ, and τ over matched contradiction evidence pairs (i.e., where intensity is defined). Here, CoT means using chain of thought prompting.

vary the number of discussion rounds D in our multi-agent deliberation framework, and let a downstream judge produce the final intensity label conditioned on the resulting dialogue. Figure 2 reports agreement with human annotations using Cohen's κ, which measures categorical agreement corrected for chance; Spearman's ρ, which captures rank-order correlation; Kendall's τ, which measures pairwise ranking consistency. These metrics capture complementary aspects of agreement: κ evaluates exact label matching, while ρ and τ assess ordinal consistency between predictions and human judgments. To summarize overall performance, we define a composite score $C = \kappa + (\rho + \tau)/2$, which balances categorical agreement with ordinal consistency.

Overall, increasing D yields consistent improvements up to a small number of rounds, after which gains saturate. The composite score improves from $C = 0.3608$ with a single round to 0.3691 with two rounds (+2.3%). Increasing to three rounds yields a pronounced jump to 0.4068 (+9.26%), the largest relative gain, accompanied by concurrent improvements across all constituent metrics. Moving from three to four rounds provides an additional but attenuated improvement (+7.76%). Beyond four rounds, the marginal benefit becomes negligible: increasing from four to five rounds yields only a +0.57% change, while extending to six rounds slightly decreases performance (-0.27%). Taken together, these results suggest that most of the benefit of deliberation is achieved within the first four rounds, indicating convergence and motivating $D = 4$ as a practical operating point. We also discuss false positive and false negative rates in Appendix C.

6.3 Ablation Study of IMPACT

Table 2 reports that without aspect conditioning, the model exhibits a relatively high FNR of 0.2969, indicating that a substantial fraction of true contradictions are missed. Introducing aspect conditioning reduces the FNR to 0.1092, corresponding to a 63% relative reduction and an increase in recall from 0.703 to 0.891. This result indicates that explicit aspect conditioning substantially improves the model's ability to identify true contradictions.

At the same time, aspect conditioning (AC) increases the FPR from 0.3661 to 0.5120. This trade-off is expected: conditioning the model on a specific aspect reduces ambiguity about relevance, but can also lead to over-identification, which can be filtered out later using other methods. These results indicate that aspect conditioning helps reduce the FNR.

Table 2 also shows that when we use a single agent[8]. For intensity scoring with examples (IS+IEx), the FNR and FPR show an average reduction of 11.7% relative to scoring without intensity examples, along with a 35.68% increase in average agreement metrics. These results demonstrate the importance of intensity examples in the framework understanding what each intensity score means.

We then combine AC and IS+IEx to achieve a

[8]We prompt a single LLM for generating contradictions and assigning them an intensity score (1-3)

DO	ACEA	IEx	IS	CVG	FNR↓	FPR↓	κ↑	ρ↑	τ↑
×	×	×	×	×	0.2969	0.3661	–	–	–
×	✓	×	×	×	0.1092	0.5120	–	–	–
×	×	×	✓	×	0.3570	0.3948	0.2488	0.3541	0.3364
×	×	✓	✓	×	0.3293	0.3346	0.3392	0.5134	0.4219
×	✓	✓	✓	✓	0.1953	0.2614	0.3115	0.4574	0.4308
✓	✓	✓	✓	✓	**0.1901**	**0.1613**	**0.3862**	**0.6193**	**0.5826**

Table 2: Ablation of IMPACT components. DO: Disagreement Orchestrator; ACEA: Aspect Conditioning Evidence Agent; IEx: intensity examples; IS: intensity scoring; CVG: Contradiction Validity Gate.

low FNR and in order to filter out the increased false positives incorporated by aspect conditioning we use CVG. Thus by combining AC, IS+IEx and CVG we reduce the FNR and FPR by 40.69% and 21.87% respectively, compared to the IS+IEx agent. This shows that this component is important part of our framework where AC increases total contradictions detected reducing FNR and CVG filtering out false positives reducing FPR.

We add DO, DIAs, and an Adjudication agent, which results in a further reduction of the false positive rate (FPR) by 38.29% and an increase in the average agreement score by 32.37%. These results indicate that iterative reasoning and discussion help further remove incorrect contradiction evidence and improve the agreement of intensity assignments.

6.4 Ablation Study of TIDE

Table 3 reports the ablation study of different components of TIDE. When prompted with identical instructions to generate contradictions from paired reviews in the test set, the LLaMA-3.1-8B-Instruct base model (×FT (Finetuned), ×IS (Intensity Scorer), ×IR (Intensity Reasoning)) exhibits high error rates, with an FNR of 0.6641 and an FPR of 0.4699, as shown in Table 3. Fine-tuning the model without incorporating intensity supervision (✓FT, ×IS, ×IR) substantially reduces these errors, achieving an average relative reduction of 33.40% in FPR and FNR.

Introducing explicit intensity scoring supervision during fine-tuning (✓FT, ✓IS, ×IR) further improves performance, lowering the FNR to 0.4111 and the FPR to 0.3555, while also enabling meaningful alignment with human intensity annotations, reflected in a 67.2% increase in average agreement metrics compared to the base model with intensity scoring. These results indicate that incorporating intensity scoring supervision helps the LLM better distinguish between contradictions and non-contradictions.

Finally, incorporating explicit intensity reasoning during training (✓FT, ✓IS, ✓IR) yields the best overall result. This full TIDE configuration achieves the lowest error rates (FNR 0.3771, FPR 0.3048) and the strongest alignment with human judgments ($\kappa = 0.2202$, $\rho = 0.3793$, $\tau = 0.3549$), demonstrating an average reduction in FPR and FNR values by 11.04% as compared to the fine-tuned model without intensity reasoning. Thus, training on intensity reasoning helps LLMs understand the logic behind the intensity scores and better distinguish between different scores, increasing the average score of alignment metrics.

6.5 Human Evaluation

Our model mainly fails due to (i) misreading vague/scalar evaluative language as contradictions and (ii) confusing which aspect is being discussed when sentences mention multiple aspects. We discuss the error analysis and the case study in detail in Appendices F and H, respectively.

7 Conclusion and Future Work

We study reviewer disagreement through fine-grained contradiction analysis over full scientific peer reviews, introducing a new task and the **RevCI** benchmark for evidence-grounded contradiction identification and graded intensity modeling. We propose **IMPACT**, a structured multi-agent framework for contradiction detection and intensity estimation, and **TIDE**, a distilled small language model that achieves competitive performance with substantially lower inference cost. Our results show that reviewer disagreement is often fine-grained and context-dependent, making binary contradiction detection over isolated sentence pairs insufficient, and that IMPACT consistently outperforms strong single-agent and generic multi-agent baselines, including settings without proprietary models, with gains driven by task-specific deliberative design rather than model scale. Finally, TIDE

FT	IS	IR	FNR	FPR	κ	ρ	τ
×	×	×	0.6641	0.4699	–	–	–
✓	×	×	0.4162	0.3390	–	–	–
×	✓	×	0.5279	0.4529	-0.0250	0.1384	0.1231
✓	✓	×	0.4111	0.3555	0.1699	0.2980	0.2530
✓	✓	✓	**0.3771**	**0.3048**	**0.2202**	**0.3793**	**0.3549**

Table 3: Ablation of TIDE components. FT: Finetuned; IS: Intensity Scoring; IR:Intensity Reasoning.

demonstrates that this fine-grained intensity reasoning can be effectively distilled into an efficient model while maintaining strong alignment with human annotations, and future work will explore generalization beyond the computer science domain and extension to multi-reviewer disagreement.

Limitation

We restrict our experiments to the most frequent aspect categories in peer reviews, i.e., *Motivation, Clarity, Soundness, Substance, Originality*, and *Meaningful Comparison*. These aspects account for the majority of explicit contradiction instances in our annotated data, enabling reliable annotation and stable evaluation. While prior work identifies additional aspect categories (Lu et al., 2025), many occur too sparsely to support robust modeling. Our framework supports extensibility: in IMPACT, new aspects can be incorporated by updating the ACEA prompt with the corresponding aspect name and definition, without retraining. In contrast, extending TIDE to unseen aspect categories requires retraining, as the student model learns aspect-specific decision boundaries during distillation; however, this retraining is lightweight due to parameter-efficient fine-tuning. Our dataset is constructed from peer reviews primarily from ICLR and NeurIPS, and may not fully capture the diversity of review practices across venues, domains, or reviewer expertise levels. Additionally, the LLM-based pre-filtering used to select candidate review pairs may bias the dataset toward more explicit contradictions, potentially underrepresenting subtle or implicit disagreements. However, all final annotations are produced by expert annotators using full review context, ensuring that the benchmark reflects human judgment rather than model-generated labels. We leave broader cross-venue validation and more diverse sampling strategies to future work.

Ethics

This work uses peer-review texts derived from publicly available sources and synthetic data generated by our models; all data are anonymized where applicable and contain no personal, sensitive, or identifying information. The proposed system is intended solely as an assistive tool for editors and area chairs to help identify potential contradictions in lengthy and complex reviews, and it is not designed to automate editorial decisions or to evaluate authors or reviewers. As with any AI-based system, the model is not perfectly accurate and may produce false positives or false negatives; therefore, all outputs must be interpreted with human oversight and verified using standard editorial judgment. For example, when one reviewer states, "The paper does not have any new findings," while another notes, "The paper is somewhat novel," such disagreement may not be immediately apparent in long reviews, and the system aims only to surface such cases to prompt further inspection. Failure to identify a contradiction does not imply that none exists, and contradictions identified outside the system's outputs should still be handled according to established review guidelines. We caution against misuse of the system for monitoring or penalizing reviewers or for exposing model outputs directly to authors, and emphasize that it should be deployed only for internal editorial support with appropriate safeguards. We used an AI-based writing assistant (ChatGPT) for minor grammatical corrections and proofreading.

Acknowledgement

Sandeep Kumar acknowledges the Prime Minister Research Fellowship (PMRF) program of the Govt of India for its support.

References

Rohan Anil, Sebastian Borgeaud, Jean-Baptiste Alayrac, Jiahui Yu, Radu Soricut, Johan Schalkwyk, Andrew M. Dai, Anja Hauth, Katie Millican, David Silver, and 1 others. 2023. Gemini: A family of highly capable multimodal models. *arXiv preprint arXiv:2312.11805*. Version v5.

Lutz Bornmann. 2011. Scientific peer review. *Annu. Rev. Inf. Sci. Technol.*, 45(1):197–245.

Lutz Bornmann and Hans-Dieter Daniel. 2010. The usefulness of peer review for selecting manuscripts for publication: a utility analysis taking as an example a high-impact journal. *PloS one*, 5(6):e11344.

Samuel R. Bowman, Gabor Angeli, Christopher Potts, and Christopher D. Manning. 2015. A large annotated corpus for learning natural language inference. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing, EMNLP 2015, Lisbon, Portugal, September 17-21, 2015*, pages 632–642. The Association for Computational Linguistics.

Elise S. Brezis and Aliaksandr Birukou. 2020. Arbitrariness in the peer review process. *Scientometrics*, 123(1):393–411.

Chi-Min Chan, Weize Chen, Yusheng Su, Jianxuan Yu, Wei Xue, Shanghang Zhang, Jie Fu, and Zhiyuan Liu. 2023. Chateval: Towards better llm-based evaluators through multi-agent debate. *arXiv preprint arXiv:2308.07201*.

Christos Christodoulopoulos, Tanmoy Chakraborty, Carolyn Rose, and Violet Peng, editors. 2025. *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics, Suzhou, China.

Jacob Cohen. 1960. A coefficient of agreement for nominal scales. *Educational and psychological measurement*, 20(1):37–46.

Pradeep Dasigi, Kyle Lo, Iz Beltagy, Arman Cohan, Noah A. Smith, and Matt Gardner. 2021. A dataset of information-seeking questions and answers anchored in research papers. *CoRR*, abs/2105.03011.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, NAACL-HLT 2019, Minneapolis, MN, USA, June 2-7, 2019, Volume 1 (Long and Short Papers)*, pages 4171–4186. Association for Computational Linguistics.

Yilun Du, Shuang Li, Antonio Torralba, Joshua B. Tenenbaum, and Igor Mordatch. 2024. Improving factuality and reasoning in language models through multiagent debate. In *Forty-first International Conference on Machine Learning, ICML 2024, Vienna, Austria, July 21-27, 2024*. OpenReview.net.

Alexander R Fabbri, Wojciech Kryściński, Bryan McCann, Caiming Xiong, Richard Socher, and Dragomir Radev. 2020. Summeval: Re-evaluating summarization evaluation. *arXiv preprint arXiv:2007.12626*.

Jill Freyne, Lorcan Coyle, Barry Smyth, and Padraig Cunningham. 2010. Relative status of journal and conference publications in computer science. *Commun. ACM*, 53(11):124–132.

Aaron Grattafiori, Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, and 1 others. 2024. The llama 3 herd of models. *arXiv preprint arXiv:2407.21783*. Version v3.

Robert E. Gropp, Scott Glisson, Stephen Gallo, and Lisa Thompson. 2017. Peer Review: A System under Stress. *BioScience*, 67(5):407–410.

Suchin Gururangan, Ana Marasovic, Swabha Swayamdipta, Kyle Lo, Iz Beltagy, Doug Downey, and Noah A. Smith. 2020. Don't stop pretraining: Adapt language models to domains and tasks. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, ACL 2020, Online, July 5-10, 2020*, pages 8342–8360. Association for Computational Linguistics.

Edward J Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, Weizhu Chen, and 1 others. 2022. Lora: Low-rank adaptation of large language models. *ICLR*, 1(2):3.

Dongyeop Kang, Waleed Ammar, Bhavana Dalvi, Madeleine van Zuylen, Sebastian Kohlmeier, Eduard Hovy, and Roy Schwartz. 2018. A dataset of peer reviews (PeerRead): Collection, insights and NLP applications. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 1647–1661, New Orleans, Louisiana. Association for Computational Linguistics.

Jacalyn Kelly, Tara Sadeghieh, and Khosrow Adeli. 2014. Peer review in scientific publications: Benefits, critiques, & a survival guide. *EJIFCC*, 25:227–43.

Maurice G Kendall. 1938. A new measure of rank correlation. *Biometrika*, 30(1-2):81–93.

Alex Kim, Keonwoo Kim, and Sangwon Yoon. 2024. DEBATE: devil's advocate-based assessment and text evaluation. In *Findings of the Association for Computational Linguistics, ACL 2024, Bangkok, Thailand and virtual meeting, August 11-16, 2024*, pages 1885–1897. Association for Computational Linguistics.

Sandeep Kumar, Tirthankar Ghosal, and Asif Ekbal. 2023. When reviewers lock horns: Finding disagreements in scientific peer reviews. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 16693–16704, Singapore. Association for Computational Linguistics.

Sandeep Kumar, Abhijit A Nargund, and Vivek Sridhar. 2025. CourtEval: A courtroom-based multi-agent evaluation framework. In *Findings of the Association for Computational Linguistics: ACL 2025*, pages 25875–25887, Vienna, Austria. Association for Computational Linguistics.

John Langford and Mark Guzdial. 2015. The arbitrariness of reviews, and advice for school administrators. *Commun. ACM*, 58(4):12–13.

Tian Liang, Zhiwei He, Wenxiang Jiao, Xing Wang, Yan Wang, Rui Wang, Yujiu Yang, Shuming Shi, and Zhaopeng Tu. 2024a. Encouraging divergent thinking in large language models through multi-agent debate. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing, EMNLP 2024, Miami, FL, USA, November 12-16, 2024*, pages 17889–17904. Association for Computational Linguistics.

Tian Liang, Zhiwei He, Wenxiang Jiao, Xing Wang, Yan Wang, Rui Wang, Yujiu Yang, Shuming Shi, and Zhaopeng Tu. 2024b. Encouraging divergent thinking in large language models through multi-agent debate. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 17889–17904, Miami, Florida, USA. Association for Computational Linguistics.

Weixin Liang, Zachary Izzo, Yaohui Zhang, Haley Lepp, Hancheng Cao, Xuandong Zhao, Lingjiao Chen, Haotian Ye, Sheng Liu, Zhi Huang, Daniel A. McFarland, and James Y. Zou. 2024c. Monitoring ai-modified content at scale: a case study on the impact of chatgpt on ai conference peer reviews. In *Proceedings of the 41st International Conference on Machine Learning*, ICML'24. JMLR.org.

Yang Liu, Dan Iter, Yichong Xu, Shuohang Wang, Ruochen Xu, and Chenguang Zhu. 2023. G-eval: NLG evaluation using gpt-4 with better human alignment. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 2511–2522, Singapore. Association for Computational Linguistics.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. Roberta: A robustly optimized BERT pretraining approach. *CoRR*, abs/1907.11692.

Ilya Loshchilov and Frank Hutter. 2019. Decoupled weight decay regularization. In *International Conference on Learning Representations (ICLR)*.

Sheng Lu, Ilia Kuznetsov, and Iryna Gurevych. 2025. Identifying aspects in peer reviews. In *Findings of the Association for Computational Linguistics: EMNLP 2025*, pages 6145–6167, Suzhou, China. Association for Computational Linguistics.

Aman Madaan, Niket Tandon, Prakhar Gupta, Skyler Hallinan, Luyu Gao, Sarah Wiegreffe, Uri Alon, Nouha Dziri, Shrimai Prabhumoye, Yiming Yang, and 1 others. 2023. Self-refine: Iterative refinement with self-feedback. *Advances in Neural Information Processing Systems*, 36:46534–46594.

Rui Mao, Guanyi Chen, Xulang Zhang, Frank Guerin, and Erik Cambria. 2024. Gpteval: A survey on assessments of chatgpt and GPT-4. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation, LREC/COLING 2024, 20-25 May, 2024, Torino, Italy*, pages 7844–7866. ELRA and ICCL.

Meta AI. 2024. LLaMA 4: Multimodal Intelligence. https://ai.meta.com/blog/llama-4-multimodal-intelligence/. Accessed March 2025.

Udgam Mishra. 2025. Challenges in the peer-review process. *Journal of Indonesian Management*, 5(1):17.

Arindam Mitra, Hamed Khanpour, Corby Rosset, and Ahmed Awadallah. 2024. Orca-math: Unlocking the potential of slms in grade school math. *arXiv preprint arXiv:2402.14830*.

Yixin Nie, Adina Williams, Emily Dinan, Mohit Bansal, Jason Weston, and Douwe Kiela. 2020. Adversarial NLI: A new benchmark for natural language understanding. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, ACL 2020, Online, July 5-10, 2020*, pages 4885–4901. Association for Computational Linguistics.

OpenAI. 2024. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*.

Timothy Parker, Simon Griffith, Judith Bronstein, Fiona Fidler, Susan Foster, Hannah Fraser, Wolfgang Forstmeier, Jessica Gurevitch, Julia Koricheva, Ralf Seppelt, Morgan Tingley, and Shinichi Nakagawa. 2018. Empowering peer reviewers with a checklist to improve transparency. *Nature Ecology & Evolution*, 2:929–935.

Anna Rogers and Isabelle Augenstein. 2020. What can we do to improve peer review in NLP? In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 1256–1262, Online. Association for Computational Linguistics.

Nihar B. Shah, Behzad Tabibian, Krikamol Muandet, Isabelle Guyon, and Ulrike von Luxburg. 2018. Design and analysis of the NIPS 2016 review process. *J. Mach. Learn. Res.*, 19:49:1–49:34.

C Spearman. 2010. The proof and measurement of association between two things. *International Journal of Epidemiology*, 39(5):1137–1150.

Ivan Stelmakh, Nihar B. Shah, and Aarti Singh. 2019. On testing for biases in peer review. In *Advances in Neural Information Processing Systems 32: Annual Conference on Neural Information Processing Systems 2019, NeurIPS 2019, December 8-14, 2019, Vancouver, BC, Canada*, pages 5287–5297.

Ivan Stelmakh, Nihar B. Shah, Aarti Singh, and Hal Daumé III. 2021. Prior and prejudice: The novice reviewers' bias against resubmissions in conference peer review. *Proc. ACM Hum. Comput. Interact.*, 5(CSCW1):75:1–75:17.

Rohan Taori, Ishaan Gulrajani, Tianyi Zhang, Yann Dubois, Xuechen Li, Carlos Guestrin, Percy Liang, and Tatsunori B Hashimoto. 2023. Alpaca: A strong, replicable instruction-following model. *Stanford Center for Research on Foundation Models. https://crfm. stanford. edu/2023/03/13/alpaca. html*, 3(6):7.

Pauli Virtanen and 1 others. 2020. Scipy 1.0: Fundamental algorithms for scientific computing in python. *Nature Methods*, 17:261–272.

Ke Wang and Xiaojun Wan. 2018. Sentiment analysis of peer review texts for scholarly papers. In *The 41st International ACM SIGIR Conference on Research & Development in Information Retrieval*, SIGIR '18, page 175–184, New York, NY, USA. Association for Computing Machinery.

J.M. Wicherts. 2016. Peer review quality and transparency of the peer-review process in open access and subscription journals. *PLOS ONE*, 11(1).

Adina Williams, Nikita Nangia, and Samuel R. Bowman. 2018. A broad-coverage challenge corpus for sentence understanding through inference. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, NAACL-HLT 2018, New Orleans, Louisiana, USA, June 1-6, 2018, Volume 1 (Long Papers)*, pages 1112–1122. Association for Computational Linguistics.

An Yang, Anfeng Li, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Gao, Chengen Huang, Chenxu Lv, Chujie Zheng, Dayiheng Liu, Fan Zhou, Fei Huang, Feng Hu, Hao Ge, Haoran Wei, Huan Lin, Jialong Tang, and 41 others. 2025. Qwen3 technical report. *arXiv preprint arXiv:2505.09388*.

Weizhe Yuan, Pengfei Liu, and Graham Neubig. 2021. Can we automate scientific reviewing? *CoRR*, abs/2102.00176.

Appendix

A Additional Implementation Details

For Gemini-based models, the maximum number of generated tokens is fixed to 4096, while OpenAI and LLaMA models have a maximum output of 8,192 new tokens.

IMPACT-P uses proprietary models for deliberation and adjudication, with LLaMA-4-Maverick and Gemini-3-flash serving as Deliberative Intensity Agents (DIA), GPT-5.2 as the adjudication agent, and ACEA providing aspect-conditioned evidence selection. IMPACT-OA replaces proprietary components with openly accessible models, using Qwen3-32b and GPT-OSS-20b as DIA agents and LLaMA-4-Maverick as the adjudication agent, while using LLaMA-3.3-70b for ACEA. We discuss the role and prompt used by ACEA, DIAs, and the adjudication agent in detail in Appendix G and H. All generic multi-agent baselines (e.g., Debate, MAD, Self-Refine, ChatEval, CourtEval) were implemented under a controlled setting to ensure fair comparison. We used the same underlying LLM (GPT-5.2) and consistent prompting across all baselines. For iterative frameworks, we followed the termination criteria specified in their respective original papers. The per-device batch size is set to 2 with gradient accumulation over 8 steps (effective batch size of 16). We use a 20% validation split and an 80% test split, with a fixed random seed of 42.

LoRA adapters are applied to the attention and feed-forward projection layers with rank 8, scaling factor 16, and dropout 0.1. Training is conducted in `bfloat16` precision with a maximum sequence length of 4096 tokens, and loss is computed only on assistant responses. The best checkpoint is selected based on validation loss. During both training and inference, the LLM has a maximum output of 768 new tokens. During inference, the fine-tuned model is evaluated using beam search with 2 beams and a temperature of 0.01. All training and inference experiments are conducted on NVIDIA A100 GPUs with 80 GB memory, using two compute nodes.

B Token Efficiency Analysis

We randomly sampled 100 instances and computed the average number of tokens used by each method. The results are shown in Table 4. We use a standard GPT tokenizer[9] for this calculation. These results show that TIDE is substantially more token-efficient than multi-agent baselines.

Method	Avg Tokens per Sample
TIDE	1892.9
CoT (Single)	2410.1
IMPACT-P	36184.9
IMPACT-OA	39412.6
Self-Refine	18741.5
Debate	26483.8
MAD	23861.4
ChatEval	34712.9
CourtEval	37184.2

Table 4: Average token usage per sample across methods.

C Result: FNR and FPR value

Across multiple experimental runs in which the number of discussion rounds in the multi-agent deliberation framework is varied as a hyperparameter, we further analyze error characteristics of the framework in detecting contradictions by examining the false negative rate (FNR), false positive

[9] https://platform.openai.com/tokenizer

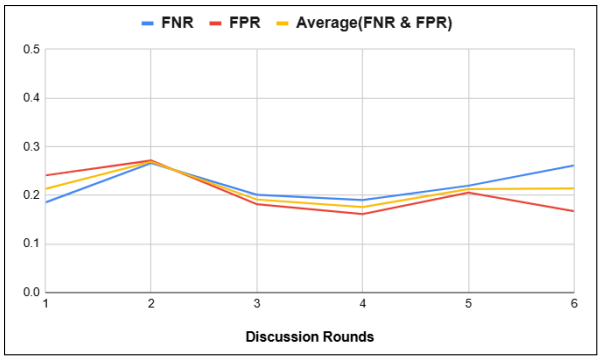

Figure 3: False Negative Rate (FNR), False Positive Rate (FPR), and their average as a function of the number of discussion rounds in the multi-agent deliberation framework.

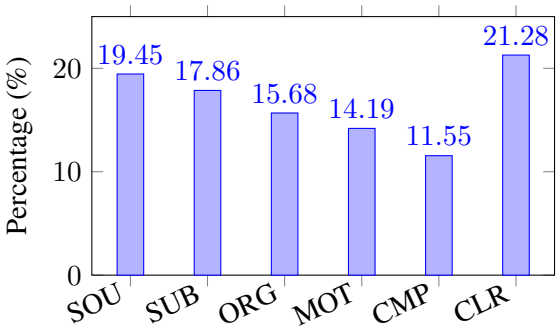

Figure 4: Percentage distribution of contradictions across aspects in the human-annotated test dataset. SOU: Soundness, SUB: Substance, ORG: Originality, MOT: Motivation, CMP: Meaningful Comparison, CLR: Clarity.

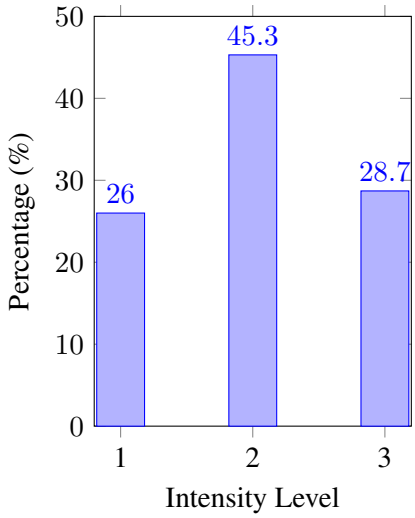

Figure 5: Percentage distribution of contradiction intensities in the human-annotated test dataset.

rate (FPR), and their average (see Figure 3). With a single discussion round, the average error rate is 0.2131. Increasing the discussion depth to two rounds leads to a sharp increase in the average error rate to 0.2689 (+26.1%), indicating that limited deliberation can amplify inconsistent or overconfident judgments. However, as the number of rounds increases beyond this point, error rates decrease substantially: the average drops to 0.1914 (-28.82%) at three rounds and reaches its minimum at four rounds (0.1757, -8.2%), corresponding to simultaneous reductions in both FNR and FPR. Further increases in discussion depth yield diminishing and unstable effects, with the average error rising again to 0.2127 at five rounds and 0.2143 at six rounds, driven by renewed increases in false negatives and fluctuations in false positives. Consistent with the agreement analysis, these results indicate that four discussion rounds achieve the best balance between missed and spurious detections.

D Annotation Details

D.1 Annotation Guidelines

The goal of the annotation task is to identify contradictions between two reviews written for the same paper, determine the severity of the contradiction, and give a brief explanation on the severity of the score given. Follow the steps below when annotating any pair of reviews.

Step 1: Understand the paper. Read the **title** and **abstract** of the paper to understand the background, motivation, and main contributions made by the paper. You may read these sections multiple times for better clarity. If a reviewer comment refers to specific technical details, you may briefly consult the paper to clarify the context.

Step 2: Read the reviews carefully. Read both reviews thoroughly. Pay attention to the opinions and judgments expressed by each reviewer, identify the aspects of the paper being discussed and the reviewers' sentiment behind the same (e.g., novelty, soundness, clarity, experiments, motivation). Do not treat individual sentences in isolation; consider the full context of each review.

Step 3: Identify contradictions. Identify pairs of statements that refer to the same aspect of the paper and determine whether they contradict each other. If multiple sentences contribute to a contradiction, then all of them must be included as evidences. A contradiction exists when the two comments express mutually incompatible judgments about the same issue. The contradiction could be explicit as well as implicit. The statements must have opposing sentiments (eg. positive vs. nega-

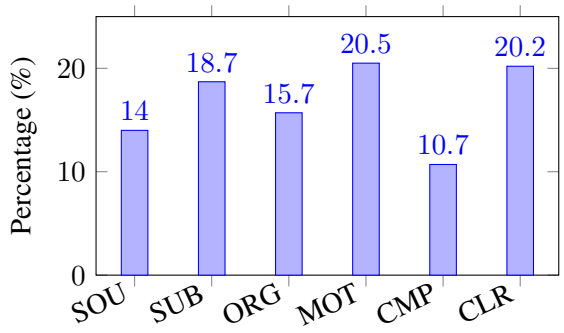

Figure 6: Aspect-wise percentage distribution of contradictions in synthetic data. SOU: Soundness, SUB: Substance, ORG: Originality, MOT: Motivation, CMP: Meaningful Comparison, CLR: Clarity.

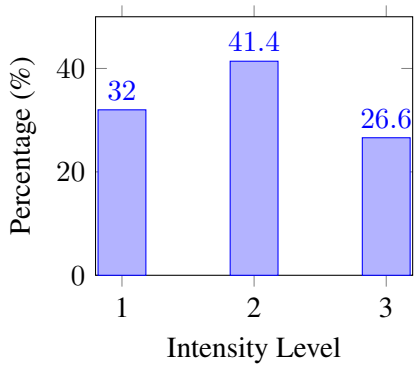

Figure 7: Intensity-wise percentage distribution of contradictions present in synthetic data.

tive) in order for them to be considered contradictory.

To decide whether a contradiction is present, ask the following question:

> "Would it be very unlikely that both of these statements could be true at the same time?"

If the answer is yes, treat the pair as a contradiction. Differences in emphasis, suggestions for improvement, or unrelated critiques should not be marked as contradictions.

Step 4: Write the contradiction statement. For each contradiction, write a **contradiction statement** explaining why the two reviews disagree. The statement should clearly describe the shared aspect and summarize the opposing viewpoints in a neutral manner, without copying text from the reviews.

Step 5: Assign contradiction severity. Assign a severity score to each contradiction using the rubric below.

- **Score 1 — Low Severity (Implicit Contradiction):** One statement is generic while the other is specific; the conflict is indirect or interpretative; the disagreement is weak or implicit; there is no strong positive or negative polarity. Eg.

  > "review1": "Section 3, where the authors describe the proposed techniques is somewhat confusing to read, because of a lack of detailed mathematical explanations of the proposed techniques."
  >
  > "review2": "The paper is clearly written and the results seem compelling."

- **Score 2 — Moderate Severity (Explicit but Mild Conflict):** Both statements refer to the same aspect; one statement is mildly critical while the other is moderately or strongly positive; the disagreement is clear but not extreme.

  > "review1": "However, the novelty is limited in the sense it is application of coordinate descent on power iterations."
  >
  > "review2": "This paper appears to be the first to solve this problem, and make a connection to coordinate decent."

- **Score 3 — High Severity (Direct Strong Contradiction):** One statement is strongly positive while the other is strongly negative; the judgments are highly polarized; the conflict is clear, direct, and fundamental.

  > "The paper is clearly written."
  >
  > "I found the presentation of the proposed measure overly confusing."

You must base the intensity judgment on the strength and polarity of the evaluative language rather than the length or technical details of the comments.

Additional Notes. A contradiction may still exist when one review provides a general evaluation and the other comments on a specific part of the paper. Focus only on the given aspect category when comparing comments, and do not introduce personal opinions or external knowledge during annotation.

D.2 Annotator Training and Annotation Process

Given the technical complexity of scientific peer reviews and their frequent use of domain-specific language, we recruited six doctoral students (4th–5th year) in Machine Learning as annotators, each with several years of experience in conducting research, publishing in peer-reviewed venues, and participating in the peer-review process. To initialize the annotation process, two domain experts with more than ten years of experience in scientific publishing independently annotated an initial set of 75 review pairs. The experts subsequently met to discuss discrepancies and reconcile their annotations, resulting in a set of expert-validated reference annotations. These reference annotations were used to support a multi-phase annotator training procedure. In the first phase, each annotator independently annotated 25 review pairs. Their annotations were then reviewed jointly with the experts to identify systematic errors in contradiction evidence selection, intensity interpretation, and guideline adherence. Based on this analysis, we refined the annotation guidelines and provided targeted feedback to the annotators. Two additional training phases followed, each involving newly sampled review pairs, enabling annotators to iteratively improve their understanding of both evidence-level contradiction identification and graded intensity assignment. From the third training round onward, most annotators exhibited stable proficiency, correctly identifying expert-aligned contradiction evidence spans and assigning matching intensity labels for at least 70% of expert-annotated contradiction instances, where correctness requires agreement with experts on both evidence selection and intensity assignment.

Throughout the full annotation process, we continuously monitored the annotated data to detect inconsistencies, ambiguous cases, and sources of systematic confusion. We employed an iterative feedback mechanism in which annotators received regular guidance informed by expert review and error analysis. In cases of disagreement or uncertainty, domain experts adjudicated and provided final decisions, ensuring consistency and coherence across the dataset. Following the completion of annotation, we assessed inter-annotator agreement separately for the two core components of the task. Agreement on contradiction evidence identification was evaluated based on semantic alignment with expert-validated evidence spans, while agreement on contradiction intensity labels was measured using Cohen's kappa (Cohen, 1960). We obtained an average kappa score of 0.64 for intensity annotation, indicating substantial agreement given the fine-grained, evidence-level nature of the task. Overall, these results demonstrate the effectiveness of the training protocol and annotation design, while reflecting the inherent difficulty of jointly annotating explicit contradiction evidence and the graded intensity of disagreement in scientific peer reviews.

D.3 Subjectivity and Borderline Cases in Intensity Annotation

Contradiction intensity annotation is inherently subjective, particularly when reviewer comments involve hedging language (e.g., "seems unclear," "might be incremental") or when reviewers weigh trade-offs differently. Although we define a formal scoring rubric (Appendix **??**), some cases remain difficult to categorize consistently across annotators.

We emphasize that RevCI annotates contradiction evidence pairing and graded contradiction intensity, while aspect categories and aspect spans are inherited from the ASAP annotations and were not re-annotated in this work. Agreement was measured on whether a span constitutes a contradiction and its corresponding intensity label.

We measure agreement using Cohen's κ over the full 4-class label space $\{0, 1, 2, 3\}$, yielding an overall $\kappa = 0.64$, indicating substantial agreement given the fine-grained, evidence-level nature of the task. To provide label-level granularity, we additionally compute one-vs-rest κ for each class: $\kappa_0 = 0.81$, $\kappa_1 = 0.47$, $\kappa_2 = 0.63$, and $\kappa_3 = 0.72$.

As expected for an ordinal severity scale, agreement is highest for null (0) and strong (3) contradictions and lowest for mild contradictions (1), which often involve hedged or scalar evaluative language. Most residual disagreements occur between adjacent levels ($1 \leftrightarrow 2$, $2 \leftrightarrow 3$), with rare extreme confusions ($1 \leftrightarrow 3$). Consistent with this ordinal structure, weighted $\kappa \approx 0.70$, indicating that disagreements primarily reflect calibration differences rather than categorical inconsistencies.

We present representative borderline cases below where contradiction intensity was difficult to define.

Case 1: Apparent doubt vs. non-contradiction. A few borderline cases emerged during annotation where apparent doubt between reviews could be misinterpreted as contradiction when no true contradiction exists. For instance, Reviewer 1 stated:

> "In the image-matching experiment, is it possible to add results for an LSSVM or other baseline besides [9]?"

while Reviewer 2 wrote:

> "Experiments show that the proposed method is as accurate but $> 10\times$ faster than traditional large-margin learning techniques on synthetic data and an image alignment problem."

Although both comments concern experimental evaluation, the first reviewer merely requests additional baselines without disputing the reported results. Some annotators initially perceived this as a low-conflict contradiction because one review sounds more cautious while the other is strongly positive. However, based on our annotation definition, the pair was removed since opposing doubt cannot be counted as a contradiction.

Case 2: Distinguishing between Scores 0 and 1. More challenging distinctions appeared between Scores 0 and 1. In one case, Reviewer 1 wrote:

> "The evaluation is a good start with comparing several base DA methods with and without the proposed TransferNorm architecture,"

while Reviewer 2 stated:

> "The experiments are extensively evaluated both qualitatively and quantitatively, demonstrating the effectiveness of the proposed TranNorm."

Here, both reviewers explicitly judge evaluation thoroughness, and the polarity difference is subtle. Some annotators initially leaned toward Score 0 because both claims appear positive. However, the first reviewer's phrasing ("good start") is hedged and implies that further work is needed rather than expressing a strongly positive judgment. Consequently, the pair was labeled Score 1.

Case 3: Distinguishing between Scores 1 and 2. A few distinctions also appeared between Scores 1 and 2. In one case, Reviewer 1 noted:

> "I suspect that most of the information is stored in the memory and only a small change of the training data is allowed... the high Inception Score cannot show the generalization ability as well..."

raising explicit concerns about generalization and the adequacy of the reported metric. Reviewer 2, however, stated:

> "Overall, I think MemoryGAN opened a new direction of GAN and is worth further exploration."

Here, the reviewers express opposing evaluative stances toward the method's quality and promise. Annotators initially considered Score 1 because the second comment is relatively generic. However, the positive judgment directly contrasts with the first reviewer's technical skepticism regarding soundness. Consequently, the pair was labeled Score 2, reflecting an explicit but moderate disagreement.

D.4 Human Annotation Protocol

We first filtered review pairs to match the annotators' domain expertise and prior reviewing experience, and then randomly sampled 800 pairs from this pool for annotation. For each selected pair, annotators read both reviews in full, identify explicit evidence spans from each review that express contradictory judgments, and write a contradiction explanation (CE) explaining the disagreement. Contradiction intensity is annotated at the evidence level, allowing a single review pair to contain multiple contradictions with varying strengths. Annotators then assign a contradiction intensity score in $\{1, 2, 3\}$ to each evidence pair using a 3-point Likert scale, where **1** denotes *low severity* (weak or implicit contradictions), **2** denotes *moderate severity* (explicit but mild contradictions), and **3** denotes *high severity* (direct and strong contradictions). The rubric and detailed annotation guidelines are described in Section D.1. Requiring explicit evidence spans together with contradiction explanation ensures the faithfulness of the annotations and encourages careful reasoning about both the source and severity of disagreement, resulting in higher-quality annotations.

D.5 Annotator Compensation

Annotators were compensated at an hourly rate consistent with standard doctoral research assistant wages, in accordance with institutional guidelines. Compensation was based on time spent on annotation rather than on a per-instance basis, reflecting the substantial variability in review length, technical complexity, and cognitive effort required to identify contradiction evidence, assign graded intensity labels, and produce faithful explanations. To support annotation quality and reduce fatigue, we imposed a maximum daily annotation limit of six hours per annotator. This constraint ensured that annotators had sufficient time to carefully read full reviews, reason about conflicting judgments, and apply annotation guidelines consistently. On average, annotating one review pair took approximately 30–40 minutes, depending on review length and technical complexity. All compensation practices complied with institutional policies and applicable ethical standards.

E Evaluation Metrics

Contradiction evidence extraction yields unordered sets of instances with variable cardinality and variable-length evidence, making count-based evaluation inadequate. For example, a ground-truth evidence span may be *"The authors claim a significant improvement in accuracy, but the reported results are not statistically significant,"* while a predicted evidence may be *"the reported results are not statistically significant."* In such cases, exact-match evaluation fails, and token-level F1 is heavily penalized due to missing tokens, whereas ROUGE-L assigns a high score by capturing the longest common subsequence.

Given the set of ground-truth contradictions G and predicted contradictions P, we enforce a one-to-one alignment by computing the maximum-weight matching using the Hungarian algorithm as implemented in `scipy.optimize.linear_sum_assignment` (Virtanen et al., 2020) at the sentence level, and averaging the aligned ROUGE-L scores. This prevents a single prediction from being aligned with multiple ground-truth instances (or vice versa) and assigns equal weight to both sentences in each evidence pair. When $|G| \neq |P|$, global matching may result in weak alignments. To mitigate this, we discard aligned pairs whose ROUGE-L score falls below a threshold λ_{match}[10]. Discarded pairs, along with other unmatched instances, are treated as unmatched and contribute to evidence-level false negatives (FN) and false positives (FP).

Finally, to evaluate the model's ability to predict evidence intensity, we compute agreement metrics over the matched evidence pairs. Specifically, we report Cohen's Kappa (κ) (Cohen, 1960), Spearman's rank correlation (ρ) (Spearman, 2010), and Kendall's Tau (τ) (Kendall, 1938).

F Error Analysis

We perform qualitative evaluation to understand where our proposed model fails.

Ambiguity in Evaluative Language: A common source of error arises from vague or scalar evaluative terms, such as "good start," "extensive," or "reasonable." In several cases, differences in emphasis for example, one reviewer describing the evaluation as "a good start" while suggesting additional comparisons and another characterizing the experiments as "extensively evaluated" do not constitute genuine contradictions but rather reflect differing thresholds or expectations. However, the framework sometimes interprets such complementary or gradational assessments as contradictory, leading to false positives.

Aspect Confusion: A recurring source of error stems from confusion in aspect identification, particularly when a single sentence references multiple aspects simultaneously (e.g., novelty, evaluation, and clarity). In such cases, the model frequently misclassifies the primary aspect of disagreement by anchoring on salient lexical cues rather than the reviewer's central evaluative intent. As a result, contradictions are often assigned to an incorrect aspect (e.g., Soundness instead of Originality).

G LLM Prompts for Evaluation

Prompt For Evaluating Baseline LLM Models

Carefully analyze the two peer reviews provided. Your objective is to identify contradictions between the reviews based on the defined contradiction criteria and within the specified evaluation aspects.

[10] We set the value of $\lambda_{\text{match}} = 0.3$ empirically.

Contradiction Definition: {contradiction_definition}
Follow the steps below:
Analyze the two peer reviews carefully.
Identify contradictions between them in accordance with the provided contradiction definition and the specified aspects.
Support each identified contradiction with explicit statements from both reviews.
Assign each contradiction to exactly one of the given aspects.
Ensure that no two contradictions rely on the same pair of evidence sentences.
Assign an intensity score for each contradiction based on the provided scoring metric.
{Aspect names}: {Aspect Definition}
{Scoring Metric}:
{Example outputs}:
{Output Format}:
{Key Rules of Contradiction Extraction}:

This appendix lists the exact prompts used to evaluate all baseline LLMs and IMPACT agents under a unified protocol.

ACEA PROMPT

CONTRADICTION DEFINITION: A contradiction occurs when two statements make claims that cannot both be true simultaneously. Contradictions can be implicit as well as explicit.
TASK: Analyze these two peer reviews and identify both implicit as well as explicit contradictions specifically related to the "{aspect_name}" aspect.
ASPECT FOCUS: {aspect_name} Description: {aspect_description}
REVIEW 1: {review_1} **REVIEW 2**: {review_2}
INSTRUCTIONS: 1. Look for statements in both reviews that relate to " {aspect_name}" 2. Identify if these statements contradict each other 3. If contradictions exist, extract them with exact evidence from both reviews 4. Extract as many contradictions you can find related to "{aspect_name}" 5. It is NOT necessary that you find contradictions for this Aspect; if none exist, return an empty list
Output format: {{ "aspect": "{aspect_name}", "contradictions": ({{ "contradiction": "brief description of contradiction", "evidence": ("exact quote from Review 1", "exact quote from Review 2") }}) }}
RULES: - Focus ONLY on the "aspect_name" aspect - Evidence array must have exactly 2 elements: [Review_1_quote, Review_2_quote] - Extract the exact complete sentences from each review that illustrate the contradiction - If no contradictions found for this aspect, return empty contradictions array - Output ONLY valid JSON, no additional text

DIA (Initial Scoring)

You are an intensity scorer for peer-review POTENTIAL contradictions.
Scoring Rubric: - Score 0 — No Contradiction (Compatible or Orthogonal Statements) Statements refer to different aspects, topics, or evaluation criteria, OR Statements discuss the same aspect but are fully compatible or consistent, OR Any differences are descriptive, complementary, or additive rather than conflicting. EXAMPLES:- {2 examples}
- Score 1 — Low Severity (Implicit Contradiction) One statement is generic, the other is specific, OR The conflict is indirect or interpretative, No strong positive/negative polarity, Weak or implicit disagreement. EXAMPLES:- {2 examples}
- Score 2 — Moderate Severity (Explicit but Mild Conflict) Both statements explicitly refer to the same aspect, One gives light criticism, the other is mildly or significantly positive, Explicit but not extreme polarity. EXAMPLES:- {2 examples}
- Score 3 — High Severity (Direct Strong Contradiction) Strongly worded positive vs. negative evaluation of the same aspect, Extremely polarized opposite judgments, Clear and fundamental extreme disagreement. EXAMPLES:- {2 examples}
REVIEW 1 CONTEXT: {review_1}
REVIEW 2 CONTEXT: {review_2}
EVIDENCE FROM REVIEW 1: {s1}
EVIDENCE FROM REVIEW 2: {s2}

Analyze these two pieces of evidence and determine the intensity of contradiction.
Output format: {{ "intensity": <0, 1, 2, or 3>, "reasoning": "detailed explanation of why you assigned this intensity score" }}

DIA (Debate Prompt)

You are an intensity scorer for peer-review POTENTIAL contradictions.
Scoring Rubric: - Score 0 — No Contradiction (Compatible or Orthogonal Statements) Statements refer to different aspects, topics, or evaluation criteria, OR Statements discuss the same aspect but are fully compatible or consistent, OR Any differences are descriptive, complementary, or additive rather than conflicting. EXAMPLES:- {2 examples}
- Score 1 — Low Severity (Implicit Contradiction) One statement is generic, the other is specific, OR The conflict is indirect or interpretative, No strong positive/negative polarity, Weak or implicit disagreement. EXAMPLES:- {2 examples}
- Score 2 — Moderate Severity (Explicit but Mild Conflict) Both statements explicitly refer to the same aspect, One gives light criticism, the other is mildly or significantly positive, Explicit but not extreme polarity. EXAMPLES:- {2 examples}
- Score 3 — High Severity (Direct Strong Contradiction) Strongly worded positive vs. negative evaluation of the same aspect, Extremely polarized opposite judgments, Clear and fundamental extreme disagreement. EXAMPLES:- {2 examples}
REVIEW 1 CONTEXT: {review_1}
REVIEW 2 CONTEXT: {review_2}
EVIDENCE FROM REVIEW 1: {s1}
EVIDENCE FROM REVIEW 2: {s2}
DEBATE HISTORY: {debate_context}
YOUR ASSIGNED SCORE: {my_score}
YOUR PREVIOUS REASONING: {my_reasoning}
OPPONENT'S SCORE: {opponent_score} **OPPONENT'S REASONING**: {opponent_reasoning}
CRITICAL INSTRUCTION: You MUST defend your score of {my_score}. You CANNOT change your score during the debate.
RULES FOR EVIDENCE-BASED DEBATE: 1. ONLY cite text that actually appears in the evidence or reviews above 2. Use direct quotes (in "quotation marks") when referencing the reviews 3. Point to specific words, phrases, or sentences that support your position 4. Counter your opponent by showing what they missed or misinterpreted in the actual text 5. Use simple language - avoid vague terms like "nuanced", "multifaceted", "complex interplay". 6. Don't use bold word or italics formatting in your response. 7. If opponent makes claims not supported by the text, point this out specifically
Your task: 1. Quote specific phrases from the evidence that support your score of {my_score} 2. Identify flaws in opponent's reasoning by showing what the text actually says 3. Explain why your intensity level {my_score} fits the criteria better than {opponent_score}
Be assertive but fair. Focus on why YOUR intensity assessment is correct.
Output format: {{ "intensity": {my_score}, "reasoning": "your defense of score {my_score} and counterarguments against opponent's score {opponent_score}" }}

Adjudication Agent Prompt

You are a judge evaluating a debate between two intensity scorers for peer-review contradictions.
INTENSITY LEVELS: - Score 0 — No Contradiction (Compatible or Orthogonal Statements) Statements refer to different aspects, topics, or evaluation criteria, OR Statements discuss the same aspect but are fully compatible or consistent, OR Any differences are descriptive, complementary, or additive rather than conflicting.
- Score 1 — Low Severity (Implicit Contradiction) One statement is generic, the other is specific, OR The conflict is indirect or interpretative, No strong positive/negative polarity, Weak or implicit disagreement.

- Score 2 — Moderate Severity (Explicit but Mild Conflict) Both statements explicitly refer to the same aspect, One gives light criticism, the other is mildly or significantly positive, Explicit but not extreme polarity.
- Score 3 — High Severity (Direct Strong Contradiction) Strongly worded positive vs. negative evaluation of the same aspect, Extremely polarized opposite judgments, Clear and fundamental extreme disagreement.
TASK: 1. Review the entire debate conversation 2. Examine the evidence from both reviews 3. Make a final judgment on the appropriate intensity score based on the reviews and the agents' debate. 4. Provide clear reasoning for your decision.
Consider: - Which arguments were most convincing? - What does the evidence actually show? - Does the contradiction meet the criteria for the claimed intensity level? - Your decision must be based solely on the evidence and debate provided.
REVIEW 1 CONTEXT: {review_1}
REVIEW 2 CONTEXT: {review_2}
EVIDENCE FROM REVIEW 1: {s1}
EVIDENCE FROM REVIEW 2: {s2}
COMPLETE DEBATE HISTORY: {debate_context}
Based on the evidence and the debate, make your final judgment on the intensity score.
Output format: {{ "intensity": <0, 1, 2, or 3>, "reasoning": "your final judgment explaining why this is the correct intensity score" }}

TIDE (System Prompt)

You are an expert at analyzing peer reviews of scientific papers and identifying contradictions between them and assign an intensity score to each contradiction. You are a contradiction extractor. Your task is not to generate contradictions, but to identify and extract ONLY the contradictions(ONLY if they exist) that are explicitly present between the two reviews provided.

TIDE (User Prompt)

TASK:
You are given TWO peer reviews of the same scientific paper. Your task is to identify where the two reviewers are contradicting. Therefore you need to EXTRACT the contradiction sentence pairs between the two reviews across multiple aspects and then assign an intensity score(1 to 3) to each contradiction evidence pair based on the severity scoring rubric provided below with a suitable reasoning for the assigned intensity score.
CONTRADICTION DEFINITION:
A contradiction occurs when two statements make claims about the same aspect but opposing sentiment that cannot both be true simultaneously. Negations or disagreements alone do not constitute contradictions unless they pertain to the same factual claim.
ASPECTS TO CONSIDER:
You must only consider contradictions that fall under one of the following aspects: 1. Substance (Insufficient experiments, weak analysis, missing ablations, lack of depth) 2. Motivation (Problem importance, relevance, impact, significance of the research) 3. Clarity (Writing quality, organization, readability, explanation of methods) 4. Meaningful Comparison (Fairness and completeness of comparisons to prior or baseline work) 5. Originality (Novelty of ideas, techniques, insights, or contributions) 6. Soundness (Correctness, validity, methodological justification, logical consistency)
SEVERITY SCORING RUBRIC:
- Score 1 — Low Severity (Implicit Contradiction) - One statement is generic, the other is specific - Conflict is indirect or interpretative - Weak or implicit disagreement - No strong positive or negative polarity
- Score 2 — Moderate Severity (Explicit but Mild Conflict) - Both statements refer to the same aspect - One is mildly critical, the other is moderately or strongly positive - Clear disagreement but not extreme
- Score 3 — High Severity (Direct Strong Contradiction) - Strongly positive vs strongly negative evaluation - Extremely polarized opposite judgments - Clear, fun-

damental, and direct conflict
RULES: - Only use the listed aspects. - Evidence must contain EXACTLY two sentences(one from review A and another one from review B) in the format: [sentence_from_Review_1, sentence_from_Review_2] - The above two sentences must directly contradict each other on the SAME aspect. They must be claims that can't be true simultaneously. - If multiple contradictions exist for the same aspect, include them all as separate entries. - Use verbatim sentences from the reviews (no paraphrasing). - Do NOT generate evidence that is not explicitly present in the reviews.(word by word evidence extraction) - Do NOT include aspects with no contradictions. - If there are no contradictions, return an empty array. - Keep the scoring rubric open in mind while assigning intensity scores.
REQUIRED OUTPUT FORMAT: Return a array where each element is an object with these fields: - "evidence": An array with exactly two strings [Review_A_sentence, Review_B_sentence] OR "evidence": [] if no contradiction is found. - "intensity_reasoning": A detailed reasoning explaining why the assigned intensity score was chosen for the contradiction evidence pair. - "intensity": An integer score (1 to 3) representing the severity of the contradiction based on the scoring rubric provided.
Following are two sample examples of contradictions from different review pairs(DONT COPY THESE EXAMPLES IN YOUR OUTPUT, THESE ARE ONLY FOR REFERENCE):
Steps to follow: 1. Read Both reviews carefully. 2. Identify contradictions between the two reviews following the contradiction definition. 3. Read the rules and validate the contradictions using the EXACT sentence pairs from both reviews while following the rules mentioned above. 4. Read the Scoring Rubric carefully. 5. Think and write the reasoning behind the intensity of contradiction according to the scoring rubric provided. 4. Assign an intensity score(1 to 3) to each contradiction evidence pair based on the reasoning you wrote and the severity scoring rubric provided. 6. Finally, return the contradictions in the REQUIRED OUTPUT FORMAT mentioned below or return empty array if no contradictions are found.
Paper ID: {paper_id}
INPUT REVIEWS:
Review A: {review_a}
Review B: {review_b}
The contradiction evidences(sentence pairs) between Review A and Review B are:

H Case Studies

H.1 Case Study 1: DIA Agreement on Initial Severity

This example illustrates a scenario in which both DIA agents independently assign the same severity score to a contradiction without requiring further deliberation.

DIA Agreement on Initial Severity Score

Evidence Statements:-

Review 1: The approach is novel and very interesting.

Review 2: All in all, the originality of the paper is lacking, the experimental setup is not convincing, and there are not much insights given by the paper into the novelty of method.

Aspect: Originality

Agent 1 (Score = 3)

The two pieces of evidence directly contradict each other on the aspect of the paper's originality. Review 1 states that the approach is novel and very interesting, while Review 2 claims that the originality of the paper is lacking. This constitutes a strong and direct contradiction, warranting a Score 3 (High Severity).

Agent 2 (Score = 3)

The two statements present a direct and fundamental contradiction regarding the core merit of the paper. Review 1 explicitly affirms novelty, whereas Review 2 explicitly denies originality and insight. This diametrically opposed evaluation of the same aspect justifies a high-severity contradiction.

The two reviews make explicitly opposing claims about the paper's originality: one affirms novelty, while the other directly denies it. Because both statements evaluate the same core aspect using strong and unambiguous language, the contradiction is immediately identifiable as high severity. As shown in the agent responses, both agents converge on a severity score of 3, demonstrating that clear, symmetric contradictions can be reliably detected and consistently scored by independent agents in the absence of ambiguity.

H.2 Case Study 2: DIA Disagreement and the Role of Discussion

This case illustrates how explicit multi-round discussion is necessary when agents disagree on the severity of a contradiction. Although both agents identify a clear conflict, they differ in interpreting the intensity of opposing evaluative language.

DIA Disagreement on Initial Severity Score

Evidence Statements

Review 1: However, the ablation study and analysis on the model is weak and the authors only show experimental observations.

Review 2: The evaluation is thorough across the board, as far as I can tell.

Aspect: Substance (Experimental Evaluation)

Round 1: Initial DIA Assessments

Agent 1 (Score = 2)

The reviews explicitly disagree on the quality of the experimental evaluation. Review 1 describes the analysis as weak, while Review 2 describes it as thorough. The contradiction is clear, but the language is not maximally polarized, suggesting a moderate-severity conflict.

Agent 2 (Score = 3)

The statements represent fundamentally opposing evaluations of the same aspect. Calling the analysis weak and observational directly conflicts with describing the evaluation as thorough across the board, indicating a high-severity contradiction.

Rounds 2–4: Discussion and Reassessment

Agent 1 (Maintains Score = 2)

Agent 1 argues that the wording in Review 1 criticizes depth rather than validity, and that Review 2's assessment is hedged. These linguistic qualifiers suggest disagreement without maximal polarization, supporting a moderate-severity interpretation.

Agent 2 (Maintains Score = 3)

Agent 2 emphasizes that phrases such as "only show experimental observations" and "thorough across the board" represent absolute and incompatible judgments, arguing that the contradiction reflects a fundamental evaluation conflict.

Final Decision

Judge (Final Score = 2)

Although the contradiction is explicit and concerns the same aspect, the negative assessment targets analytical depth rather than overall validity, and the positive assessment is subjectively hedged. The disagreement is therefore clear but not maximally polarized, leading to a final moderate-severity classification.

Through repeated discussion rounds, this case highlights how surface-level polarity can mask important differences in linguistic intensity. The deliberation process allows the framework to distinguish between strongly polarized contradictions and those where interpretive flexibility remains, resulting in a more calibrated severity judgment.

H.3 Case Study 3: DIA Disagreement and the Role of Multi-Round Discussion

This case study illustrates a scenario in which the DIA agents agree on the presence of a contradiction but disagree on its severity. Multiple discussion rounds are required to surface and resolve differences in how linguistic intensity and evaluative scope are interpreted.

DIA Disagreement on Initial Severity Score

Evidence Statements

> *Review 1:* However, the ablation study and analysis on the model is weak and authors only show experimental observations.
>
> *Review 2:* The evaluation is thorough across the board, as far as I can tell.

Aspect: Substance (Experimental Evaluation)

Round 1: Initial Severity Assessment

Agent 1 (Score = 2)

The two pieces of evidence refer to the same aspect of the paper and express opposite opinions. Review 1 describes the ablation and analysis as weak, while Review 2 characterizes the evaluation as thorough. The contradiction is explicit, but the language is not extremely polarized, supporting a moderate-severity classification.

Agent 2 (Score = 3)

Review 1 explicitly criticizes the ablation study and analysis as weak and merely observational, while Review 2 claims the evaluation is thorough across the board. These statements are interpreted as fundamentally incompatible evaluations of the same aspect, indicating a high-severity contradiction.

What is gained in Round 1:
Both agents agree that the contradiction is explicit and concerns the same aspect. The disagreement is localized to the interpretation of severity rather than contradiction existence.

Round 2

Agent 1 (Locked Score = 2)

The language used is not maximally severe. The term "weak" suggests insufficient analysis rather than a fundamental flaw, and "thorough" reflects a positive but not extreme assessment. Therefore, the contradiction remains of moderate severity.

Agent 2 (Locked Score = 3)

The phrases "only show experimental observations" and "across the board" are interpreted as strong and absolute qualifiers. In the context of peer review, these represent extremely polarized judgments about evaluation quality.

What is gained in Round 2:
This round makes explicit the disagreement over whether the evaluative language should be interpreted as moderately critical or fundamentally dismissive.

Round 3

Agent 1 (Locked Score = 2)

The criticism in Review 1 targets the depth of ablation and analysis rather than declaring the entire evaluation invalid. The contradiction is clear, but not maximally polarized.

Agent 2 (Locked Score = 3)

Describing the analysis as weak and only observational undermines the scientific depth of the evaluation. When contrasted with "thorough across the board," this represents a fundamental conflict in assessment.

What is gained in Round 3:
The discussion clarifies whether the contradiction concerns partial inadequacy or a complete failure of the evaluation, which is central to severity classification.

Round 4

Agent 1 (Locked Score = 2)

The positive claim is hedged by "as far as I can tell," leaving room for missed weaknesses. Combined with the non-maximal negativity of "weak," the contradiction remains moderately severe rather than extreme.

Agent 2 (Locked Score = 3)

The absolutist phrasing "across the board" leaves no room for the weaknesses described in Review 1. This level of opposition is interpreted as maximally polarized.

What is gained in Round 4:
This round consolidates all previously surfaced cues-linguistic strength, scope of criticism, and hedging-into a stable disagreement boundary.

Final Decision

Judge (Final Score = 2)

Both reviews address the same aspect and clearly disagree, forming an explicit positive–negative contradiction. However, the negative wording critiques analytical depth rather than declaring the evaluation invalid, and the positive assessment is explicitly hedged by the phrase "as far as I can tell," indicating limited confidence and leaving room for unobserved weaknesses. As a result, the polarity is clear but not extremely polarized, supporting a moderate-severity classification.

Across four discussion rounds, the framework progressively surfaces and evaluates linguistic intensity, evaluative scope, and hedging, with each round grounding the agents' arguments more explicitly in the textual evidence. This iterative process moves the discussion from high-level explanations toward evidence-backed reasoning, enabling the judge to integrate all relevant signals and arrive at a calibrated final severity judgment that neither initial assessment alone could conclusively justify.