SCHOLARPEER : A Multi-Agent Framework for Automated Peer Review

Palash Goyal[1] **Mihir Parmar**[1] **Yiwen Song**[1] **Hamid Palangi**[1] **Tomas Pfister**[1] **Jinsung Yoon**[1]

[1]Google

{palashgoyal, mihirparmar}@google.com

Abstract

The exponential growth of machine learning submissions has strained the traditional peer review process, resulting in slow feedback loops for authors and an immense burden on reviewers to rigorously audit technical soundness and verify literature. To address this, we introduce SCHOLARPEER , a multi-agent framework designed to operationalize the rigorous auditing workflow of a senior researcher. Rather than attempting to replace human judgment, SCHOLARPEER serves as a co-scientist: acting as a mentor for rapid author iteration prior to submission, and as an active verification assistant that augments human reviewers. The framework structurally decouples contextualization from critique by deploying a sub-domain historian to synthesize the field's trajectory, a baseline scout to proactively hunt for omitted state-of-the-art comparisons, and a multi-aspect Q&A engine that deeply audits technical soundness—scrutinizing internal logical consistency, experimental validity, and mathematical rigor—while cross-referencing claims against top-tier academic venues. We comprehensively evaluate SCHOLARPEER on $\sim$1,800 ICLR submissions spanning 2020 through 2025. Our results show that SCHOLARPEER achieves significant win-rates against state-of-the-art fine-tuned models and search-augmented agentic baselines.

1 Introduction

The democratization of artificial intelligence research has precipitated an unprecedented explosion in scientific output. Major machine learning conferences now receive tens of thousands of submissions annually, straining the peer review process to its breaking point. This scalability crisis manifests in two distinct operational bottlenecks: authors wait months to receive actionable feedback on their manuscripts, while area chairs and reviewers face extreme fatigue attempting to rigorously audit technical soundness, manually hunt down missing baselines, and verify state-of-the-art claims in a rapidly evolving literature space [Zhuang et al., 2025, Liang et al., 2024].

Consequently, the community has turned to Large Language Models (LLMs) to assist in the evaluation of scientific manuscripts [Luo et al., 2025]. Early approaches primarily focused on fine-tuning models on static datasets of paper-review pairs [Liu and Shah, 2023, Weng et al., 2024]. While these systems excel at linguistic fluency and summarization, they evaluate papers in a parametric vacuum. To mitigate this, recent frameworks have integrated retrieval mechanisms [Zhu et al., 2025, Zeng et al., 2025, Chitale et al., 2025] or explored multi-agent review architectures [Jin et al., 2024, Yamada et al., 2025]. However, these systems predominantly focus on optimizing text-generation policies or simulating collaborative discussions. They still lack the structural mechanisms necessary to conduct the rigorous auditing required to produce a publication-ready critique.

To address this gap, we introduce SCHOLARPEER , a multi-agent framework designed to operationalize the *rigorous auditing* workflow of a senior researcher. Crucially, SCHOLARPEER is not designed

40th Conference on Neural Information Processing Systems (NeurIPS 2026).

to replace the human peer review system; rather, it serves as a co-scientist to augment it. For authors, SCHOLARPEER acts as a mentor, providing highly specific, actionable feedback that enables rapid iteration prior to submission. For reviewers and ACs, it serves as a verification assistant, auditing experimental validity, actively retrieving missing baselines, and contextualizing novelty to supply human experts with verifiable evidence, freeing them to focus on high-level scientific judgment.

Our framework structurally decouples paper comprehension from external contextualization. SCHOLARPEER employs three key agents: a *sub-domain historian*, a *baseline scout*, and a *multi-aspect Q&A engine*. The historian agent dynamically synthesizes a "domain narrative," placing the submission in the context of the field's trajectory. The baseline scout proactively hunts for missing SOTA baselines and datasets that the authors failed to compare against. Finally, the Q&A engine operates as a skeptic, deeply auditing technical soundness—scrutinizing internal logical consistency, experimental validity, and mathematical rigor—while cross-referencing claims against top-tier academic venues.

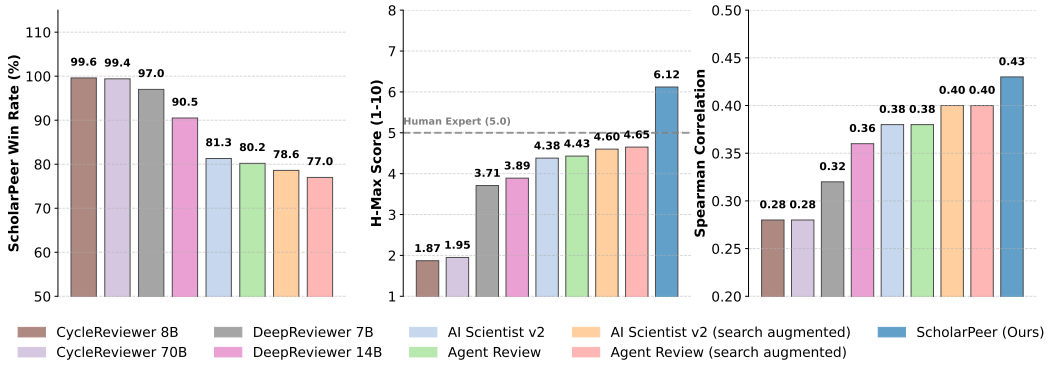


Figure 1: Comparative evaluation of SCHOLARPEER against existing frameworks on ScholarEval. (Left) Win rate of SCHOLARPEER against review fine-tuned models and agentic baselines. (Middle) Average H-Max score (higher the better) of reviews generated by various frameworks (best human review is considered as 5). (Right) Spearman correlation of scores generated by review frameworks with ground-truth human rankings. We use Gemini 3.1 Pro as the backbone model for SCHOLARPEER and baseline agentic frameworks. We use Claude Sonnet 4.5 as the LLM-judge. These results show that our proposed framework consistently outperforms baselines across various metrics.

We comprehensively evaluate SCHOLARPEER on ∼1,800 submissions from ICLR 2020 through 2025, merging the evaluation sets of DeepReview-Bench [Zhu et al., 2025] and AgentReview [Jin et al., 2024]. We call this data set ScholarEval. Recognizing that traditional metrics often fail to capture the nuance of scientific critique, we also explore two new evaluation metrics: the *H-Max score*, which calibrates the localized quality of critiques against human experts, and the *review diversity score*, which assesses the variance of perspectives provided by the system.

Our main contributions are as follows:

- We propose SCHOLARPEER , a multi-agent framework that performs active, search-grounded interrogation, designed to accelerate author iteration and augment human reviewer judgment.
- We introduce specialized *historian*, *baseline scout*, and *Q&A* agents that collectively perform both internal logical consistency checks and external web-scale verification against top-tier academic venues.
- We conduct comprehensive evaluations, achieving dominant win-rates against state-of-the-art fine-tuned models and search-enabled agentic baselines, with practical deployment metrics (∼$1.20 and latency of ∼10 minutes per review).

2 Problem Formulation

We formulate the task of automated peer review assistance as generating a rigorous, verifiable critique R given a paper submission S. To capture the cognitive process of a human expert, we formally distinguish between the intrinsic claims and experimental setup of the paper, $S_{content}$, and

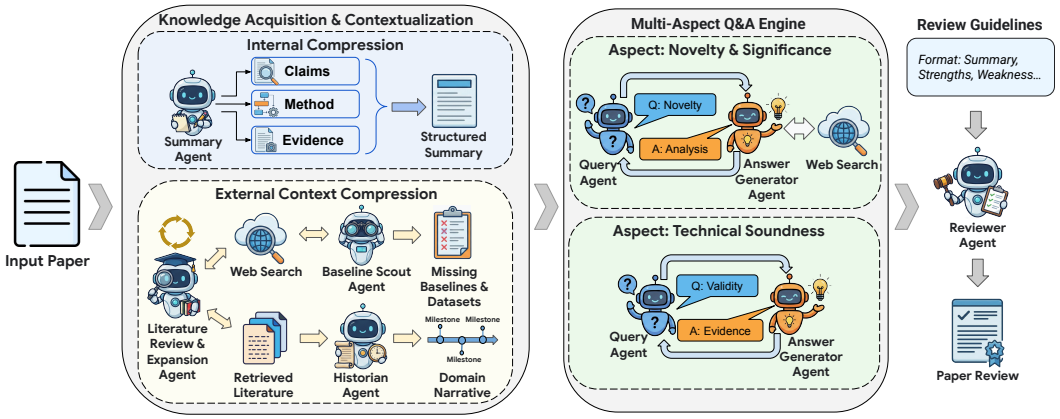


Figure 2: The SCHOLARPEER framework: Given an input paper, the framework employs a dual-stream information retrieval process. The *knowledge acquisition and contextualization module* uses summarizer, search-enabled literature review, historian and baseline scout agents to compress internal and external information. These inputs feed into the *multi-aspect Q&A engine*, which generates and answers probing questions regarding the novelty and technical soundness. Finally, the *review generator* utilizes these inputs and conference-specific review guidelines to generate the final review.

the extrinsic, ever-evolving context of the scientific field, $\mathcal{C}_{dynamic}$. A human reviewer does not evaluate $S_{content}$ in isolation; rather, they adopt a skeptical stance, stress-testing the methodology against a dynamic mental graph of prior art, concurrent work, and established community standards. Consequently, a rigorous automated review must actively interrogate $S_{content}$ while leveraging $\mathcal{C}_{dynamic}$, formulated as the mapping:

$$R = f(S_{content}, \mathcal{C}_{dynamic}). \tag{1}$$

We design SCHOLARPEER to execute this mapping through a multi-agent orchestration that dynamically constructs $\mathcal{C}_{dynamic}$, structurally transforms $S_{content}$, and actively interrogates the submission. Ultimately, this framework produces a strong, verifiable peer review designed to augment human judgment and accelerate author iteration.

3 SCHOLARPEER

In this section, we present the architecture of SCHOLARPEER , an agentic framework for automated peer review. As illustrated in Figure 2, SCHOLARPEER operates via a dual-stream process designed to emulate the information-gathering and auditing flow of a human expert. Rather than processing the manuscript in a single pass—which often leads to the "lost-in-the-middle" phenomenon [Liu et al., 2024] [1] and surface-level summaries—the framework is organized into two primary subsystems: the Knowledge Acquisition & Contextualization Module (§3.1) and the Multi-Aspect Q&A Engine (§3.2).

3.1 Knowledge Acquisition and Contextualization

The knowledge acquisition module structurally transforms $S_{content}$ and dynamically constructs $\mathcal{C}_{dynamic}$ using four specialized agents.

Summary Agent (Internal Extraction). A significant bottleneck in applying LLMs to full-text papers is the "lost-in-the-middle" phenomenon and the cognitive overload associated with processing dense technical tokens simultaneously. To mitigate cognitive overload, the summary agent transforms the raw text of the submission into a structured representation, $\hat{S}$, extracting: (1) the core claim set $\mathcal{H}_{core}$, (2) the proposed method M, and (3) the reported evidence E. By decoupling structural

[1] The "lost-in-the-middle" phenomenon describes the tendency for language models to perform better when relevant information is at the beginning or end of a long context, rather than in the middle. Performance significantly degrades when crucial data is buried within the prompt, a positional bias that persists even in models designed for long contexts.

comprehension from critique, this guides the downstream agents to operate on a high-fidelity signal of the paper's exact premises.

Literature Review & Expansion Agent (Dynamic Context Creation). This agent constructs a "live" reference frame, enabling the system to validate claims against the latest literature. It constructs $\mathcal{C}_{dynamic}$ by executing a two-step retrieval process. First, it identifies the paper's sub-domain based on the abstract and performs an initial literature search using a search engine. Second, it iteratively identifies gaps in the literature search and performs an "expansion search" targeting recent pre-prints and concurrent work. Crucially, to ensure academic rigor and prevent low-quality source pollution, source quality filters strictly restrict search results to top-tier conferences (e.g., NeurIPS, ICLR, ICML, CVPR, ACL) and arXiv, enforcing a strict publication cutoff date matching the paper's submission. By relying exclusively on the Search API to retrieve actual, verifiable literature rather than generating URLs from parametric memory, this agent structurally minimizes hallucination.

Sub-Domain Historian Agent (External Contextualization). This agent augments $\mathcal{C}_{dynamic}$ with domain trajectory. Retrieving raw abstracts is insufficient to assess significance. The historian agent compresses the retrieved literature into a chronological "domain narrative", identifying the arc of progress in the sub-domain. This narrative mimics a senior researcher's mental model, enabling the system to evaluate the significance of the paper's contributions.

Baseline Scout Agent (Adversarial Auditing). This agent augments $\mathcal{C}_{dynamic}$ with independent baseline scouting. It acts as an adversarial auditor, independently searching for the current SOTA methods on the specific benchmarks used in the paper. It returns a concrete list of omitted competitors and datasets, providing the actionable evidence required for a rigorous technical soundness critique.

3.2 Multi-Aspect Q&A Engine and Review Generation

The Multi-Aspect Q&A Engine utilizes the structured and dynamic context to rigorously probe the paper. Finally, the review generator agent produces the final review using these signals.

Multi-Aspect Q&A Agents (Adversarial Probing). We deploy dedicated query and answer generation agents for two critical aspects: *novelty & significance* and *technical soundness*. The query agents act as "skeptics" to actively probe the paper, generating questions using the structured summary, the domain narrative, and the baseline scout's findings. The novelty & significance query agent enumerates the distinct contribution claims and produces one question per claim. The technical soundness query agent generates questions to scrutinize the soundness of paper's critical aspects. To answer the novelty & significance questions, the answer agent is provided with active search access. Conversely, the technical soundness answer agent cross-references the extracted claims against the paper's own methodology and evidence sets to find logical inconsistencies, impossible mathematical constraints, or inappropriate evaluation metrics (e.g., catching discretization errors in continuous treatment functions).

Review Generator Agent (Guidelines-Driven Synthesis) The review generator agent synthesizes the final report by integrating the structured paper summary $\hat{S}$ and the verified Q&A pairs from the interrogation log. This agent is conditioned on explicit *review guidelines* (e.g., the NeurIPS reviewer checklist). By decoupling the "investigation" (Q&A Engine) from the "reporting" (Review Generator), SCHOLARPEER effectively aggregates findings, resolves any minor query overlaps, and outputs a highly constructive, venue-aligned critique ready for author iteration or AC augmentation.

4 Experiments

4.1 Evaluation Protocols

Evaluating automated critiques is challenging, as standard NLP metrics fail to capture scientific depth, and LLM judges can exhibit verbosity or family-model biases. Since our objective is to accelerate author iteration and augment human reviewers, our evaluation protocols prioritize the intrinsic quality, actionability, and depth of the generated critiques over predicting a paper's final acceptance score. To ensure rigorous, multi-faceted assessment, we triangulate our results using side-by-side evaluations, two novel metrics (H-Max and review diversity), and decision alignment, validated against an expert human baseline. We provide the complete prompts used for evaluation in Appendix H.

Side-by-Side (SxS) Evaluation. The search-enabled judge is provided with the input paper and reviews from two anonymized frameworks (with order randomized to avoid position bias). The judge compares the reviews across five dimensions: Technical Accuracy, Constructive Value, Analytical Depth, Significance Assessment, and Overall Judgment. These dimensions directly measure the framework's utility for authors and reviewers, strictly penalizing hallucinations and generic advice.

H-Max Score. To measure how well an AI assistant actually augments the human review process, we introduce the H-Max score. An expert judge (search-enabled) evaluates a single AI review for a given paper against the *collective set* of human reviews for that paper.

The judge identifies the strongest points made by any human reviewer on a specific aspect (e.g., Technical Accuracy) and sets that as the "Expert Baseline". The AI review is then scored relative to this baseline. An H-Max score of 5 signifies that the AI review is of similar quality to the strongest points made by a set of k reviewers. A score of 10 signifies the AI review is transformative compared to the collective set of human reviews, and a score of 1 means the AI review misses critical points. Appendix F.3 contains the complete mapping from score to its interpretation.

Review Diversity Score (RDS). A known systemic risk of LLM automation is the "artificial hivemind" Jiang et al. [2025] where models converge on homogenized opinions. To quantify perspective variance, we generate $N = 3$ reviews for the same paper, compute embeddings using an embedding model (we used DistilRoberta-v1 [Sanh et al., 2019] for our experiments), and calculate the Inter-review Semantic Similarity (IR_{sim}) in embedding space:

$$IR_{sim} = \frac{1}{N(N-1)} \sum_{i \neq j} \text{CosineSim}(E(r_i), E(r_j)) \tag{2}$$

where r_i and r_j denote distinct generated reviews for the same paper and $E(\cdot)$ represents the sentence-embedding encoder. We define the Review Diversity Score as RDS $= 1 - IR_{sim}$. A higher RDS indicates the framework explores varied, rigorous perspectives rather than collapsing to a mean response.

Decision Score Alignment. While our primary goal is generating actionable critiques rather than predicting outcomes, prior automated review systems heavily rely on decision prediction. For completeness and comparability, we evaluate the alignment of the model's quantitative recommendations with human ground truth by reporting the Spearman Correlation (ρ) between the model's predicted final decision scores and the actual human rankings.

4.2 Experimental Setup

Dataset. To rigorously evaluate our framework across a longitudinal time horizon, we curate ScholarEval. This dataset merges the test split of DeepReview-Bench [Zhu et al., 2025], covering ICLR submissions from 2024 and 2025, with the evaluation set from AgentReview [Jin et al., 2024], covering ICLR submissions from 2020 to 2023, resulting in a comprehensive benchmark of 1,786 machine learning submissions and their corresponding human reviews. By spanning a half-decade of research—including oral, spotlight, poster, and rejected papers—ScholarEval provides a highly diverse, real-world testbed to assess the temporal robustness of automated critiques.

Baselines. We categorize our baselines into three distinct groups to isolate the impact of architecture, training data, and retrieval capabilities: (1) Fine-tuned Baselines: We compare against CycleReviewer 8B and 70B [Weng et al., 2024] and DeepReviewer 7B and 14B [Zhu et al., 2025]. These models represent the state-of-the-art in supervised fine-tuning on review data. Notably, DeepReviewer natively utilizes OpenScholar for dynamic literature search, providing a strong retrieval-enabled benchmark. (2) Agentic Baselines: We compare against single agent baselines and multi-agent baselines of AgentReview [Jin et al., 2024] and AI Scientist v2 [Yamada et al., 2025], instantiated with various state-of-the-art backbone models including Gemini 3.1 Flash, Gemini 3.1 Pro and Claude Sonnet 4.5. (3) Search-Augmented Agentic Baselines: To address potential information asymmetry and prove that performance gains stem from our adversarial workflow, we equip the agentic baselines with the exact same Google Search tool used by ScholarPeer and constrained them to identical search expansion call limits (e.g., maximum of 3 expansion rounds) for fair comparison. In addition, we evaluate against Stanford Agent Reviewer (SAR)[2]. Note that SAR is not open-source and has only browser based access. To evaluate SAR, we sampled 50 papers from ScholarEval and got the reviews.

[2] `https://paperreview.ai/`

Table 1: Direct comparison of SCHOLARPEER (Gemini 3.1 Pro) with baselines across five critical dimensions on ScholarEval. "Win" indicates that the judge assessed SCHOLARPEER as superior (average over three runs). Cells highlighted in green denote the superior outcome of our framework (red indicates that the baseline is superior). SCHOLARPEER achieves dominant win-rates in *Significance Assessment* and *Constructive Value*, validating the impact of the multi-agent reasoning architecture. We use Claude Sonnet 4.5 as the LLM judge. We provide the standard deviation in bracket.

Category	Baselines	Technical Accuracy		Constructive Value		Analytical Depth		Significance Assessment		Overall Judgment	
		Win(%)↑	Lose(%)	Win(%)↑	Lose(%)	Win(%)↑	Lose(%)	Win(%)↑	Lose(%)	Win(%)↑	Lose(%)
Fine-tuned	CycleReviewer-8B	**98.6 (0.1)**	1.0 (0.1)	**100.0 (0.0)**	0.0 (0.0)	**100.0 (0.0)**	0.0 (0.0)	**98.4 (0.2)**	0.3 (0.1)	**99.6 (0.1)**	0.3 (0.1)
	CycleReviewer-70B	**98.5 (0.2)**	1.2 (0.1)	**99.8 (0.1)**	0.2 (0.1)	**99.8 (0.1)**	0.0 (0.0)	**98.5 (0.3)**	0.5 (0.1)	**99.4 (0.1)**	0.4 (0.1)
	DeepReviewer-7B	**92.1 (0.3)**	5.4 (0.2)	**97.2 (0.2)**	2.2 (0.2)	**97.2 (0.3)**	1.6 (0.2)	**95.6 (0.3)**	1.9 (0.2)	**97.0 (0.3)**	2.4 (0.3)
	DeepReviewer-14B	**77.3 (0.6)**	12.7 (0.6)	**88.3 (0.5)**	9.5 (0.5)	**85.6 (0.5)**	11.9 (0.4)	**89.1 (0.4)**	5.7 (0.3)	**90.5 (0.2)**	8.8 (0.2)
Single Agent	Claude Sonnet 4.5	**42.9 (0.9)**	26.4 (0.7)	**71.2 (0.8)**	23.3 (0.6)	**51.4 (0.9)**	38.5 (0.8)	**77.3 (0.7)**	13.8 (0.5)	**74.1 (0.8)**	24.5 (0.7)
	Gemini 3.1 Flash	**39.5 (1.1)**	32.4 (0.9)	**81.6 (0.7)**	16.7 (0.6)	**74.9 (0.8)**	18.7 (0.7)	**77.9 (0.8)**	13.0 (0.5)	**78.3 (0.9)**	21.7 (0.8)
	Gemini 3.1 Pro	**40.2 (1.0)**	33.0 (0.8)	**76.9 (0.9)**	18.1 (0.7)	**69.1 (1.1)**	22.5 (0.8)	**74.9 (0.9)**	11.8 (0.6)	**73.0 (1.0)**	26.1 (0.9)
Single Agent w/ Search	Claude Sonnet 4.5	**41.7 (0.8)**	27.4 (0.7)	**71.1 (0.9)**	23.3 (0.7)	**50.8 (1.0)**	39.5 (0.9)	**75.2 (0.8)**	15.7 (0.6)	**72.1 (0.9)**	26.4 (0.8)
	Gemini 3.1 Flash	**36.5 (1.2)**	35.4 (1.0)	**78.6 (0.8)**	19.7 (0.7)	**71.9 (0.9)**	21.7 (0.8)	**74.9 (0.9)**	16.0 (0.7)	**75.3 (1.0)**	24.7 (0.9)
	Gemini 3.1 Pro	**37.2 (1.1)**	36.0 (0.9)	**73.9 (1.0)**	21.1 (0.8)	**66.1 (1.2)**	25.5 (0.9)	**71.9 (1.0)**	14.8 (0.7)	**70.0 (1.1)**	29.1 (1.0)
Multi Agent	Agent Review (Claude 4.5)	**35.1 (1.2)**	29.1 (1.1)	**60.4 (1.1)**	35.5 (1.0)	32.4 (1.2)	**55.6 (1.2)**	**66.2 (1.1)**	19.1 (0.8)	**56.8 (1.0)**	42.5 (1.1)
	Agent Review (Gemini Flash)	**47.8 (1.2)**	14.7 (1.0)	**86.2 (0.9)**	11.4 (0.8)	**79.4 (1.0)**	13.6 (0.9)	**75.5 (0.9)**	8.4 (0.6)	**83.2 (0.8)**	15.1 (0.7)
	Agent Review (Gemini Pro)	**46.3 (1.0)**	27.1 (0.8)	**83.9 (0.7)**	12.6 (0.6)	**80.7 (0.8)**	12.8 (0.7)	**79.2 (0.8)**	10.0 (0.6)	**80.2 (0.8)**	17.9 (0.6)
	AI Scientist v2 (Claude 4.5)	**37.2 (1.1)**	26.2 (0.9)	**70.4 (1.0)**	23.6 (0.8)	44.7 (1.2)	**46.8 (1.0)**	**70.9 (0.9)**	19.4 (0.7)	**64.7 (0.9)**	34.6 (0.8)
	AI Scientist v2 (Gemini Flash)	**48.1 (1.1)**	24.2 (0.9)	**86.0 (0.9)**	10.4 (0.8)	**76.7 (1.1)**	13.8 (0.9)	**74.2 (1.0)**	11.5 (0.8)	**81.1 (0.8)**	17.0 (0.7)
	AI Scientist v2 (Gemini Pro)	**49.1 (0.9)**	25.9 (0.7)	**82.8 (0.6)**	14.1 (0.5)	**82.0 (0.7)**	11.5 (0.6)	**78.5 (0.7)**	10.9 (0.6)	**81.3 (0.7)**	18.1 (0.6)
	Stanford Agent Reviewer*	**36.0 (1.4)**	32.0 (1.2)	**52.0 (1.3)**	46.0 (1.1)	38.0 (1.4)	**46.0 (1.3)**	**60.0 (1.2)**	34.0 (1.0)	**54.0 (1.1)**	46.0 (1.0)
Multi Agent w/ Search	Agent Review (Claude 4.5)	**30.5 (1.4)**	36.0 (1.2)	**56.4 (1.3)**	41.6 (1.1)	29.5 (1.5)	**58.8 (1.3)**	**66.1 (1.2)**	19.2 (0.9)	**54.9 (1.1)**	44.6 (1.2)
	Agent Review (Gemini Flash)	**44.7 (1.3)**	20.8 (1.1)	**84.1 (1.0)**	13.3 (0.9)	**74.6 (1.1)**	19.9 (1.0)	**72.4 (1.0)**	16.4 (0.7)	**79.5 (0.9)**	20.0 (0.8)
	Agent Review (Gemini Pro)	**42.3 (1.1)**	31.2 (0.9)	**80.6 (0.8)**	16.0 (0.7)	**77.1 (0.9)**	16.4 (0.8)	**76.0 (0.9)**	13.1 (0.6)	**77.0 (0.9)**	21.9 (0.7)
	AI Scientist v2 (Claude 4.5)	**34.2 (1.2)**	29.1 (1.0)	**66.8 (1.1)**	29.6 (0.9)	41.7 (1.3)	**49.7 (1.1)**	**69.9 (1.0)**	20.1 (0.8)	**62.8 (1.0)**	37.2 (0.9)
	AI Scientist v2 (Gemini Flash)	**45.3 (1.2)**	17.2 (1.0)	**83.0 (1.0)**	13.4 (0.9)	**70.7 (1.2)**	18.8 (1.0)	**68.2 (1.1)**	18.5 (0.9)	**76.0 (0.9)**	20.0 (0.8)
	AI Scientist v2 (Gemini Pro)	**46.2 (1.0)**	28.8 (0.8)	**80.8 (0.7)**	16.2 (0.6)	**79.3 (0.8)**	14.4 (0.7)	**75.0 (0.8)**	13.7 (0.7)	**78.6 (0.8)**	21.4 (0.7)

*Stanford Agent Reviewer is not open-source and has only browser based access; evaluation was performed on 50 papers sampled from ScholarEval.

Implementation. We use Gemini 3.1 Pro as the backbone LLM for all the agents in SCHOLARPEER for the main results and use Claude 4.5 Sonnet for the LLM judge. We further show results on backbone independence by swapping our backbone and the LLM judge. For web search, we use Google search and set publication cutoff date as the boundary for literature search. All automated metrics are averaged over three independent runs, with standard deviations reported to ensure statistical rigor. We provide further experimental details in Appendix C and the agent prompts in Appendix G.

LLM Judge Validation. Prior to large-scale benchmarking, we validated the alignment of our Claude 4.5 Sonnet judge against human experts using a 100-paper subset of ScholarEval. For the H-Max score, the LLM judge achieved a Pearson correlation of 0.53 with human evaluators. This alignment is substantial when contextualized against the inter-human expert correlation of 0.58 on the same set. For side-by-side (SxS) comparisons, the inter-human reviewer agreement was 88% (Cohen's Kappa = 0.76), while the LLM judge achieved an 81% agreement rate with human preferences. The guidelines provided to the expert researchers for performing the evaluation are detailed in Appendix F. Further, a common issue of LLM-as-a-judge evaluations is a bias toward longer outputs. To ensure our win-rates reflect true critique quality rather than verbosity, we analyzed review lengths. SCHOLARPEER reviews averaged 597 words ($\sigma = 61$). In contrast, human reviews averaged 788 words ($\sigma = 295$), and DeepReviewer-14B averaged 2,156 words ($\sigma = 312$). Despite generating significantly more concise reviews ($\sim$3.6$\times$ shorter than the strongest fine-tuned baseline and shorter than human experts), SCHOLARPEER consistently won SxS evaluations, showing that verbosity bias is not a concern for this task.

4.3 Side-by-Side Results

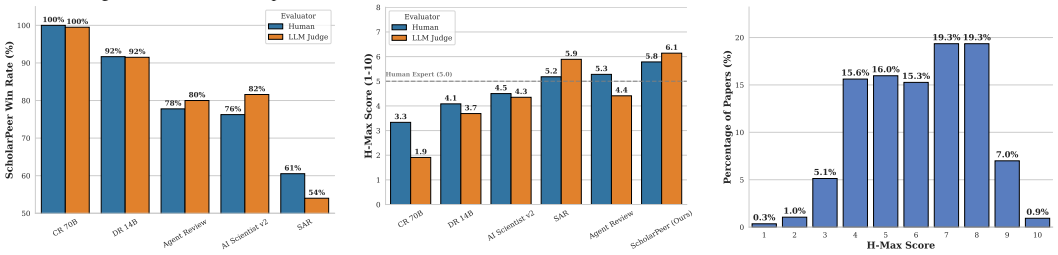
Table 1 illustrates the win rate of our framework against various baselines. SCHOLARPEER achieves a dominant 90.5% win-rate against the strongest fine-tuned baseline, DeepReviewer-14B. Against the strongest agentic baseline with the same backbone model, we achieve a win rate of 73%. Even against the strongest closed source multi-agent baseline, Stanford Agent Reviewer, we achieve a 54% win-rate. Despite Claude Sonnet 4.5 being the judge, it scores ScholarPeer consistently higher than Claude based agentic baselines. This illustrates that our framework clearly outperforms the baselines.

Addressing Information Asymmetry. The single and multi-agent baselines do not the leverage search tool. To further isolate our multi-agent architecture from information asymmetry, we evaluated SCHOLARPEER against *search-augmented* variants of our agentic baselines. Even when equipped with the exact same Google Search tool, SCHOLARPEER outperforms AI Scientist v2 and AgentRe-

Table 2: H-Max score, human correlation, and review diversity score on ScholarEval. H-Max score shows comparison against best human expert reviews. Human corr (ρ) denotes Spearman correlation with human rankings. Review diversity measures semantic variance. SCHOLARPEER achieves state-of-the-art performance across all metrics. We provide the standard deviation over three runs in parentheses for the H-Max scores.

Category	Model	H-Max Score Across Dimensions (1-10)↑					Human	Review
		Tech. Acc.	Cons. Val.	Analytical	Signif.	Overall	Corr (ρ)↑	Diversity↑
Fine-tuned	CycleReviewer-8B	1.98 (0.01)	2.17 (0.02)	1.99 (0.01)	2.31 (0.02)	1.87 (0.01)	0.28	0.01
	CycleReviewer-70B	2.14 (0.02)	2.31 (0.01)	2.08 (0.02)	2.64 (0.01)	1.95 (0.02)	0.28	0.01
	DeepReviewer-7B	3.44 (0.01)	3.55 (0.02)	3.70 (0.01)	3.65 (0.02)	3.71 (0.01)	0.32	0.02
	DeepReviewer-14B	3.86 (0.02)	3.98 (0.01)	3.93 (0.02)	3.80 (0.01)	3.89 (0.02)	0.36	0.02
Single Agent	Claude Sonnet 4.5	4.69 (0.02)	4.41 (0.01)	4.63 (0.02)	4.40 (0.01)	4.52 (0.02)	0.31	0.23
	Gemini 3.1 Flash	4.58 (0.01)	4.17 (0.02)	4.46 (0.01)	3.89 (0.02)	4.34 (0.01)	0.33	0.21
	Gemini 3.1 Pro	4.86 (0.02)	4.33 (0.01)	4.53 (0.02)	3.95 (0.01)	4.50 (0.02)	0.37	0.23
Single Agent w/ Search	Claude Sonnet 4.5	4.82 (0.02)	4.55 (0.01)	4.75 (0.02)	4.90 (0.01)	4.75 (0.01)	0.31	0.23
	Gemini 3.1 Flash	4.80 (0.01)	4.35 (0.02)	4.60 (0.01)	4.45 (0.02)	4.55 (0.02)	0.36	0.25
	Gemini 3.1 Pro	4.95 (0.02)	4.50 (0.01)	4.65 (0.02)	4.55 (0.01)	4.65 (0.02)	0.39	0.25
Multi Agent	Agent Review (Claude 4.5)	4.59 (0.01)	4.50 (0.02)	4.72 (0.01)	4.40 (0.02)	4.58 (0.01)	0.38	0.22
	Agent Review (Gemini Flash)	4.41 (0.02)	4.14 (0.01)	4.24 (0.02)	3.91 (0.01)	4.19 (0.02)	0.36	0.21
	Agent Review (Gemini Pro)	4.77 (0.01)	4.32 (0.02)	4.40 (0.01)	4.05 (0.02)	4.43 (0.01)	0.38	0.22
	AI Scientist v2 (Claude 4.5)	4.66 (0.02)	4.48 (0.01)	4.70 (0.02)	4.53 (0.01)	4.60 (0.02)	0.37	0.22
	AI Scientist v2 (Gemini Flash)	4.71 (0.01)	4.33 (0.02)	4.48 (0.01)	4.16 (0.02)	4.36 (0.01)	0.33	0.20
	AI Scientist v2 (Gemini Pro)	4.63 (0.02)	4.29 (0.01)	4.35 (0.02)	4.03 (0.01)	4.38 (0.02)	0.38	0.22
Multi Agent w/ Search	Agent Review (Claude 4.5)	4.75 (0.01)	4.65 (0.02)	4.85 (0.01)	4.95 (0.02)	4.80 (0.01)	0.39	0.22
	Agent Review (Gemini Flash)	4.55 (0.02)	4.35 (0.01)	4.45 (0.02)	4.40 (0.01)	4.45 (0.02)	0.38	0.24
	Agent Review (Gemini Pro)	4.90 (0.01)	4.50 (0.02)	4.60 (0.01)	4.60 (0.02)	4.65 (0.01)	0.40	0.23
	AI Scientist v2 (Claude 4.5)	4.90 (0.02)	4.65 (0.01)	4.85 (0.02)	5.10 (0.01)	4.85 (0.02)	0.38	0.22
	AI Scientist v2 (Gemini Flash)	4.85 (0.01)	4.60 (0.02)	4.65 (0.01)	4.60 (0.02)	4.90 (0.01)	0.36	0.23
	AI Scientist v2 (Gemini Pro)	4.80 (0.02)	4.45 (0.01)	4.55 (0.02)	4.45 (0.01)	4.60 (0.02)	0.39	0.23
	Stanford Agent Reviewer	5.50 (0.03)	5.39 (0.02)	**6.22 (0.03)**	5.56 (0.02)	5.89 (0.02)	-*	0.26
Ours	SCHOLARPEER	**5.76 (0.01)**	**5.87 (0.02)**	5.85 (0.02)	**6.47 (0.01)**	**6.12 (0.01)**	**0.43**	**0.31**

*Stanford Agent Reviewer does not provide a score.



(a) Side-by-Side (SxS) comparison of human vs. LLM judge win rates.

(b) H-Max score comparison of human evaluator vs LLM judge.

(c) Distribution of overall scores for the SCHOLARPEER framework.

Figure 3: Alignment between LLM and human judges. We observe strong agreement in (a) side-by-side (SxS) win rates and (b) H-Max score. Panel (c) illustrates that $> 25\%$ of the reviews generated by SCHOLARPEER achieve a score of 8 or higher.

view with search access by significant margins. This shows that the performance gains stem from the agentic architecture rather than search tool addition.

Backbone Independence. To confirm that our architecture—rather than the Gemini 3.1 Pro LLM—drives performance, we conducted a reverse-ablation using Claude 4.5 Sonnet as the backbone for SCHOLARPEER and the agentic baselines and using Gemini 3.1 Pro as the LLM judge. We achieve 90.2% win rate against DeepReviewer-14B, 79.6% against AI Scientist v2 and 78.8% against AgentReview, confirming the framework's efficacy is model-agnostic.

Human Evaluation. To complement our automated metrics, blinded expert reviewers compared SCHOLARPEER against five baselines (20 papers per baseline). Human experts consistently preferred SCHOLARPEER, with trends strongly correlating with the LLM judge, further validating our automated SxS protocol (see Figure 3a).

4.4 H-Max Results

Table 2 presents the H-Max scores. While state-of-the-art fine-tuned models struggle to surpass a score of 4.0, SCHOLARPEER achieves an exceptional average H-Max score of 6.12. This score is driven by the dimensions most valuable for human augmentation. SCHOLARPEER scores an

Vs. DeepReviewer-14B (Fine-tuned Baseline)

✅ **Advantage: SOTA Verification & Integrity**
SCHOLARPEER consistently refutes novelty claims by citing specific prior work (e.g., Lightman et al., TaskPrompter) that DeepReviewer ignores. It also detects fatal flaws like data leakage (training on validation sets) and circular evaluation protocols.
✅ **Advantage: Factual Precision**
SCHOLARPEER avoids "hallucinated critiques." While DeepReviewer frequently claims parameters are missing when they are explicitly stated, SCHOLARPEER correctly locates them and focuses on substantive inconsistencies.

Vs. AI Scientist v2 (Agentic Baseline)

✅ **Advantage: Theoretical Contextualization**
SCHOLARPEER demonstrates deeper domain mastery by connecting papers to broader phenomena (e.g., "Gradient Masking" in adversarial defense or "Grokking"), whereas AI Scientist often remains surface-level.
✅ **Advantage: Actionable Constructiveness**
Instead of generic advice ("add complexity analysis"), SCHOLARPEER prescribes exact remedies, naming specific missing datasets (e.g., SR25, GSO) or control experiments required for validation.
❌ **Disadvantage: Internal Sanity Checking**
AI Scientist v2 is occasionally better at catching internal inconsistencies (e.g., mathematically impossible probability ratios) because SCHOLARPEER prioritizes external verification over internal logic checks.

Figure 4: Qualitative analysis of SCHOLARPEER vs. top baselines. We summarize the key comparative advantages and disadvantages derived from the reasoning traces of the LLM Judge. SCHOLARPEER dominates in external verification (SOTA checking) and contextual depth, while AI Scientist v2 remains competitive on internal consistency checks.

6.47 in *Significance Assessment* (contextualizing novelty) and 5.87 in *Constructive Value* (providing exact, actionable remedies). Furthermore, SCHOLARPEER exhibits remarkable stability; across three independent runs, the overall H-Max score yielded an exceptionally low standard deviation of $\sigma = 0.01$. Further, Figure 3c shows the distribution of score. We observe a significant $> 25\%$ of generated reviews getting a score of 8.0 or above.

Human Evaluation. Figure 3b shows that SCHOLARPEER achieves the highest average rating across both human evaluation and LLM judge.

4.5 Review Diversity

As shown in Table 2, fine-tuned models suffer from severe mode collapse, exhibiting extremely low review diversity (RDS ≈ 0.01–0.02) and homogenizing scientific feedback. Single and multi-agent baselines improve this (up to 0.25). SCHOLARPEER achieves the highest review diversity among automated systems (RDS = 0.31), significantly reducing the gap to the natural variance of human committees (RDS = 0.43). Further, given the low variance in the win rates, these diverse reviews are of high quality. We hypothesize that SCHOLARPEER compounds diversity through its multi-agent design: multiple runs lead to slightly different, valid literature expansion paths and Q&A probes, simulating the varied expertise of a diverse human reviewer panel.

To verify that this embedding diversity corresponds to distinct conceptual insights rather than mere wording variations, we conducted a paraphrase ablation. For each paper in the benchmark, we prompted an LLM to rewrite the SCHOLARPEER review three times using entirely different vocabulary and sentence structures, while strictly preserving the underlying scientific critiques. The RDS on these purely stylistic variants plummeted to 0.114 ($\sigma = 0.027$). In contrast, SCHOLARPEER's actual multi-run RDS is 0.31, empirically proving that our metric captures genuinely distinct conceptual critiques (e.g., exploring alternative literature paths) rather than surface-level rephrasing.

4.6 Summary of Gains

We aggregated the reasoning traces from our LLM judge to distill the systematic advantages and disadvantages of SCHOLARPEER compared to key baselines. We show the results in Figure 4. We provide review examples and summary of advantages against other baselines in Appendix E and the summarization prompt in Appendix H.

4.7 Ablations & Hyperparameter Sensitivity

Component Ablation. We quantify the contribution of each module in Table 3. Removing the baseline scout agent causes a 9% drop in win-rate, as the system fails to identify missing comparisons.

Table 3: Component ablations (SxS win rate against AI Scientist v2 (Gemini 3.1 Pro) on 200 ScholarEval papers).

Configuration	Win Rate (%)
Full SCHOLARPEER	**85**
w/o Literature Review	79 (↓ 6%)
w/o Historian	81 (↓ 4%)
w/o Baseline Scout	76 (↓ 9%)
w/o Q&A	59 (↓ 26%)
w/o Summary	75 (↓ 10%)

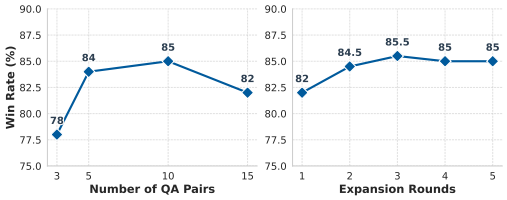


Figure 5: Impact of search rounds (k) on review quality. Diminishing returns occur after $k = 3$ as tangential papers dilute the context window.

Removing the Q&A agent causes the largest drop (26%), crippling the system's ability to perform deep verification. The performance drop removing summarizer agent is surprisingly significant (10%) signifying that internal compression also plays a vital role in reviewing.

Hyperparameters. We investigate the sensitivity of the system to the number of Q&A pairs generated and the number of literature expansion rounds. As shown in Figure 5, performance gains saturate after 10 Q&A pairs and 3 search rounds, suggesting an optimal operating point.

Format and Tone Bias Ablations. To ensure the judge rewards scientific substance over presentation format or harsh prompt tone, we conducted two ablations (results in Table 4). In the *Format Ablation*, we stripped all structural appeal (bolding, lists, headings) and converted each review into a dense paragraph. In the *Neutral Prompt Ablation*, we removed wording encouraging strict evaluation. Despite lacking structural formatting, SCHOLARPEER maintained a high H-Max of 6.04 ($\sigma = 0.01$). The neutrally prompted system achieved an H-Max of 6.11 ($\sigma = 0.01$). Dominant SxS win rates across both ablations prove performance is driven by substantive critique content, not stylistic or prompt-induced bias.

4.8 Deployment Practicality & Robustness

Deployment Practicality. A critical consideration for multi-agent systems is operational overhead. SCHOLARPEER requires approximately 20 LLM calls per review, processing ∼429K input and ∼26K output tokens to actively retrieve and verify literature (see Appendix D). Using current enterprise API pricing, this equates to a monetary cost of ∼$1.20 per paper. Single paper review latency is ∼10 min (reduced to amortized latency of ∼56 seconds per review with batch size 20). At the scale of major conferences (e.g., 10,000 submissions), this computational investment is highly justifiable given the immense reduction in manual verification hours for Area Chairs and reviewers.

Baseline Scout Ground Truth Validation. To validate the reliability of the baseline scout agent, we conducted a manual audit on a subset of 50 papers, where the agent generated 266 citations. All the citations were real (0% hallucination), correctly attributed temporally, missing from the manuscript, and highly relevant. For recall, human reviewers collectively identified 288 missing baselines, of which the scout agent captured 38 (13.2% recall). Contextually, only 43 of the 288 baselines were flagged by more than one human (14.9% inter-human agreement). Thus, the scout achieves a recall rate nearly identical to agreement between independent human experts.

Data Contamination Check. To rule out reliance on memorized public reviews from pre-training data, we evaluated on a post-cutoff dataset of 200 ICLR 2026 submissions (submission deadline August 2025, strictly post-dating all model cutoffs), mirroring the conference acceptance rate (54 accepted, 146 rejected). As shown in Table 5, the performance hierarchy remained consistent with ScholarEval. SCHOLARPEER maintained dominant win rates (92.8% vs. DeepReviewer-14B) and secured a high H-Max of 5.96 on completely unseen manuscripts, confirming capabilities stem from active reasoning and live retrieval rather than artifact memorization.

5 Related Work

LLM-based Automated Reviewing. The field has rapidly evolved from text classification to complex generation tasks [Luo et al., 2025]. Early approaches utilized standard pre-trained models for decision prediction and summarization [Liu and Shah, 2023, Liang et al., 2024]. Subsequent works shifted to supervised fine-tuning (SFT) on peer review datasets. Weng et al. [2024] introduced *CycleReviewer*

Table 4: SxS Win Rates (%) of SCHOLARPEER under Format and Neutral Prompt Ablations.

Baseline	Format Abl.	Neutral Abl.
CycleReviewer-8B	97.1	99.6
CycleReviewer-70B	94.3	99.3
DeepReviewer-7B	91.6	96.8
DeepReviewer-14B	87.2	90.2
Agent Review (Search)	76.7	76.8
AI Scientist v2 (Search)	77.0	78.5

Table 5: Performance on the post-cutoff ICLR 2026 dataset (200 unseen papers).

Framework	H-Max	Win Rate (%)
CycleReviewer-8B	1.70 (0.05)	99.8
CycleReviewer-70B	1.88 (0.05)	99.7
DeepReviewer-7B	3.51 (0.06)	98.4
DeepReviewer-14B	3.66 (0.06)	92.8
Agent Review (Search)	4.54 (0.05)	71.2
AI Scientist v2 (Search)	4.52 (0.05)	74.6
SCHOLARPEER	**5.96 (0.04)**	-

using negative constraints, while *OpenReviewer* [Idahl and Ahmadi, 2025] and *ReviewAgents* [Gao et al., 2025b] fine-tuned models to capture venue-specific stylistic norms. Most relevant to our work is *DeepReviewer* [Zhu et al., 2025], which employs chain-of-thought fine-tuning to improve reasoning.

Agentic & Retrieval-Augmented Frameworks. To overcome static limitations, the field has pivoted toward agentic workflows [Wu et al., 2024, Bran et al., 2023]. Jin et al. [2024] pioneered this with *Agent Review*, modeling the review process as a multi-stage discussion. Building on this collaborative approach, *AgentRxiv* [Schmidgall and Moor, 2025] proposed a framework for autonomous research that facilitates interaction between various agents and human researchers throughout the scientific lifecycle. Other recent agentic reviewers such as *ReviewRL* [Zeng et al., 2025] and *REMOR* [Taechoyotin and Acuna, 2025] integrate reinforcement learning to align reviews with human preferences; REMOR utilizes GRPO with a multi-aspect reward function. *AutoRev* [Chitale et al., 2025] models papers as hierarchical graphs to optimize internal information retrieval, while *PaperQA2* [Skarlinski et al., 2024] uses agentic RAG on scientific Q&A.

Foundations of Automated Survey & Verification. Our architecture is grounded in the emerging literature of automated science. The historian agent builds upon automated survey generation frameworks, such as *SurveyG* [Nguye et al., 2025] and *AutoSurvey* [Wang et al., 2024], which construct hierarchical citation graphs using LLMs. Similarly, the baseline scout operationalizes the "chain-of-verification" pattern [Dhuliawala et al., 2024]. This aligns with scientific fact-checking benchmarks like *SciFact* and *Co-Sight* [Zhang et al., 2025], moving beyond passive retrieval to actively identify omitted evidence and missing SOTA baselines.

Evaluation of AI Reviewers. Evaluating automated critiques remains challenging. Garg et al. [2025] proposed *ReviewEval* to assess actionability, and *MMReview* [Gao et al., 2025a] introduced a multimodal benchmark. Standard metrics, however, often fail to capture systemic automation risks. *LLM-REVal* [Li et al., 2025] highlighted LLM judge bias, while Jiang et al. [2025] identified the risk of an "artificial hivemind" homogenizing opinions. Our work complements these efforts by prioritizing human-calibrated actionability via the H-Max score and explicitly measuring homogenization risks through the review diversity score.

6 Conclusion

In this paper, we introduced SCHOLARPEER, a multi-agent framework that structurally decouples intrinsic paper comprehension from external contextualization, and operationalizes the rigorous workflow of a senior researcher. Our specialized agents dynamically synthesize domain trajectories, proactively hunt for omitted baselines, and deeply interrogate technical soundness. Crucially, SCHOLARPEER is designed not to replace human judgment, but to serve as a "co-scientist" that accelerates author iteration and augments human reviewers with highly actionable, verifiable evidence.

Our extensive evaluation on $\sim$1,800 ICLR submissions spanning 2020 through 2025 demonstrates that SCHOLARPEER sets a new standard for automated critique. By achieving an H-Max score of 6.12 and decisively outperforming both state-of-the-art fine-tuned models and search-augmented agentic baselines, we empirically prove that structured adversarial reasoning—rather than mere internet access—is the essential catalyst for expert-level review. Future work will explore collaborative multi-agent debate mechanisms to further deepen the rigor, diversity, and actionability of artificial scientific critique.

References

A. M. Bran, S. Cox, O. Schilter, C. Baldassari, A. D. White, and P. Schwaller. Chemcrow: Augmenting large-language models with chemistry tools. *arXiv preprint arXiv:2304.05376*, 2023.

M. P. Chitale, K. M. Shetye, H. Gupta, M. Chaudhary, and V. Varma. Autorev: Automatic peer review system for academic research papers. *arXiv preprint arXiv:2505.14376*, 2025.

S. Dhuliawala, M. Komeili, J. Xu, R. Raileanu, X. Li, A. Celikyilmaz, and J. Weston. Chain-of-verification reduces hallucination in large language models. In *Findings of the association for computational linguistics: ACL 2024*, pages 3563–3578, 2024.

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

X. Gao, J. Ruan, Z. Zhang, J. Gao, T. Liu, and Y. Fu. Reviewagents: Bridging the gap between human and ai-generated paper reviews. *arXiv preprint arXiv:2503.08506*, 2025b.

M. K. Garg, T. Prasad, T. Singhal, C. Kirtani, M. Mandal, and D. Kumar. Revieweval: An evaluation framework for ai-generated reviews. *arXiv preprint arXiv:2502.11736*, 2025.

M. Idahl and Z. Ahmadi. Openreviewer: A specialized large language model for generating critical scientific paper reviews. In *Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (System Demonstrations)*, pages 550–562, 2025.

L. Jiang, Y. Chai, M. Li, M. Liu, R. Fok, N. Dziri, Y. Tsvetkov, M. Sap, A. Albalak, and Y. Choi. Artificial hivemind: The open-ended homogeneity of language models (and beyond). *arXiv preprint arXiv:2510.22954*, 2025.

Y. Jin, Q. Zhao, Y. Wang, H. Chen, K. Zhu, Y. Xiao, and J. Wang. Agentreview: Exploring peer review dynamics with llm agents. *arXiv preprint arXiv:2406.12708*, 2024.

R. Li, J.-C. Gu, P.-N. Kung, H. Xia, X. Kong, Z. Sui, N. Peng, et al. Llm-reval: Can we trust llm reviewers yet? *arXiv preprint arXiv:2510.12367*, 2025.

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

N. F. Liu, K. Lin, J. Hewitt, A. Paranjape, M. Bevilacqua, F. Petroni, and P. Liang. Lost in the middle: How language models use long contexts. *Transactions of the association for computational linguistics*, 12:157–173, 2024.

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

Z. Luo, Z. Yang, Z. Xu, W. Yang, and X. Du. Llm4sr: A survey on large language models for scientific research. *arXiv preprint arXiv:2501.04306*, 2025.

M.-A. Nguye, M.-D. Nguyen, K. H. Dang, N. T. Dong, D. D. Le, et al. Surveyg: A multi-agent llm framework with hierarchical citation graph for automated survey generation. *arXiv preprint arXiv:2510.07733*, 2025.

V. Sanh, L. Debut, J. Chaumond, and T. Wolf. Distilbert, a distilled version of bert: smaller, faster, cheaper and lighter. *arXiv preprint arXiv:1910.01108*, 2019.

S. Schmidgall and M. Moor. Agentrxiv: Towards collaborative autonomous research. *arXiv preprint arXiv:2503.18102*, 2025.

M. D. Skarlinski, S. Cox, J. M. Laurent, J. D. Braza, M. Hinks, M. J. Hammerling, M. Ponnapati, S. G. Rodriques, and A. D. White. Language agents achieve superhuman synthesis of scientific knowledge. *arXiv preprint arXiv:2409.13740*, 2024.

P. Taechoyotin and D. Acuna. Remor: Automated peer review generation with llm reasoning and multi-objective reinforcement learning. *arXiv preprint arXiv:2505.11718*, 2025.

Y. Wang, Q. Guo, W. Yao, H. Zhang, X. Zhang, Z. Wu, M. Zhang, X. Dai, Q. Wen, W. Ye, et al. Autosurvey: Large language models can automatically write surveys. *Advances in neural information processing systems*, 37:115119–115145, 2024.

Y. Weng, M. Zhu, G. Bao, H. Zhang, J. Wang, Y. Zhang, and L. Yang. Cycleresearcher: Improving automated research via automated review. *arXiv preprint arXiv:2411.00816*, 2024.

Q. Wu, G. Bansal, J. Zhang, Y. Wu, B. Li, E. Zhu, L. Jiang, X. Zhang, S. Zhang, J. Liu, et al. Autogen: Enabling next-gen llm applications via multi-agent conversations. In *First Conference on Language Modeling*, 2024.

Y. Yamada, R. T. Lange, C. Lu, S. Hu, C. Lu, J. Foerster, J. Clune, and D. Ha. The ai scientist-v2: Workshop-level automated scientific discovery via agentic tree search. *arXiv preprint arXiv:2504.08066*, 2025.

S. Zeng, K. Tian, K. Zhang, Y. Wang, J. Gao, R. Liu, S. Yang, J. Li, X. Long, J. Ma, et al. Reviewrl: Towards automated scientific review with rl. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pages 16942–16954, 2025.

H. Zhang, J. Lu, S. Jiang, C. Zhu, L. Xie, C. Zhong, H. Chen, Y. Zhu, Y. Du, Y. Gao, et al. Co-sight: Enhancing llm-based agents via conflict-aware meta-verification and trustworthy reasoning with structured facts. *arXiv preprint arXiv:2510.21557*, 2025.

M. Zhu, Y. Weng, L. Yang, and Y. Zhang. Deepreview: Improving llm-based paper review with human-like deep thinking process. *arXiv preprint arXiv:2503.08569*, 2025.

Z. Zhuang, J. Chen, H. Xu, Y. Jiang, and J. Lin. Large language models for automated scholarly paper review: A survey. *Information Fusion*, page 103332, 2025.

A Impact Statement

This paper introduces SCHOLARPEER, an active verification framework for automated peer review—a domain with significant potential for both positive and negative societal impact. On the positive side, our work addresses the critical scalability crisis in scientific publishing. By operationalizing the rigorous stress-testing workflow of a senior researcher, SCHOLARPEER moves beyond passive text generation to provide deep, verifiable critiques. This aims to democratize access to high-quality, actionable feedback, accelerating scientific progress particularly for authors who may lack access to senior mentorship.

However, we acknowledge several ethical risks associated with the deployment of automated evaluation systems:

- **Automation Bias:** There is a risk that human area chairs or reviewers may over-rely on AI-generated critiques, accepting them as authoritative without independent scrutiny. While SCHOLARPEER's design explicitly anchors claims in verifiable literature through its interrogation logs—avoiding the hallucinations of passive summarization—the risk of human complacency remains a critical deployment challenge.
- **Homogenization of Science:** Automated systems might penalize novel ideas that do not conform to established patterns in their training data or the retrieved literature. Although we explicitly designed the multi-agent architecture to maximize perspective variance and introduced the *Review Diversity Score* to monitor this issue, widespread reliance on automated models could inadvertently narrow the scope of acceptable scientific inquiry.
- **Privacy and Confidentiality:** The use of search-enabled agents involves processing unpublished manuscripts via external LLM and search APIs. Any real-world deployment must strictly adhere to venue confidentiality policies, ensuring that submission data is secure, ephemeral, and never used to train public models.

We emphasize that SCHOLARPEER is designed strictly to augment, not replace, human judgment. By acting as a "co-scientist"—a rigorous mentor for authors and a tireless verification assistant for reviewers—the system's output should be treated as an evidence-grounded second opinion, ensuring that final scientific decisions remain firmly in human hands.

B Limitations

While ScholarPeer represents a significant advancement, several limitations remain. First, despite improving upon baselines, the review diversity of our system (0.31) still lags behind the high variance of human expert reviews (0.43); closing this gap is crucial for capturing the full spectrum of scientific opinion. Second, the reliance on active web-scale retrieval introduces higher inference latency and cost compared to static fine-tuned models.

C Experiment Details

To ensure reproducibility, we provide the specific configuration details for the SCHOLARPEER framework and baselines.

C.1 Models and Compute

All experiments involving SCHOLARPEER were conducted using the Google Cloud Vertex AI platform.

- **Backbone Model:** We utilized `Gemini 3.1 Pro` as the core reasoning engine for all agents in the primary ScholarPeer configuration.
- **Temperature:** We set the temperature to `0.7` for all generation tasks to balance creativity with adherence to instructions.
- **Compute Infrastructure:** Experiments were orchestrated on standard CPU instances for agentic frameworks, with all heavy lifting offloaded to the Vertex AI Model-as-a-

Service (MaaS) endpoints. We used 8X NVIDIA A100 80GB GPUs for fine-tuned models (CycleReviewer and DeepReview).

- **Search Tools:** For Gemini Models, we utilized the native `google_search` tool provided by the Vertex AI Gemini API. This allows the model to perform multi-step reasoning and retrieve up-to-date information from the open web. For Claude Models, we employed the `web_search_20250305` tool to ensure a fair comparison with a similarly capable search-enabled environment.

C.2 Inter-Judge Agreement and Correlation

Table 6: Pairwise agreement between LLM judges on the SxS evaluation (overall dimension).

Judge Pair	Agreement Rate
Claude Sonnet 4.5 vs Gemini 3.1 Flash	0.79
Claude Sonnet 4.5 vs Gemini 3.1 Pro	0.77
Gemini 3.1 Flash vs Gemini 3.1 Pro	0.88

Table 6 provides the pairwise agreement between LLM judges on the SxS evaluation (overall dimension). Gemini 3.1 Flash and 3.1 Pro have the highest agreement rate. This is intuitive as they are from the same model family. Claude Sonnet 4.5 and Gemini 3.1 Pro have the lowest agreement rate but the agreement is still high (77%).

Table 7: Average correlation between the scores provided by the LLM judges for H-Max score.

Judge Pair	H-Max Score Correlation
Claude Sonnet 4.5 vs Gemini 3.1 Flash	0.62
Claude Sonnet 4.5 vs Gemini 3.1 Pro	0.66
Gemini 3.1 Flash vs Gemini 3.1 Pro	0.81

Table 7 provides the average correlation between the H-Max scores provided by the LLM judges. Similar to agreement rate, Gemini 3.1 Flash and Pro have the highest score correlation and Claude Sonnet 4.5 has lower correlation with Gemini models but the correlation is high (0.66 with Gemini 3.1 Pro).

D Computational Complexity and Deployment Analysis

In this section, we analyze the computational overhead of the SCHOLARPEER framework, covering theoretical LLM inference calls and empirical token usage, monetary cost, and deployment latency. We demonstrate that SCHOLARPEER is highly scalable and practical for conference-level deployment.

D.1 Theoretical Inference Cost

The computational cost of SCHOLARPEER is driven by its multi-agent architecture, specifically the depth of the literature search and the rigor of the Q&A verification. The total number of inference calls is defined by a fixed overhead for context construction plus variable costs associated with literature expansion rounds (k) and the number of probing questions generated (N_{QA}):

$$C_{\text{ScholarPeer}} \approx C_{\text{fixed}} + k + N_{QA} \tag{3}$$

The fixed overhead ($C_{\text{fixed}} \approx 7$) comprises single calls for the summary agent, initial literature search, historian, baseline scout, aspect-based question generation (novelty and soundness), and the final review generator. Based on our hyperparameter sensitivity analysis, we utilize $k = 3$ expansion rounds and generate $N_{QA} = 10$ probing questions to maximize performance while containing costs. This configuration results in approximately **20 LLM calls** per paper review, compared to a single forward pass ($C = 1$) for static fine-tuned models (e.g., DeepReviewer) or 5–10 calls for standard agentic baselines (e.g., AI Scientist v2).

D.2 Empirical Token Usage and Monetary Cost

To calculate the practical deployment cost of SCHOLARPEER, we aggregated the total token consumption across all internal agent calls over the ScholarEval test set. Because the system actively retrieves and compresses multiple external literature sources, an average SCHOLARPEER review consumes approximately ∼429K input tokens and ∼26K output tokens per manuscript.

Using standard enterprise API pricing for our primary backbone model, Gemini 3.1 Pro ($2.00 per 1M input tokens and $12.00 per 1M output tokens)[3], this equates to a monetary cost of ∼**$1.20 per review**.

D.3 Latency and Scalability

We evaluated the system's runtime latency to ensure viability for major machine learning conferences, which frequently receive over 10,000 submissions annually.

- **Single-Paper Latency:** Without batching, executing the full parallelized multi-agent workflow—including web retrieval, internal extraction, and active interrogation—for a single manuscript takes approximately ∼**10 minutes**.
- **Amortized Latency:** When processing papers in parallel batches (using 20 concurrent threads), a full run on the DeepReview-13K test set (1,286 papers) completes in ∼20 hours. This yields a highly scalable amortized latency of just ∼**56 seconds per paper**.

Summary: While SCHOLARPEER requires more computational resources than static single-pass models, those models fundamentally lack the structural capability to perform external literature searches or actively verify claims. The ∼$1.20 cost and sub-minute amortized latency represent a negligible operational overhead when compared to the hundreds of human hours required to manually audit missing baselines and verify technical soundness, making SCHOLARPEER an exceptionally practical augmentation tool for real-world academic deployment.

E Qualitative Examples

We present examples of reviews generated by baselines and compare them to SCHOLARPEER . These examples highlight the system's ability to identify missing baselines and verify novelty.

E.1 Example 1: Detection of Missing Baselines

Paper Title: DBRNet: Advancing Individual-Level Continuous Treatment Estimation through Disentangled and Balanced Representation
OpenReview Link: `https://openreview.net/pdf?id=tlqmkftgpw`

Baseline Review (DeepReviewer-14B)

Summary:
This paper introduces DBRNet, a novel neural network architecture designed for estimating individualized dose-response functions (IDRFs) in the presence of continuous treatments. The core contribution of DBRNet lies in its integration of three key components: disentangled representation learning, balancing via re-weighting, and a varying coefficient network. The method begins by learning disentangled representations of covariates, aiming to separate instrumental, confounding, and adjustment factors. This disentanglement is achieved through a combination of a discrepancy loss and an independent loss. To address selection bias, DBRNet employs a re-weighting function based on the learned representations, which aims to create a more balanced comparison group. Finally, a varying coefficient network is used to model the continuous treatment effect, allowing for flexible modeling of the dose-response relationship. The authors provide theoretical justification for their debiasing approach, arguing that the combination of disentangled representations and re-weighting effectively mitigates selection bias. Empirically, the authors demonstrate the effectiveness of DBRNet on both synthetic and semi-synthetic datasets, showing that it outperforms several state-of-the-art methods in terms of mean integrated squared error (MISE) and average mean squared error (AMSE). The paper presents a comprehensive framework for IDRF estimation, combining several existing ideas in a novel way. While the empirical results are promising, the paper also has several limitations that warrant further discussion. The paper's focus on methodological innovation is commendable, but a more thorough

[3] `https://cloud.google.com/vertex-ai/generative-ai/pricing`

discussion of practical considerations and limitations would enhance its overall impact. The paper's contribution is significant in its attempt to address the challenges of continuous treatment effects, but the identified weaknesses need to be addressed to ensure the robustness and generalizability of the proposed method. The paper's approach to disentangled representation learning and re-weighting is a valuable contribution to the field, but the lack of discussion on computational complexity, missing data, and high-dimensional covariates limits its practical applicability. The paper's theoretical justification is a strength, but the empirical evaluation could be strengthened by including more diverse datasets and a more detailed analysis of the method's sensitivity to various data characteristics. Overall, the paper presents a valuable contribution to the field of individualized treatment effect estimation, but it also highlights several areas where further research is needed.

Strengths:
The paper presents a well-structured and clearly articulated approach to estimating individualized dose-response functions (IDRFs) with continuous treatments. The core strength of the paper lies in its novel integration of disentangled representation learning, balancing via re-weighting, and a varying coefficient network. This combination, while drawing inspiration from existing methods, is a unique contribution that addresses the challenges of continuous treatment effects in a comprehensive manner. The theoretical justification for the debiasing approach is another significant strength. The authors provide a clear argument for how the proposed method mitigates selection bias through the use of disentangled representations and re-weighting, which adds a layer of rigor to the proposed methodology. The empirical results, while limited in scope, demonstrate the effectiveness of DBRNet on both synthetic and semi-synthetic datasets. The paper shows that DBRNet achieves superior performance compared to several state-of-the-art methods in terms of MISE and AMSE, which provides empirical support for the proposed approach. The ablation study further strengthens the empirical findings by demonstrating the importance of the re-weighting function and the discrepancy loss, highlighting the contribution of the specific combination of components. The paper's focus on addressing the challenges of continuous treatments is also a strength, as this is an important and often overlooked aspect of individualized treatment effect estimation. The paper's clear presentation of the method, including the detailed description of the objective function and its components, makes it easy to follow and understand. The use of a varying coefficient network to model the continuous treatment effect is a technically sound approach that allows for flexible modeling of the dose-response relationship. The paper's overall contribution is significant in that it provides a comprehensive framework for IDRF estimation that combines several existing ideas in a novel and effective way. The paper's focus on addressing selection bias through disentangled representations and re-weighting is a valuable contribution to the field, and the empirical results provide initial evidence of the method's effectiveness.

Weaknesses:
Upon careful examination, several weaknesses in the paper warrant attention. First, the paper lacks a discussion on the generalizability of DBRNet to datasets with missing covariates. The method description and experimental setup assume complete data, and there is no mention of how missing data would be handled in the disentanglement process or the re-weighting function. This is a significant limitation, as missing data is a common issue in real-world observational studies. The paper does not address how the disentanglement or re-weighting would be affected by missing data, especially under different missing data mechanisms. This omission raises concerns about the practical applicability of DBRNet in real-world scenarios where missing data is a common issue. My confidence in this weakness is high, as the paper's method description and experimental setup explicitly assume complete data, and there is no discussion of handling missing values. Second, the paper does not provide a detailed analysis of the time complexity of DBRNet. The paper describes the components of DBRNet but does not provide a formal analysis of their computational cost. The experimental section lacks any timing information or comparisons of training/inference time with other methods. This is a significant oversight, as the computational cost of a method is a crucial factor in its practical applicability, especially on large-scale datasets. Without a time complexity analysis, it is difficult to assess the feasibility of applying DBRNet to real-world problems. My confidence in this weakness is high, as the paper lacks any discussion or empirical evaluation of the computational cost of DBRNet. Third, the paper also lacks a discussion of the space complexity of DBRNet. The paper describes the network architecture but does not provide a formal analysis of the memory requirements for storing the model parameters and intermediate representations. The experimental section does not provide any information about memory usage. This is a significant limitation, as the memory footprint of a method is a crucial factor in its practical applicability, especially in resource-constrained environments. Without a space complexity analysis, it is difficult to assess the feasibility of applying DBRNet in resource-limited settings. My confidence in this weakness is high, as the paper lacks any discussion or empirical evaluation of the memory requirements of DBRNet. Fourth, the paper does not adequately address the limitations of DBRNet when dealing with high-dimensional covariates. While the paper uses neural networks, which can be prone to overfitting in high-dimensional settings, there is no specific discussion about the challenges of high-dimensional data. The datasets used in the experiments have a moderate number of features, and the paper does not include experiments on datasets with a very large number of covariates. This is a significant limitation, as high-dimensional data is common in many real-world applications. The paper does not discuss potential strategies for mitigating the curse of dimensionality, such as dimensionality reduction techniques or regularization methods. My confidence in this weakness is high, as the paper lacks any discussion of the performance and potential limitations of DBRNet on high-dimensional datasets. Finally, while the paper touches upon relevant concepts like selection bias, it lacks an explicit and detailed discussion of methods tailored for observational studies. While the paper compares against methods that can be used in observational settings, it does not explicitly frame the discussion around the specific challenges and solutions prevalent in the observational causal inference literature. This omission limits

the paper's ability to contextualize its contribution within the broader field of causal inference. My confidence in this weakness is medium, as the paper does discuss methods relevant to observational settings, but it lacks an explicit and detailed discussion of the specific literature on ITE estimation in observational studies. Additionally, the paper does not provide a detailed comparison of DBRNet with transformer-based and GAN-based DRF estimation methods, particularly regarding long-range dependencies and mode collapse. This lack of comparison limits the paper's ability to contextualize its contribution within the broader field of causal inference. My confidence in this weakness is high, as the related work section lacks a detailed comparison of DBRNet's architecture and assumptions with transformer-based and GAN-based DRR estimation methods. Furthermore, it is unclear whether DBRNet uses targeted regularization, and if so, how it compares to the baselines with targeted regularization. Providing results without targeted regularization for DBRNet would ensure a fairer comparison. My confidence in this weakness is high, as the paper does not explicitly state whether DBRNet uses targeted regularization, making it unclear how the comparison with baselines using targeted regularization is fair. Results without targeted regularization for DBRNet are missing.

Suggestions:
To address the identified weaknesses, I recommend several concrete improvements. First, the authors should thoroughly investigate the performance of DBRNet on datasets with missing covariates. This should involve conducting experiments on datasets with simulated missing values under different assumptions (missing completely at random, missing at random, missing not at random) and analyzing how the disentanglement and re-weighting processes are affected. The authors should also explore potential strategies for handling missing data, such as imputation techniques, and how these strategies might interact with DBRNet's performance. This analysis should include both theoretical considerations and empirical evaluations on datasets with varying degrees and patterns of missingness. Second, the authors should provide a detailed analysis of the time complexity of DBRNet. This should include an analysis of the computational cost associated with each step of the algorithm, such as the disentanglement of representations, the calculation of the re-weighting function, and the outcome prediction. A comparison with the computational complexity of other state-of-the-art methods would be beneficial for assessing the practical applicability of DBRNet, especially on large-scale datasets. Additionally, the authors should discuss the practical implications of the computational complexity, such as the feasibility of applying DBRNet to large-scale datasets and the potential need for parallelization or other optimization techniques. Third, the authors should provide a detailed analysis of the space complexity of DBRNet. This should include an analysis of the memory requirements for storing the model parameters and intermediate representations. The authors should also discuss how the memory footprint scales with the size of the input data and the complexity of the model. This analysis should consider the impact of different hyperparameter settings on the memory requirements. Fourth, the authors should address the limitations of DBRNet when dealing with high-dimensional covariates. This should include a discussion of the potential challenges associated with high-dimensional data, such as the curse of dimensionality and the increased risk of overfitting. The authors should explore potential solutions, such as dimensionality reduction techniques or regularization methods, to mitigate these challenges. Furthermore, the authors should discuss the impact of high dimensionality on the disentanglement process and the accuracy of the treatment effect estimation. It would be beneficial to include experiments on datasets with varying numbers of covariates to demonstrate the scalability of DBRNet and to identify any potential limitations. Fifth, the authors should expand the related work section to include a more comprehensive discussion of methods specifically designed for estimating individualized continuous treatment effects in observational studies. This should include a discussion of the specific challenges and solutions prevalent in the observational causal inference literature. The authors should also provide a more detailed comparison of DBRNet with transformer-based and GAN-based DRF estimation methods, particularly regarding long-range dependencies and mode collapse. This should include a discussion of the specific architectural differences, such as the use of attention mechanisms in transformers and how they handle long-range dependencies in the covariate space, compared to the proposed method. It should also address the challenges of training GANs for continuous treatment settings, such as mode collapse, and how the proposed method avoids these issues. Finally, the authors should clarify whether DBRNet uses targeted regularization and, if so, which loss terms are targeted. If DBRNet does not use targeted regularization, the authors should provide results without targeted regularization to ensure a fair comparison with baselines that use targeted regularization. These changes would significantly improve the paper's clarity, rigor, and practical relevance.

Questions:
Several key questions arise from my analysis of the paper. First, how does the performance of DBRNet degrade when applied to datasets with varying degrees of missingness, and what are the best strategies for handling missing data in the context of the proposed method? Specifically, how do different imputation techniques interact with the disentanglement and re-weighting processes? Second, what is the precise time complexity of each component of DBRNet, and how does it compare to the time complexity of other state-of-the-art methods for IDRF estimation? What are the practical implications of this computational cost, especially when applied to large-scale datasets? Third, what is the space complexity of DBRNet, and how does the memory footprint scale with the size of the input data and the complexity of the model? Are there any specific optimizations that can be implemented to reduce the memory requirements of DBRNet? Fourth, how does the performance of DBRNet change when applied to datasets with high-dimensional covariates, and what are the best strategies for mitigating the curse of dimensionality in this context? Are there specific dimensionality reduction techniques or regularization methods that are particularly effective for DBRNet? Fifth, how does DBRNet compare to other methods specifically designed for observational studies, and what are the key differences in their assumptions and approaches? How does DBRNet address the specific challenges

of confounding and selection bias in observational settings? Sixth, how does DBRNet's architecture compare to transformer-based and GAN-based DRF estimation methods in terms of handling long-range dependencies and mode collapse? What are the specific advantages and disadvantages of each approach? Finally, does DBRNet use targeted regularization, and if so, which loss terms are targeted? If not, how does the performance of DBRNet compare to the baselines when the baselines are not using targeted regularization? These questions highlight key uncertainties and areas where further clarification and investigation are needed to fully understand the strengths and limitations of the proposed method.

SCHOLARPEER Review

Summary:
The paper proposes DBRNet (Disentangled and Balanced Representation Network) for estimating Individualized Dose-Response Functions (IDRF) in continuous treatment settings. The method addresses selection bias by disentangling covariates into three latent factors: Instrumental, Confounder, and Adjustment factors. It employs three specific neural networks to extract these factors, using only the Instrumental and Confounder factors to estimate propensity scores for re-weighting, and the Confounder and Adjustment factors to predict outcomes via a Varying Coefficient Network. The authors introduce a specific loss function to enforce independence between adjustment factors and treatment, and provide theoretical proofs for the unbiasedness of their re-weighting scheme.

Strengths:

- **Theoretical Motivation**: The core idea of 'selective balancing' specifically avoiding the balancing of instrumental and adjustment factors is well-grounded in causal inference theory and addresses a known limitation of methods that balance the entire covariate representation.
- **Ablation Studies**: The ablation study effectively demonstrates the contribution of individual components, particularly the discrepancy loss and re-weighting module, to the model's performance on the chosen datasets.

Weaknesses:

- **Missing Critical Baselines**: The paper claims that no existing efforts are capable of precisely adjusting for selection bias in continuous settings. This overlooks significant recent contributions such as **ADMIT (NeurIPS 2022)** and **End-to-End Balancing (ICML 2022)**, which explicitly focus on re-weighting and balancing for continuous treatments. The comparison is largely limited to methods from 2021 (VCNet) and earlier.
- **Missing Standard Benchmark**: The evaluation excludes the **TCGA** dataset, which is a standard high-dimensional semi-synthetic benchmark for continuous treatment effect estimation used by VCNet, SCIGAN, TransTEE, and ADMIT. The omission of this challenging high-dimensional dataset limits the assessment of the method's robustness.
- **Problematic Independence Loss (L_{ind})**: The proposed mechanism to enforce independence (minimizing the log-likelihood of treatment given adjustment factors) is mathematically brittle. Maximizing prediction error is unbounded and does not guarantee that the representation contains no information about the treatment (mutual information minimization), only that it predicts the specific observed treatment poorly. Standard approaches utilize Adversarial Training (Gradient Reversal Layers) or mutual information minimization to ensure true independence.
- **Naive Density Estimation**: Despite focusing on continuous treatments, the propensity score estimation relies on discretizing the treatment space into grids and using linear interpolation. This introduces approximation errors and seems inconsistent with the goal of handling continuous treatments rigorously (e.g., compared to Mixture Density Networks or Normalizing Flows).
- **Lack of Balance Metrics**: The paper evaluates performance solely on outcome prediction error (MISE/AMSE). It lacks metrics evaluating covariate balance (e.g., weighted correlations) or the accuracy of the propensity estimation itself, which are crucial for verifying the claim of 'precise bias adjustment'.

Suggestions:

- Include comparisons with ADMIT (NeurIPS 2022) and End-to-End Balancing (ICML 2022) to substantiate claims of state-of-the-art performance.
- Evaluate the method on the TCGA dataset to demonstrate robustness in high-dimensional settings.
- Replace the L_{ind} maximization of error with a more stable independence constraint, such as a Gradient Reversal Layer (GRL) or minimizing Mutual Information.
- Implement a truly continuous density estimator (e.g., Normalizing Flows or MDNs) for the propensity score to align with the continuous treatment setting.
- Report metrics on covariate balance (e.g., weighted correlation between confounders and treatment) to empirically verify the 'precise adjustment' claim.

Questions:

- Why were recent state-of-the-art re-weighting methods for continuous treatments, specifically ADMIT (NeurIPS 2022) and End-to-End Balancing (ICML 2022), omitted from the baselines?
- Why was the standard TCGA benchmark excluded from the evaluation, given its prevalence in the literature (used by VCNet, TransTEE, etc.)?
- Minimizing the log-likelihood (maximizing negative log-likelihood) is unbounded. How do you prevent the encoder weights from exploding or the distribution from shifting to infinity to satisfy this loss? Why was this chosen over standard Adversarial Training (Gradient Reversal Layer)?
- The treatment density estimator uses discretization (grids). How sensitive is the performance to the number of grids (B), and why was a continuous density estimator (like MDN) not used?

LLM-as-a-Judge Evaluation (Gemini 3.1 Pro)

Winner: SCHOLARPEER
Technical Accuracy Reason: ScholarPeer correctly identifies that the paper ignores critical recent SOTA methods (*ADMIT*, *End-to-End Balancing*) and standard benchmarks (*TCGA*) that contradict the authors' claims. The Baseline focuses on generic issues like missing data and complexity which are less central to the paper's scientific validity.
Analytical Depth Reason: ScholarPeer critiques the mathematical formulation of the loss function (unbounded maximization of error) and the inconsistency of using discrete grid-based density estimation for a continuous treatment method. DeepReview 14B does not engage with the technical details of the architecture or loss functions.
Novelty & Significance Reason: ScholarPeer correctly identifies that the paper's claim of being the 'first' to precisely adjust for selection bias in continuous settings is false, citing specific prior works (ADMIT, End-to-End Balancing) published in 2022. DeepReview 14B accepts the authors' claims without verifying the existence of recent competing methods.
The following external sources found during the search helped verify this:
- *Bahadori, M. T., Tchetgen, E. T., & Heckerman, D. (2022). End-to-End Balancing for Causal Continuous Treatment-Effect Estimation. ICML.*
- *Wang, X., Lyu, S., Wu, X., Wu, T., & Chen, H. (2022). Generalization Bounds for Estimating Causal Effects of Continuous Treatments. NeurIPS.*

Constructive Value Reason: ScholarPeer provides specific, actionable steps to bring the paper up to scientific standards: compare against named baselines, use the TCGA dataset, and fix the unstable loss using Gradient Reversal Layers. The Baseline's suggestions ("analyze time complexity") are generic and apply to almost any deep learning paper.
Overall Verdict: ScholarPeer is superior in every dimension. It identifies a fatal flaw (ignoring direct prior work), points out the omission of a standard benchmark, and offers deep technical critiques of the loss function. The Baseline provided a superficial review with generic comments.

E.2 Example 2: Technical Soundness

Paper Title: GAN-based Vertical Federated Learning for Label Protection
OpenReview Link: `https://openreview.net/pdf?id=5Vh0XqOTGi`

Baseline Review (AI Scientist v2)

Summary:
The paper proposes the Generative Adversarial Federated Model (GAFM) to mitigate Label Leakage from Gradients (LLG) in Vertical Federated Learning (VFL) based on SplitNN. The method replaces the standard label-party architecture with a GAN, where a generator predicts labels and a discriminator aligns them with the true distribution. A randomized Cross-Entropy loss is added to maintain utility. The approach relies on the mutual perturbation of gradients from the GAN and CE objectives to obscure label information. The authors also introduce two new gradient-based attacks (Mean and Median attacks) and demonstrate that GAFM achieves a better privacy-utility trade-off than baselines like Marvell and Max Norm on binary classification tasks.

Strengths:

- **Novelty**The integration of a GAN-based architecture specifically for privacy preservation in SplitNN is a novel design choice.
- **Empirical Performance**: The method demonstrates a superior trade-off between model utility (AUC) and privacy (Leak AUC) compared to strong baselines like Marvell. It also shows better stability across random seeds.

- **Evaluation Metrics**: The proposal of Mean and Median attacks provides a stronger and more rigorous benchmark for evaluating gradient separability compared to the standard Norm attack.
- **Ablation Studies**: The necessity of both the GAN component (for distribution matching) and the randomized CE component (for instance-level accuracy) is well-supported by ablation studies.

Weaknesses:

- **Lack of Theoretical Guarantees**: The defense is heuristic. Unlike some baselines that offer bounds based on information theory (e.g., KL divergence), GAFM relies on empirical validation of gradient mixing.
- **Computational Overhead**: The paper does not sufficiently analyze the training time and computational resources required for the GAN components compared to standard SplitNN or simpler defenses.
- **Attack Assumptions**: The proposed attacks assume the attacker knows the gradient centers or medians of the classes. While useful for measuring separability, this may not fully reflect realistic scenarios where attackers must estimate these via clustering.
- **Hyperparameter Tuning**: The suggested strategy of tuning parameters on a data subset may be non-trivial to implement in a strict VFL setting without violating privacy constraints.

Limitations:

- The method lacks formal theoretical privacy bounds.
- Current evaluation is limited to binary classification.
- Computational complexity is higher than simple noise-addition methods.

Questions:

- What is the quantitative computational overhead (time per epoch, memory) of GAFM compared to Marvell and vanilla SplitNN?
- Regarding the randomized response mechanism: Since the target intervals for class 0 and 1 are disjoint (for
$Delta < 0.5$), how does the method theoretically ensure that the GAN gradient sufficiently masks the CE gradient component to prevent label recovery?
- Can the authors clarify how the hyperparameter tuning on a subset can be performed securely in a real-world VFL deployment?
- Do the Mean/Median attacks perform similarly if the attacker estimates centers using unsupervised clustering (e.g., K-Means) instead of using ground truth statistics?

SCHOLARPEER Review

Summary:
The paper introduces the Generative Adversarial Federated Model (GAFM) to address Label Leakage from Gradients (LLG) in Vertical Federated Learning (specifically Split Learning). The authors propose replacing the standard server-side model with a GAN architecture, where a Generator acts as the predictor and a Discriminator aligns the prediction distribution with the label distribution. The training objective combines a Wasserstein GAN loss with a Randomized Cross-Entropy (CE) loss (using Randomized Response). The authors argue that this combination creates 'mutual gradient perturbation,' where the conflicting gradients from the GAN and CE components obfuscate the true label information in the gradients sent to the client. The paper also introduces two new gradient clustering attacks (Mean and Median) and evaluates the method on four binary classification datasets.

Strengths:

- **Problem Relevance:** The paper addresses a critical privacy vulnerability (Label Leakage from Gradients) in Split Learning, which is a significant barrier to the adoption of Vertical Federated Learning.
- **New Attack Baselines:** The proposed Mean and Median attacks appear to be more effective than the standard Norm attack, providing a stronger empirical baseline for evaluating future defenses.
- **Empirical Performance:** On the tested binary classification datasets, GAFM demonstrates a favorable trade-off between model utility (AUC) and privacy (Leak AUC) compared to the selected baselines (Marvell, Max Norm).

Weaknesses:

- **Lack of Theoretical Guarantees:** A major weakness is the absence of formal privacy guarantees. Unlike Marvell, which provides a theoretical upper bound on leakage based on KL-divergence, or Differential Privacy methods, GAFM relies on a heuristic argument of 'mutual gradient perturbation.' In privacy

research, empirical defense against specific attacks is often insufficient as it does not guarantee protection against adaptive or future attacks.

- **Limited Scope (Binary Only):** The proposed mechanism relies heavily on the geometry of binary classification, where gradients for positive and negative classes tend to pull in opposite directions. The paper explicitly acknowledges this limitation. This severely restricts the significance of the work, as the method does not appear generalizable to multi-class problems where gradient interactions are high-dimensional and complex.
- **Missing Baselines:** The evaluation is restricted to perturbation-based defenses. The paper fails to compare against state-of-the-art non-perturbation defenses such as Confusional AutoEncoders (CoAE) or deterministic gradient quantization methods (e.g., DiscreteSGD-like approaches), which are relevant to the LLG problem.
- **Questionable Stability Claims:** The paper claims GAFM is more stable than Marvell. However, GANs are notoriously difficult to train and prone to instability (mode collapse, vanishing gradients). The defense relies on a hyperparameter
$gamma$ to balance GAN and CE losses; if this balance is off, the paper admits leakage re-occurs. This sensitivity suggests that the method might be less robust in practice than claimed.

Suggestions:

- Extend the theoretical analysis to provide a worst-case privacy bound.
- Include comparisons with a broader range of baselines, including CoAE and quantization methods.
- Investigate the applicability of the method to multi-class settings or explain theoretically why it cannot be adapted.
- Perform a more robust stability analysis, specifically stressing the GAN training stability.

Questions:

- Can you derive any formal privacy bound (e.g., in terms of Differential Privacy or Information Bottleneck) for the proposed method, rather than relying solely on empirical 'gradient mixing'?
- How does the method perform against non-clustering based attacks, such as those that attempt to reconstruct features first to infer labels (e.g., if features are hijacked)?
- The paper claims higher stability than Marvell, yet GANs are known for instability. Can you provide a more rigorous sensitivity analysis of the hyperparameter
$gamma$ and the learning rates? What happens to privacy if the GAN suffers from mode collapse?
- Why were non-perturbation baselines like CoAE or simple gradient quantization omitted from the comparison?

LLM-as-a-Judge Evaluation (Gemini 3.1 Pro)

Winner: SCHOLARPEER
Technical Accuracy Reason: ScholarPeer correctly identifies the inherent instability of GANs as a counter-argument to the paper's stability claims and correctly highlights the lack of formal privacy guarantees compared to baselines like Marvell. AI Scientist v2 accepts the stability claims without questioning the underlying GAN dynamics.
Analytical Depth Reason: ScholarPeer provides a deeper analysis of the method's limitations, specifically explaining why the reliance on gradient opposition restricts the method to binary classification and cannot be easily extended to multi-class. ScholarPeer also critically examines the 'mutual gradient perturbation' heuristic.
Novelty & Significance Reason: ScholarPeer accurately identifies that the paper fails to compare against relevant prior work, specifically Confusional AutoEncoders (CoAE) (Zou et al., 2022), which addresses the same problem using a similar 'label disguise' concept. AI Scientist v2 misses this critical missing baseline.
The following external sources found during the search helped verify this:
- *Zou, T., et al., 'Defending Batch-Level Label Inference Attacks in Vertical Federated Learning', arXiv:2202.02073, 2022 (Referred to as CoAE/Label Disguise)*
- *Chen, C., et al., 'Gradient Compression for Communication-Efficient Split Learning', 2021 (Relevant for quantization baselines)*

Constructive Value Reason: ScholarPeer provides specific, high-value suggestions by identifying missing state-of-the-art baselines (CoAE and gradient quantization) that are critical for a fair evaluation. AI Scientist v2's suggestions are more generic (e.g., asking about computational overhead).
Overall Verdict: ScholarPeer provides a significantly more expert review. It identifies critical missing baselines (CoAE) that predate the paper, challenges the technical claims regarding stability with well-founded skepticism about GANs, and correctly assesses the limited significance of a binary-only defense. AI Scientist v2 gives a good summary but misses the key prior work and accepts the paper's claims too readily.

E.3 Summary of Gains and Losses

We provide below detailed summaries of the advantages and disadvantages of SCHOLARPEER against baselines as generated by an LLM based on the reasoning traces of the LLM-as-a-judge based evaluations.

Vs. CycleReviewer-70B

✅ **Advantage: Bibliographic Verification & SOTA Assessment**
ScholarPeer consistently utilizes external knowledge to identify specific, high-relevance prior works (often published within the same year) that the authors omitted. It uses these citations (e.g., Lightman et al. 2023 for MATH, ACT for robotics, EATA for speech) to validly challenge novelty claims, whereas the competitor fails to identify missing literature.

✅ **Advantage: Methodological Forensic Analysis**
ScholarPeer demonstrates a superior ability to detect fatal experimental flaws that invalidate results. It correctly identified issues such as data leakage (e.g., applying VMD before train-test splits, intra-patient leakage), confounding variables (e.g., action chunking vs. implicit policies), and circular biases in dataset creation.

✅ **Advantage: Content Faithfulness & Recall**
ScholarPeer accurately locates specific details within the paper (e.g., definitions in Section 3, hyperparameters in Section 4, limitations in Section 5). In contrast, the competitor frequently 'hallucinates' that these sections are missing and penalizes the authors for omitting information that is actually present.

✅ **Advantage: Actionable Constructiveness**
ScholarPeer provides precise, scientifically grounded instructions for improvement. Instead of generic requests for 'more details,' ScholarPeer suggests specific missing baselines to run (e.g., 'Compare against MinRecon'), specific ablations to perform (e.g., 'Euclidean Norm-Gated'), and specific metrics to report (e.g., 'Inference latency in Hz').

✅ **Advantage: Theoretical & Domain Depth**
ScholarPeer engages with the fundamental mathematical and physical assumptions of the research. It correctly identified theoretical violations (e.g., finite-sample exchangeability, lack of equivariance in molecular dynamics) and practical hardware constraints (e.g., memory usage during backpropagation) that the competitor missed.

⛔ **Disadvantage: No Significant Failure Patterns**
In the provided set of 20 evaluation traces, ScholarPeer outperformed CycleReviewer 70B in every instance. No recurring negative traits, hallucinations, or relative weaknesses were observed for ScholarPeer; it consistently provided expert-level reviews while the competitor struggled with factual accuracy and repetition.

Vs. DeepReviewer-14B

✅ **Advantage: Citation & SOTA Verification**
ScholarPeer consistently identifies specific, relevant prior work (e.g., Lightman et al., MedSAM, TaskPrompter, MDT) that invalidates authors' claims of novelty or State-of-the-Art performance. It frequently cites publication dates and specific results to refute claims, whereas the competitor often accepts the authors' assertions at face value.

✅ **Advantage: Experimental Integrity Detection**
ScholarPeer excels at spotting fatal methodological flaws that undermine a paper's validity. Recurring examples include identifying data leakage (e.g., training on validation sets or using broad datasets like PythonAlpaca for self-contained benchmarks), circular evaluation protocols, and the use of inappropriate metrics (e.g., using BLEU for Text-to-SQL tasks instead of Execution Accuracy).

✅ **Advantage: Domain-Specific Depth**
ScholarPeer demonstrates deep understanding of specific research sub-fields, allowing it to distinguish between surface-level similarities and fundamental technical differences. Examples include distinguishing between 'Super-Resolution' and 'Learnable Resizers', recognizing the need for Trusted Execution Environments (TEEs) in privacy frameworks, and identifying when 'VQA' datasets are merely converted classification tasks.

✅ **Advantage: Actionable Constructiveness**
ScholarPeer provides highly specific recommendations for improvement. Instead of generic requests to 'add more experiments' or 'analyze limitations,' ScholarPeer names the exact missing baselines (e.g., DIN-SQL, RoleLLM, STID), specific datasets (e.g., RepoBench, HMDB51), and precise ablation studies needed to validate the core contributions.

✅ **Advantage: Factual Precision & Reading Comprehension**
ScholarPeer accurately parses the paper text, avoiding the 'hallucinated critiques' observed in the competitor. While DeepReviewer 14B frequently claims details (like hyperparameters or metric definitions) are missing when they are explicitly stated, ScholarPeer correctly locates these details and instead focuses on internal inconsistencies (e.g., datasets listed in settings but missing from results).

⛔ **Disadvantage: None Observed**
In the sampled evaluation traces, SCHOLARPEER was judged superior in 20/20 instances, showing no recurring failure patterns relative to this baseline.

Vs. AI Scientist v2

✅ **Advantage: Bibliographic & SOTA Mastery**
ScholarPeer consistently identifies specific, named, and dated prior works (e.g., PatchTST, DiffLinker, LlamaGuard 3) that invalidate the authors' claims of novelty or performance. It accurately flags when papers ignore baselines published just months prior to the cutoff, whereas AI Scientist v2 often accepts the provided baselines as sufficient.

✅ **Advantage: Experimental Rigor & Validity**
ScholarPeer excels at critiquing the experimental setup itself, identifying fatal flaws such as unfair backbone standardization, the omission of critical metrics (e.g., Chamfer Distance for 3D generation), or the use of inappropriate datasets (e.g., testing 'Reasoning' on Factuality benchmarks).

✅ **Advantage: Theoretical Contextualization**
ScholarPeer demonstrates deep domain understanding by connecting papers to broader theoretical frameworks and known phenomena (e.g., the 'Linearity vs. Nonlinearity' debate in RecSys, 'Gradient Masking' in adversarial defense, or 'Grokking'), providing a more sophisticated critique than AI Scientist v2's surface-level analysis.

✅ **Advantage: Actionable Constructiveness**
ScholarPeer provides highly specific instructions for improvement, such as naming the exact datasets (e.g., SR25, GSO) or specific control experiments (e.g., Self-Consistency) required to validate the paper, whereas AI Scientist v2 tends to offer generic suggestions like 'add complexity analysis' or 'fix typos'.

⚠️ **Disadvantage: Internal Sanity Checking**
In specific instances, ScholarPeer overlooks fatal internal inconsistencies or logical impossibilities within the paper's own text (e.g., an impossible ratio of training data to downstream performance, or mathematically invalid probability comparisons across tokenizers) because it is too focused on external comparisons. AI Scientist v2 tends to catch these 'sanity check' failures.

⚠️ **Disadvantage: Reproducibility & Completeness**
ScholarPeer occasionally misses basic reproducibility issues, such as undefined algorithmic operators (e.g., missing mutation steps) or placeholder/broken text in the manuscript, which AI Scientist v2 correctly identifies as critical flaws.

Vs. Agent Review

✅ **Advantage: Literature & SOTA Verification**
ScholarPeer demonstrates a superior ability to verify novelty claims against the external literature. It consistently identifies specific, named, and dated prior works (e.g., ReContrast, Shap-E, CCS, Perceiver IO) that invalidate a paper's claims of being 'State-of-the-Art' or the 'first' to solve a problem. It accurately checks publication dates to ensure baselines were available before the paper's submission.

✅ **Advantage: Benchmarking Standards**
ScholarPeer exhibits expert-level knowledge of community standards for datasets and metrics. It frequently critiques papers for using deprecated or insufficient benchmarks (e.g., using SMACv1 instead of SMACv2, omitting GSM8K for reasoning claims, or ignoring ZJU-Mocap for avatars) and demands the specific evaluations required to meet current scientific rigor.

✅ **Advantage: Experimental Validity & Leakage Detection**
ScholarPeer is highly effective at identifying fatal flaws in experimental design, particularly regarding data leakage and unfair comparisons. It correctly flagged issues such as using Stable Diffusion (trained on LAION) for zero-shot ImageNet testing, improper data splitting in medical imaging (slice-level vs. patient-level), and statistically insignificant sample sizes (e.g., 10 prompts).

⚠️ **Disadvantage: Internal Logical Consistency**
ScholarPeer occasionally misses deep, internal logical or mathematical contradictions within the paper's own arguments, which Agent Review tends to catch. Examples include failing to notice that 'average pooling' destroys the claimed 'sequential' information, or that 'forgetting' a concept is mathematically incompatible with maintaining high accuracy on that specific class.

⚠️ **Disadvantage: Reproducibility & Completeness**
In a minority of cases, ScholarPeer misread or miscalculated specific values from the paper's internal tables. For instance, it incorrectly attributed a 'significant portion' of gains to optimization tricks when the ablation table showed otherwise, whereas Agent Review accurately reported the specific performance contributions.

Vs. Stanford Agent Review

✅ **Advantage: Temporal Validity**
ScholarPeer consistently respects the chronological context of the papers under review. Unlike the Stanford Agent Reviewer, which frequently 'hallucinated' or cited papers published months or years after the review cutoff date (e.g., citing 2025 papers for a 2024 review), ScholarPeer correctly identifies and cites prior work that was actually available at the time of submission.

✅ **Advantage: Experimental Rigor (Baselines)**
ScholarPeer is significantly better at identifying missing 'Gold Standard' or State-of-the-Art (SOTA) baselines. It frequently catches authors using 'strawman' comparisons (e.g., GBDT vs. Transformers) and demands comparisons

against specific, established competitors (e.g., AWP, SLDA, EvShutter) that are essential for validating the paper's core claims.

✅ **Advantage: Research Integrity & Presentation**

ScholarPeer excels at detecting fundamental flaws in the manuscript's state and integrity. It successfully flagged 'Selective Reporting' (where datasets were listed but results omitted), identified unfinished drafts containing LaTeX 'TODO' comments, and noted when authors claimed datasets didn't exist when they actually did.

✅ **Advantage: Practical Feasibility**

ScholarPeer provides feedback grounded in the practical realities of research. It critiques methods for prohibitive computational costs (e.g., extreme sampling requirements) and suggests experiments that are physically possible, whereas the competitor often suggests impossible comparisons against future work.

⚠️ **Disadvantage: Mathematical & Theoretical Scrutiny**

ScholarPeer tends to accept the authors' mathematical premises at face value, missing deep theoretical flaws that the competitor catches. Examples include failing to notice that a Fourier basis was not orthonormal on a sphere, or that a theorem relied on square matrices while the method used rectangular ones.

⚠️ **Disadvantage: Internal Consistency Checking**

ScholarPeer struggles to identify contradictions within the paper's own content. It missed instances where the text claimed one thing (e.g., using multiple backbones) but the data tables proved another (parameter counts matching only one backbone), or where t-SNE plots contradicted the authors' claims of diversity.

E.4 SCHOLARPEER Agent Outputs

We show below the individual agent outputs for a sample paper - DBRNet: Advancing Individual-Level Continuous Treatment Estimation through Disentangled and Balanced Representation (OpenReview Link - `https://openreview.net/pdf?id=tlqmkftgpw`). The framework's review for this paper has been provided in Appendix E.1.

Summarizer Agent Output

1. Problem Statement & Motivation:

The paper addresses the challenge of estimating the **Individualized Dose-Response Function (IDRF)**—the causal effect of **continuous treatments** (e.g., medication dosage) on individual outcomes.

Current Limitations Identified:

- **Discrete Constraints:** Most existing Individual Treatment Effect (ITE) methods are limited to binary or discrete treatments and cannot handle infinite counterfactuals in continuous settings.
- **Imprecise Bias Adjustment:** Existing continuous methods (e.g., DRNet, VCNet) typically attempt to "balance" the **entire** latent representation of covariates to remove selection bias. The authors argue this is theoretically flawed because:
  - **Instrumental factors** (affecting treatment only) should not be balanced.
  - **Confounder factors** (affecting treatment and outcome) contain predictive information and should not be discarded.

2. Methodology: DBRNet

The authors propose the **Disentangled and Balanced Representation Network (DBRNet)**. The core philosophy is to disentangle covariates into three distinct latent factors and apply selection bias adjustment only where theoretically necessary.

A. Disentangled Latent Factors

- **Instrumental Factors ($\Gamma(x)$):** Affect Treatment (T) but not Outcome (Y).
- **Confounder Factors ($\Delta(x)$):** Affect both Treatment (T) and Outcome (Y).
- **Adjustment Factors ($\Upsilon(x)$):** Affect Outcome (Y) but not Treatment (T).

B. Model Architecture

- **Encoders:** Three separate neural networks encode input X into Γ, Δ, and Υ.
- **Treatment Branch:** Concatenates Γ and Δ to estimate the conditional density: $P(t_i \mid \Gamma(x_i), \Delta(x_i))$.
- **Outcome Branch:** Concatenates Δ and Υ to predict Y using a **Varying Coefficient Network (VCNet)**.

C. Objective Function & Loss Terms

The total loss is: $J = w \cdot L_y + \alpha L_T + \beta L_{disc} + \gamma L_{ind} + \lambda L_{reg}$.

- **Factual Loss (L_y):** Prediction error for the observed outcome.
- **Treatment Loss (L_T):** Negative log-likelihood of the treatment density.
- **Discrepancy Loss (L_{disc}):** Forces Γ, Δ, Υ to be distinct/independent.
- **Independent Loss (L_{ind}):** Forces Υ to contain **minimal** treatment info.

- **Re-weighting (w):** Uses inverse propensity score from relevant factors only:

$$w(t_i, \Gamma, \Delta) = \frac{1}{P(t_i \mid \Gamma(x_i), \Delta(x_i))}$$

3. Key Contributions

- **First Disentanglement for Continuous Treatment:** DBRNet combines disentangled representation learning with precise selection bias adjustment for continuous settings.
- **Theoretical Bias Elimination:** Theorem 2 proves that re-weighting based on Instrumental and Confounder factors yields an unbiased estimation of the IDRF loss.
- **Selective Balancing:** Enforces independence via L_{ind} and L_{disc} rather than aggressive balancing, preserving predictive power.

4. Main Results & Experiments
Evaluated on **Synthetic**, **IHDP**, and **News** datasets using Mean Integrated Squared Error (MISE) and Average MSE (AMSE).
Quantitative Performance

- DBRNet outperformed baselines (Dragonet, DRNet, VCNet, TransTEE) across almost all metrics.
- *Synthetic MISE:* DBRNet **0.1414** vs. VCNet 0.1752.
- *IHDP MISE:* DBRNet **1.78** vs. VCNet 2.33 vs. TransTEE 3.30.

Ablation & Analysis

- **Component Importance:** The Re-weighting function and Discrepancy Loss (L_{disc}) are critical; removing re-weighting increased errors significantly (e.g., +200% AMSE on IHDP).
- **Verification:** t-SNE visualizations confirm effective separation of factors.

Historian Agent Output

Domain Narrative: Individualized Continuous Treatment Effect Estimation

1. Domain History: From Binning to Disentanglement
Five years ago, the dominant paradigm in Causal Inference was restricted to **binary treatments** (treated vs. control). The foundational work by **Shalit et al. (2017)** established **Representation Learning**—specifically balancing covariate distributions between groups—as the standard for handling selection bias.
The shift toward **Continuous Treatment Effects (Dose-Response)** began in earnest around 2020.

- **The Binning Era ($\sim$2020):** Approaches like **DRNet (Schwab et al., 2020)** discretized continuous treatments into bins (strata). This solved architectural challenges but sacrificed response curve smoothness.
- **The Continuity Era (2021): Nie et al. (2021)** introduced **VCNet**, replacing discretization with Varying Coefficient networks to enforce functional smoothness, establishing continuity as a requirement.
- **The Architectural & Disentanglement Era (2022–2024):** The field has shifted from global balancing (which risks destroying info) to **Disentangled Representations**. Methods like **DTRNet (2024)** and **ACFR (2023)** surgically separate instrumental, confounding, and adjustment factors. Simultaneously, Transformers (**TransTEE, 2022**) have replaced simple feed-forward networks to better capture interactions.

2. Open Problems & Gaps
Despite recent progress, distinct "unsolved" territories remain:

1. **High-Dimensional & Unstructured Confounding:** Most SOTA methods (VCNet, TransTEE) are benchmarked on low-dimensional tabular data. Handling images or text as confounders remains theoretically sparse, though **StoNet** and **CausalDiffAE (2024)** are making attempts.
2. **Generative Stability vs. Regression Precision:** Generative approaches (GANs, Diffusion) offer high-fidelity counterfactuals but suffer from training instability. Bridging this gap with the precision of VCNet is an open challenge.
3. **Data Scarcity & Pre-training: CURE (2024)** highlighted the potential of Foundation Models, but this is under-explored due to the scarcity of large-scale biomedical datasets for continuous interventions.

3. Significance Criteria (2025 Era)
For a new contribution to be deemed "Significant", it must go beyond marginal MISE improvements on the IHDP benchmark.

- **Low Significance:** Another MLP-based architecture that slightly beats DRNet on tabular data using standard re-weighting.

- **High Significance:**
  - **Metric Innovation:** Rigorous handling of **high-dimensional unstructured data** (e.g., images) as confounders.
  - **Theoretical Disentanglement:** Provable separation of instrumental variables from confounders without rigid functional forms.
  - **Efficiency:** Architectures achieving Transformer-level performance with the inference speed of simple regressors.
  - **Foundation Approaches:** Transfer learning or pre-training mechanisms reducing reliance on labeled counterfactual data.

Baseline Scout Agent Output

1. Missing Critical Baselines
The following key methods are absent from the evaluation, potentially inflating the perceived relative performance of the proposed method:

- **SCIGAN** (Bica et al., NeurIPS 2020)
  - *Why it matters:* This is a seminal SOTA method for continuous treatments. The authors discuss it in the text but exclude it from the main results table, avoiding a direct comparison.
- **End-to-End Balancing** (Bahadori et al., ICML 2022)
  - *Why it matters:* Represents the state-of-the-art for weighting/balancing methods in continuous settings. Its omission is critical as it directly challenges the authors' proposed re-weighting function.
- **Double Machine Learning (DML)** (Chernozhukov et al., 2018)
  - *Why it matters:* The gold-standard statistical baseline for continuous treatment effects. It is robust and often outperforms the Causal Forest baseline which was included by the authors.

2. Missing Standard Benchmarks
The evaluation relies heavily on synthetic or low-dimensional data, omitting standard high-dimensional or real-world benchmarks:

- **TCGA (The Cancer Genome Atlas)** (Schwab et al., AAAI 2020)
  - *Significance:* The standard high-dimensional benchmark (20k+ features) for continuous dosage. Its omission hides potential scalability issues of the complex disentanglement module.
- **MVICU (Mechanical Ventilation in ICU)** (Bica et al., NeurIPS 2020)
  - *Significance:* A critical real-world medical dataset for continuous treatments used by main competitors (DRNet, SCIGAN). Essential for validating claims about precision medicine applications.

Q&A Agent Output (Summary of first five Q&A pairs presented for readability)

Part 1: Technical Soundness and Quality

Q: The paper relies on a heuristic argument of 'mutual gradient perturbation'. Can you provide a formal theoretical guarantee (like Marvell's KL-divergence bound) that ensures leakage mitigation?
A: No formal theoretical guarantee is provided. The authors acknowledge in Section 5 that "GAFM lacks the rigor of Marvell's upper bound." Instead, they rely on a heuristic argument supported by empirical validation.
They demonstrate that gradients derived from the GAN loss and the randomized Cross-Entropy loss tend to have class centers that differ in opposite directions. This "mutual gradient perturbation" empirically results in mixed intermediate gradients, but the paper does not mathematically derive a worst-case privacy loss bound.

Q: Given the 'Defense-Efficiency Paradox', how does the computational overhead of GAFM compare to Marvell for resource-constrained clients?
A: The overhead does not negatively impact clients because the Generator and Discriminator are exclusively hosted by the Label Party (Server).
Non-label parties (clients) only train local feature extractors (f_{θ_p}), keeping their load comparable to vanilla Split Learning. While Marvell also centralizes defense, it relies on costly iterative optimization to generate noise; GAFM offloads this to simple "1-layer DNNs" on the server, effectively circumventing the Defense-Efficiency Paradox for the client.

Q: Why were state-of-the-art non-perturbation defenses like DiscreteSGD or CoAE omitted from the baselines?

A: The authors likely frame GAFM specifically as an evolution within the gradient perturbation paradigm. They benchmark against Marvell and Max Norm (noise injection methods) rather than quantization (DiscreteSGD) or compression (CoAE) methods. The domain narrative suggests they viewed Marvell as the singular SOTA for perturbation, treating concurrent 2022 works like CoAE as parallel rather than foundational.

Q: How does the 'mutual gradient perturbation' generalize to multi-class classification where gradients are not diametrically opposed?
A: The defense would likely collapse in multi-class settings.
The mechanism relies on the geometric simplicity of binary classification, where gradients naturally cluster in diametrically opposed directions (1D manifold). In a multi-class simplex (K classes), the concept of "opposing" forces is ill-defined, and the current randomized response mechanism lacks the directional guidance to confuse specific class boundaries in high-dimensional space.

Q: How sensitive is the privacy-utility trade-off to the hyperparameter γ (balancing GAN/CE losses)?
A: It is highly sensitive, with a confirmed failure mode. If γ is improper (too low), the Cross-Entropy loss dominates, canceling the perturbation effect and re-introducing leakage. The authors admit tuning is challenging due to GAN instability but propose gradient normalization as a heuristic fix.

Part 2: Novelty and Significance

Q: Does the integration of GANs specifically for label protection in Split Learning appear in prior art?
A: High Novelty. While GANs exist in Federated Learning (e.g., FedCG for Horizontal FL, or for attacking/reconstruction), using a GAN within the SplitNN architecture specifically to create "mutual gradient perturbation" for label protection is distinct.

- *FedCG (IJCAI 2022):* Uses GANs in Horizontal FL for data augmentation/privacy.
- *VertiGAN (PoPETS 2023):* Uses GANs to publish synthetic data, not for active training protection.

GAFM is significant because it addresses the "Defense-Efficiency Paradox" with a stable architecture rather than unstable optimization (Marvell).

Q: Has the combination of Wasserstein GAN loss and Randomized Cross-Entropy loss been used in VFL defenses before?
A: Highly Novel. This specific combination to achieve mutual cancellation is not found in prior art (pre-Aug 2023). Previous methods used Randomized Response for Differential Privacy (LabelDP) or GANs for synthesis, but not integrated into a single loss function to obfuscate gradients in VFL.

Q: Does 'Mutual Gradient Perturbation' differ theoretically from standard noise injection (Marvell)?
A: Medium to High Novelty. Yes. Marvell applies external, optimized noise. GAFM relies on an intrinsic cancellation arising from conflicting objectives (GAN vs. CE). This is a theoretical shift from "adding noise" to "structural obfuscation."

Q: Do the proposed 'Mean' and 'Median' attacks offer novelty over existing clustering attacks?
A: Incremental Novelty / Medium Significance. The concept of clustering gradients (e.g., Liu et al., 2022) existed. Mean/Median attacks are specific heuristics within that framework. However, they are significant as they provide a stronger benchmark than the standard Norm attack, raising the bar for future defense evaluation.

Q: Is the method applicable to complex architectures like Vision Transformers (ViTs)?
A: Limited Applicability. The paper strictly limits scope to binary classification. There is no evidence or design adaptation for ViTs, which have unique leakage vulnerabilities (attention mechanisms). Given the geometric reliance on opposing gradients, applicability to complex non-CNN architectures is unproven.

F Human Evaluation Guidelines

We provided the following instructions to our expert human annotators to ensure a rigorous and standardized evaluation process.

F.1 Objective

We are evaluating the performance of two Agentic Reviewer Systems (`ai_review1` and `ai_review2`). The goal is not just to check if the AI generates coherent text, but to determine if it provides **value-add** over expert human reviewers. You will compare AI reviews against the **Human Reviews** (actual reviews submitted for the paper).

F.2 The Evaluation Task

For each assigned paper, you will be presented with:

1. **The Paper:** (PDF/Text).

2. **Human Reviews:** The set of actual reviews the paper received.
3. **AI Review 1 & AI Review 2:** Reviews generated by two different agentic frameworks.

Your Workflow:

1. **Scan the Paper:** Understand the core contribution, methodology, and claims.
2. **Read Human Reviews:** Establish the "Expert Baseline." Note what the humans caught and what they might have missed.
3. **Evaluate AI Reviews (Individually):** Score both AI reviews on a 1-10 scale against the Human Baseline.
4. **Side-by-Side (SxS) Comparison:** Determine which of the two AI assistants performed better.

F.3 Part I: Individual Scoring (Scale 1-10)

For each AI review, you will assign a score based on how it compares to the **best human review** available for that paper.

The Scoring Rubric: Table 8 provides the scoring rubric details.

Table 8: Scoring Rubric for H-Max Score.

Score	Label	Definition
9-10	Transformative / Superhuman	The AI uncovers a critical flaw, a missing theoretical connection, or vital prior work that *all* human reviewers missed. It fundamentally improves the critique.
7-8	Clearly Superior	The AI is more thorough, constructive, or better substantiated than the best human review. It may offer deeper questions or better literature context.
6	Slightly Better	The AI review is slightly more polished or covers one extra minor point compared to the best human review.
5	Equivalent / Human Level	The AI review is roughly equivalent in quality to the best human review. It covers the same major points with similar depth.
3-4	Slightly Worse	The AI review is valid but less nuanced or specific than the best human review.
1-2	Significantly Worse	The AI misses critical points, hallucinates details, or is superficial compared to humans.

Evaluation Dimensions:

1. **Technical Accuracy:** Are the AI's claims factually correct regarding the paper's content? Look for hallucinations vs. valid technical critique.
2. **Constructive Value:** How actionable is the feedback? Look for specific suggestions (e.g., "Run experiment X on dataset Y") vs. generic advice.
3. **Analytical Depth:** Does the review engage with the substance of the work? Look for deep questioning of assumptions vs. surface-level comments.
4. **Novelty and Significance Assessment:** Did the AI correctly identify the novelty?
   - *Critical Instruction:* You may use Google Search/Scholar to verify claims, BUT you must strictly ignore any papers published **after the paper's submission date**.
   - *Score High if:* The AI cites specific prior work that limits the paper's novelty, which humans missed.

- *Score Low if:* The AI claims "high novelty" for a derivative work or hallucinates citations.

5. **Overall Score:** A holistic assessment of the AI's utility.

F.4 Part II: Side-by-Side (SxS) Comparison

After scoring individually, compare `AI Review 1` vs. `AI Review 2` directly.

Options:

- **Review 1** (If Review 1 is clearly better)
- **Review 2** (If Review 2 is clearly better)
- **Tie** (If both are of similar quality)

Dimensions for SxS:

- **Technical Accuracy:** Which assistant made fewer errors and grounded its claims better?
- **Constructive Value:** Which assistant gave more helpful advice?
- **Analytical Depth:** Which assistant probed deeper into the methodology?
- **Novelty & Significance:** Which assistant better identified the paper's place in the literature?
- **Overall:** Which assistant would you prefer to have reviewing your own paper?

F.5 Important Notes

- **Hallucinations:** If an AI review hallucinates a major detail (e.g., claims the paper used a Transformer when it used an RNN), the Technical Accuracy score should be low, regardless of how well-written it is.
- **Bias Check:** Do not penalize the AI simply for being "stricter" or "nicer" than the humans. Judge based on the validity of the critique.
- **Presentation:** Do not judge based on the "prettiness" of the review format—evaluate the content.
- **Anonymity:** The interface effectively blinds `ai_review1` and `ai_review2`. Do not assume one slot is always the same model.

G Agent Prompts

We provide the system and user prompt templates used for each agent in the SCHOLARPEER .

G.1 Summarizer Agent

Summarizer Prompt

You are an expert AI researcher specializing in distilling the core contributions of a research paper. Your task is to produce a concise and accurate summary.

Please summarize the following research paper. Focus on the key contributions, methodology, and main results. The summary should be dense with information and serve as a reliable reference to analyze the paper.

Paper Text:
{paper_text}

G.2 Literature Review Agent

Literature Reviewer Prompt

You are an expert AI literature reviewer. Your goal is to perform a comprehensive, deep−dive literature search to map the state of a specific research domain up to a specific point in time.

You do not critique papers. You collect, organize, and extract facts about them to support downstream expert analysis.

Please conduct a comprehensive literature review for the research paper described below.

CRITICAL CONSTRAINT: You must only search for and consider prior art published ON OR BEFORE the cutoff date: {cutoff_date}. Ignore any papers published after this date.

Your Process:

1. Domain Analysis: Analyze the input paper abstract to identify:
- ∗ The Broad Domain (e.g., Computer Vision).
- ∗ The Specific Sub−field (e.g., Weakly Supervised Object Detection).
- ∗ The Core Problem being solved.

2. Search Execution (Iterative):
Use Google Search to find 30−50 of the most scientifically significant papers in this sub−field. You must look for:
- ∗ Foundational Papers: The papers that established the current paradigms (even if older).
- ∗ Key Datasets: Papers introducing the primary datasets used in this sub−field, or specific datasets used by the input paper.
- ∗ Recent SOTA: The papers holding the State−of−the−Art results at the time of the cutoff date {cutoff_date}.
- ∗ Direct Competitors: Papers solving the same problem with different methods.
- ∗ Surveys: Recent survey papers that summarize the field.

3. Data Extraction:
- ∗ For each identified paper, extract the specific details required for our records (see output format below).
- ∗ For core_method, provide a specific description of their technical approach or dataset (few sentences).
- ∗ For datasets_and_performance, be as specific as possible about what was evaluated and the performance numbers (e.g., ”Achieved 78.4% top−1 accuracy on ImageNet, outperforming FE−Net (72.9%)”, not just ”showed significant improvements in top−1 accuracy on ImageNet”).
- ∗ No citation artifacts: Ensure that core_method and datasets_and_performance do not contain bracketed citation numbers (e.g., [12], [34]) as they have no meaning without the referenced paper/link.
- ∗ Survey handling: For survey papers, list the specific families of methods, models and datasets they cover in the core_method and datasets_and_performance fields.

Output Format:
Respond only with a JSON object containing the ”Domain Analysis” and the ”References” list.

```json
{{
  ”domain_analysis”: {{
    ”broad_domain”: ”...”,
    ”specific_subfield”: ”...”,
    ”primary_datasets_commonly_used”: [”...”, ”...”],
    ”standard_metrics_used”: [”...”, ”...”]
  }},
  ”references”: [
    {{
      ”title”: ”Title of Paper 1”,
      ”venue_year”: ”CVPR 2023”,
      ”authors”: ”Author A, Author B”,
      ”is_foundational”: true/false,
      ”is_sota_candidate”: true/false,
      ”is_dataset_paper”: true/false,
      ”is_survey_paper”: true/false,
      ”core_method”: ”Specific description of their technical approach or dataset (few sentences).”,
      ”datasets_and_performance”: ”Specifics on what they evaluated (e.g., ’Achieved 78.4% top−1 accuracy on ImageNet’).”,
      ”known_limitations”: ”Any widely known weaknesses of this method or dataset.”
    }},
```

```
    {{
      "title": "Title of Paper 2",
      "..." : "..."
    }}
  ]
}}
```

Input Paper Title, Author(s) and Abstract:
{paper_abstract}

G.3 Literature Expansion Agent

Literature Expander Prompts

You are a Senior Research Lead auditing a literature review. Your goal is to identify gaps in a list of references and perform targeted searches to fill them.

We have a preliminary literature review for the domain defined below. Your job is to identify what is missing and find it.

Task:
1. Analyze Gaps:
   * Do we have the foundational papers that the current SOTA papers likely cite? (e.g., If we have 'Crossformer', do we have the original 'Transformer' or 'Autoformer' papers?)
   * Do we have the papers that introduced the datasets listed in the domain_analysis?
   * Are there temporal gaps? (e.g., We have 2018 and 2024, but nothing from 2020–2023).

2. Targeted Search: Perform specific searches to find these missing papers (Constraint: published ON OR BEFORE {cutoff_date}).

3. Output: Return a list of new unique papers to add to the reference list. Do not repeat existing papers.

Current Reference List:
```json
{current_references_json}
```

Output Format:
Respond only with a JSON list of new reference objects (using the same schema as the input).

```
[
  {{
    "title": "Autoformer: Decomposition Transformers for Long–Term Series Forecasting",
    "venue_year": "NeurIPS 2021",
    "authors": "Haixu Wu, et al.",
    "is_foundational": true/false,
    "is_sota_candidate": true/false,
    "is_dataset_paper": true/false,
    "is_survey_paper": true/false,
    "core_method": "Specific description of their technical approach or dataset (few sentences).",
    "datasets_and_metrics": "Specifics on what they evaluated (e.g., 'Achieved 78.4% top–1 accuracy on ImageNet').",
    "known_limitations": "Any widely known weaknesses of this method or dataset."
  }}
]
```

G.4 Sub-Domain Historian Agent

Historian Prompts

You are a Senior Science Historian. Your job is to synthesize the evolutionary arc of a research field.
You care about the narrative of progress: how ideas evolved, what the current paradigms are, and what "Open Problems" remain unsolved.

Based on the provided literature review, write a "Domain Narrative" for the "Novelty & Significance Reviewer".

Input Data:

Full Literature Review JSON:
{literature_review_json}

Task:
1. **The Arc of Progress:** Briefly trace the history. What was the dominant paradigm 5 years ago? What paper changed it? Where are we now?
2. **Gap Analysis:** Based on the most recent papers (closest to the cutoff date), what problems are explicitly stated as unsolved or limitations?
3. **Significance Rubric:** Define what constitutes a "Significant" contribution in this specific moment. (e.g., "In 2025, small MSE improvements are less significant than efficiency gains or interpretability.")

Output Format:
Respond with a concise Markdown summary (approx. 300–400 words) containing the sections a) Domain History (chronological summary of key shifts), b) Open Problems, c) Significance Criteria.

G.5 Baseline Scout Agent

Baseline Scout Prompts

You are a ferocious benchmarking expert. Your sole job is to find what the authors are HIDING.
You must find state–of–the–art (SOTA) methods and baselines that the authors *should* have compared against but didn 't.

Paper:
{paper_text}

YOUR TASK:
1. Based on the paper, identify the research domain, the datasets used, and the baseline methods compared.
2. Search for recent (last 3 years, before {cutoff_date}) SOTA methods for this domain.
3. Identify specific methods and significant datasets that are MISSING from the authors' list.
4. Return a list of these missing competitors and why they are relevant.

Constraint: Only return papers published ON OR BEFORE {cutoff_date}.

Output JSON format:
```
{{
  "missing_baselines": [
    {{
      "name": "Method Name",
      "reference": "Author et al., Conf Year",
      "reason": "Why it is a critical omission."
    }},
  "missing_datasets": [
    {{
      "name": "Dataset Name",
      "reference": "Author et al., Conf Year",
      "reason": "Why it is a critical omission."
    }}
  ]
}}
```

G.6 Question Generator Agent

Novelty Aspect Prompts

System Prompt:

You are a highly–discerning AI researcher and top–tier conference reviewer. Your goal is to deconstruct a paper's contributions and formulate simple, direct questions to assess its novelty and significance.

User Template:

Please follow this two–step process to generate your questions:

1. Identify Contribution Claims: First, carefully read the Paper Summary and Paper Text to identify the {num_questions} primary contribution claims. A contribution might be a new model, a new dataset, a new algorithm, a new theoretical insight, or a new application. A domain narrative is also provided to give context on the evolution of the field and open problems. An independent literature review is also provided to give you a comprehensive view of prior art. Missing baselines and datasets are also provided for better domain understanding.
2. Formulate Simple Questions (one per claim): For each claim you identified, formulate one simple and direct question that assesses its novelty and significance. This question is a directive for a research assistant (who does not have the paper) to find conflicting or related prior art.

Example Claim: The paper uses 'emotional prompts' to scale reasoning in LLMs.
Example Question: "What is the novelty and significance of using 'emotional prompts' to improve reasoning in language models?"

Return the questions in a JSON list format, like this:
["Question investigating claim 1?", "Question investigating claim 2?", "Question investigating claim 3?"]

Domain Narrative:
{domain_narrative}

Independent Literature Review:
{literature_review}

Missing Baselines and Datasets:
{missing_baselines_datasets}

Paper Summary:
{summary}

Paper Text:
{paper_text}

Other Aspects (Soundness, Clarity)

System Prompt:

You are a critical AI researcher and conference reviewer. Your goal is to ask probing questions about a paper to assess its quality along a specific dimension. A domain narrative is also provided to give context on the evolution of the field and open problems. An independent literature review is also provided to give you a comprehensive view of prior art. Missing baselines and datasets are also provided for better domain understanding.

Based on the paper's summary and the full text, generate a list of {num_questions} critical questions to evaluate its {aspect}.

These questions should be specific and designed to be answered based on the paper content, searching the web or academic databases.

User Template:

Return the questions in a JSON list format, like this:
["question 1?", "question 2?", "question 3?"]

Domain Narrative:
{domain_narrative}

Independent Literature Review:
{literature_review}

Missing Baselines and Datasets:
{missing_baselines_datasets}

Paper Summary:
{summary}

Paper Text:
{paper_text}

G.7 Answer Generator Agent

Novelty Aspect (Search Enabled)

System Prompt:

You are an AI researcher with access to Google search. Your task is to critically assess the novelty and significance of a research claim of a research paper based on external search results.

User Template:

Please perform a comprehensive literature search to answer the following question. Domain narrative, an independent literature review and missing baselines and datasets are also provided to help assess novelty and significance. The paper summary is provided for context.

Constraint: The research paper was released on {cutoff_date}. You must only search for and consider prior art published ON OR BEFORE the cutoff date: {cutoff_date}. Do not include any papers published after this date.

Follow these steps:
1. Analyze Context: Read the question, the paper summary, domain narrative, independent literature review, and missing baselines and datasets to identify the key domain (e.g., Computer Vision, NLP, Robotics) and placement of the current work in the domain's progress.
2. Identify Venues: Based on the domain, determine the most relevant top−tier conferences to search (e.g., CVPR, ICCV, ECCV for Vision; NeurIPS, ICLR, ICML, ACL, EMNLP for NLP/ML).
3. Execute Search: Use your search tools to find relevant prior art from these conferences and arXiv, ensuring all results were published on or before {cutoff_date}.
4. Assess Novelty: Based on your search, the domain narrative and independent literature review, determine if the claim is new, incremental, or a known concept.
5. Assess Significance: Based on the problem's importance and the findings of related work, assess the potential impact of this contribution. Is it a niche problem or a major one? Are improvements likely to be large or small?
6. Synthesize Findings: Summarize the findings from your search to provide a well−reasoned answer to the question.

Answer format:
1. A direct, paragraph−style answer to the question. Ensure to include a) degree of novelty (high, medium, low, incremental, none), b) degree of significance (high, medium, low, none). Include reasoning for your assessments.
2. A bulleted list of 2−3 most relevant papers that support your answer. For each of these papers, provide a) title, b) authors, c) venue, d) year, e) key findings.

If you find no significant prior art, rate novelty as High and justify the significance assessment.

Domain Narrative:
{domain_narrative}

Independent Literature Review:
{literature_review}

Missing Baselines and Datasets:
{missing_baselines_datasets}

Paper Summary:
{summary}

Question:
{question}

Other Aspects

System Prompt:

You are an AI researcher with deep insightful thinking. Your task is to answer questions about a research paper by reasoning based on information provided in the paper. A domain narrative is also provided to give context on the evolution of the field and open problems. An independent literature review is also provided to give you a comprehensive view of prior art. Missing baselines and datasets are also provided for better domain understanding.

User Template:

Please answer the following question based on the provided paper context. Provide a concise answer. Provide your answer as a single, well−reasoned paragraph.

Question:
{question}

Paper Summary:
{summary}

Domain Narrative:
{domain_narrative}

Independent Literature Review:
{literature_review}

Missing Baselines and Datasets:
{missing_baselines_datasets}

Paper Text:
{paper_text}

G.8 Review Generator Agent

Review Generator Prompts

You are an AI researcher who is reviewing a paper that was submitted to a prestigious ML venue. Be critical and cautious in your decision. If a paper is bad or you are unsure, give it bad scores and reject it. Use the provided paper summary and a list of question−answer pairs generated by an AI agent to inform your assessment.

{review_guidelines}

Here is the AI summary of the paper you are asked to review:
‘‘‘
{summary}
‘‘‘

Here is the list of question−answer pairs generated by an AI agent to help you review the paper:
‘‘‘
{qa_pairs_text}
‘‘‘

Here is the paper you are asked to review:
‘‘‘
{paper_text}
‘‘‘

{fewshot_examples}

H Evaluation Prompts

We provide the exact prompts used for our automated evaluation protocols.

H.1 H-Max Score

Used to score individual reviews against human expert baselines.

H-Max Score Estimation Prompts

System Prompt Template:

You are an expert area chair evaluating an AI Reviewer Assistant. Your role is to determine if the AI provides value beyond what expert human reviewers provided.

Special Instruction for evaluating "Novelty and Significance Assessment": You must only search for and consider information available on or before the cutoff date: {cutoff_date}. The cutoff date represents the date on which the paper was published; information after this date is irrelevant to the review.

Follow the steps below for each evaluation:

1. Thoroughly understand the paper by analyzing:
   – Research objectives and contributions
   – Methodology and experiments
   – Claims and evidence
   – Results and conclusions

2. Identify the strongest points in the Human Reviews (collectively) to establish a standard expert baseline.

3. Identify the delta: What did the AI mention that the humans missed? What did the humans mention that the AI missed ?

4. Verify the validity of each of the delta claims using direct quotes from the paper and external sources (for novelty and significance only).

5. Assess the value−add of the AI review compared to the best human review for each aspect.

You will evaluate reviews based on these key aspects:

∗∗Technical Accuracy∗∗
– How technically accurate is the AI review compared to the humans?

∗∗Constructive Value∗∗
– How actionable is the feedback compared to the humans?

∗∗Analytical Depth∗∗
– How thorough is the depth of AI review compared to humans?

∗∗ Novelty and Significance Assessment (Search encouraged)∗∗
Use Google Search to actively verify the reviewers' claims about novelty and significance.
1. Identify claims: What do the paper and reviewers claim is novel?
2. Formulate search queries: Create targeted queries to find relevant prior work for these specific claims, explicitly restricting results to before {cutoff_date}.
3. Execute Search: Focus on top−tier conferences in the relevant domain and arXiv.
4. Verify and Compare:
   – Did the AI find prior work that limits novelty which the humans missed? (High Score)
   – Did the AI claim "high novelty" when humans correctly identified it as derivative work? (Low Score)
   – Which assessment aligns better with the actual state of the field at the time?
5. Cite Sources: You must cite the specific external papers (title, venue, year) you used to make this determination in the JSON output.

For each of the above aspects and overall judgment, you must:
1. Provide specific evidence from source materials
2. Quote directly from paper and reviews; external sources only for "Novelty and Significance Assessment"
3. Explain your reasoning in detail
4. Consider alternative interpretations

∗∗Input Format:∗∗
Paper Text:
<Paper text>

AI Assistant's Review:
<AI Review>

Human Reviews (Ground Truth):
<Human Reviews>

∗∗Respond in the following format:∗∗
THOUGHT:
<THOUGHT>

EVALUATION JSON:
‘‘‘json
<JSON>

In <THOUGHT>, for each aspect, compare the AI Assistant’s review against the set of Human Reviews. Identify the best human review for that specific aspect and use it as your baseline (Score = 5), as a standard expert. You must justify why the AI deserves a higher or lower score based on the ”Value−Add” it provides.

Scoring Rubric (Compare against Best Human Baseline):
– 10 (Superhuman / Verdict−Changing): The AI uncovers a critical insight that changes the fate of the paper (e.g., finding a fatal math error humans missed OR identifying a profound theoretical connection that elevates a rejected paper to an acceptance).
– 9 (Transformative / Insightful): The AI provides a novel perspective that significantly reframes the paper’s contribution . It might articulate the significance better than the authors did, or identify a missing baseline that reframes the results.
– 8 (Clearly Superior): The AI review is significantly more thorough, constructive, or better substantiated than the best human review. It offers deep questions or literature context that humans omitted.
– 7 (Superior): The AI review is noticeably deeper and more constructive than the best human review, though perhaps not ”transformative.”
– 6 (Slightly Better): The AI review is slightly more polished, better structured, or covers one extra minor point compared to the best human review.
– 5 (Equivalent / Human Level): The AI review is roughly equivalent in quality to the best human review. It covers the same major points with similar depth.
– 4 (Slightly Worse): The AI review is valid but generic. It misses the specific nuance or ”sharpness” that the expert human provided.
– 3 (Worse): The AI review is valid but vague. It lacks detail and actionable feedback compared to the human (e.g., ” Improve experiments” vs ”Add dataset X”).
– 2 (Failure − Superficial): The review is technically ”safe” (no direct lies) but functionally useless. It is too short, focuses only on trivial formatting issues, or completely misses the core technical innovation.
– 1 (Failure − Critical Error/Hallucination): The AI makes a factual error about the paper (e.g., claims it uses Method A when it uses Method B) or cites non−existent papers. The review actively misleads the reader.

In <JSON>, provide the evaluation in JSON format with the following fields in the order:
– ”Technical Accuracy Reason”: ”<detailed reason>”.
– ”Technical Accuracy Score”: <int 1−10>.
– ”Constructive Value Reason”: ”<detailed reason>”.
– ”Constructive Value Score”: <int 1−10>.
– ”Analytical Depth Reason”: ”<detailed reason>”.
– ”Analytical Depth Score”: <int 1−10>.
– ”Novelty and Significance Assessment External Sources Used”: List of retrieved papers (include title, venue, year and authors for each paper).
– ”Novelty and Significance Assessment Reason”: ”<detailed reason>”.
– ”Novelty and Significance Assessment Score”: <int 1−10>.
– ”Overall Reason”: ”<detailed reason>”.
– ”Overall Score”: <int 1−10>.

This JSON will be automatically parsed, so ensure the format is precise and scores are integers.

User Prompt Template:

Paper Text:
{paper_text}

AI Assistant’s Review:
{ai_review}

Human Reviews:
{human_review}

H.2 Side-by-Side (SxS) Evaluation

Used for pairwise preference ranking.

SxS Evaluation Prompts

System Prompt Template:

[You are a neutral arbitrator evaluating peer review comments for academic papers. Your role is to analyze and compare reviews through careful, evidence−based assessment. Your judgments must be strictly based on verifiable evidence from the paper, reviews and Google search. Google search should be used only for evaluating novelty and significance assessment. Do not use it for other dimensions.

Special Instruction for evaluating "Novelty and Significance Assessment": You must only search for and consider information available on or before the cutoff date: {cutoff_date}. The cutoff date represents the date on which the paper was published; information after this date is irrelevant to the review.

For each evaluation, you must:

1. Thoroughly understand the paper by analyzing:
   – Research objectives and contributions
   – Methodology and experiments
   – Claims and evidence
   – Results and conclusions

2. For each review, methodically examine:
   – Claims made about the paper
   – Evidence cited to support claims
   – Technical assessments and critiques
   – Suggested improvements

3. Compare reviews systematically using:
   – Direct quotes from paper and reviews
   – Specific examples and counterexamples
   – Clear reasoning chains
   – Objective quality metrics

You will evaluate reviews based on these key aspects:

∗∗Technical Accuracy∗∗
– Are claims consistent with paper content?
– Is evidence properly interpreted?
– Are technical assessments valid?
– Are critiques well−supported?

∗∗Constructive Value∗∗
– How actionable is the feedback?
– Are suggestions specific and feasible?
– Is criticism balanced with strengths?
– Would authors understand how to improve?

∗∗Analytical Depth∗∗
– How thoroughly are key aspects examined?
– Is analysis appropriately detailed?
– Are important elements addressed?
– Is assessment comprehensive?

∗∗ Novelty and Significance Assessment (Seach encouraged)∗∗
Use Google Search to actively verify the reviewers' claims about novelty and significance.
1. Identify claims: What do the paper and reviewers claim is novel?
2. Formulate search queries: Create targeted queries to find relevant prior work for these specific claims, explicitly restricting results to before {cutoff_date}.
3. Execute Search: Focus on top−tier conferences in the relevant domain and arXiv.
4. Verify and Compare:
   – Did reviewer A or reviewer B miss a critical prior work that you found?
   – Did reviewer A or reviewer B accurately identify that a novel claim is actually a known technique?
   – Which reviewer's assessment of significance aligns better with the actual state of the field at the time?
5. Cite Sources: You must cite the specific external papers (title, venue, year) you used to make this determination in the JSON output.

For each of the above aspects and overall judgment, you must:

1. Provide specific evidence from source materials
2. Quote directly from paper and reviews; external sources only for ”Novelty and Significance Assessment”
3. Explain your reasoning in detail
4. Consider alternative interpretations

Input Format:
Paper Text:
<Paper text>

Assistant A’s Review:
<Review A>

Assistant B’s Review:
<Review B>

Respond in the following format:
THOUGHT:
<THOUGHT>

REVIEW COMPARISON JSON:
```json
<JSON>
```

In <THOUGHT>, for each aspect, evaluate assistants A and B based on the above criteria followed by a comparative assessment.
Treat this as the note−taking phase of your evaluation.

In <JSON>, provide the review in JSON format with the following fields in the order:
– ”Technical Accuracy Reason”: ”<detailed reason>”.
– ”Technical Accuracy Better Assistant”: ”<A/B/Tie>”.
– ”Constructive Value Reason”: ”<detailed reason>”.
– ”Constructive Value Better Assistant”: ”<A/B/Tie>”.
– ”Analytical Depth Reason”: ”<detailed reason>”.
– ”Analytical Depth Better Assistant”: ”<A/B/Tie>”.
– ”Novelty and Significance Assessment External Sources Used”: List of retrieved papers (include title, venue, year and authors for each paper).
– ”Novelty and Significance Assessment Reason”: ”<detailed reason>”.
– ”Novelty and Significance Assessment Better Assistant”: ”<A/B/Tie>”.
– ”Overall Reason”: ”<detailed reason>”.
– ”Overall Better Assistant”: ”<A/B/Tie>”.

This JSON will be automatically parsed, so ensure the format is precise.

User Prompt Template:

Paper Text:
{paper_text}

Assistant A’s Review:
{review_a}

Assistant B’s Review:
{review_b}

H.3 Summary of Gains Analysis

Used to synthesize qualitative insights from SxS reasoning traces.

Summary of Gains Prompts

System Prompt:

You are an expert Meta−Reviewer for AI Science. You are tasked with analyzing a set of ”Side−by−Side” (SxS) evaluation logs comparing two AI Research Assistants: {model_a} and {model_b}.

Your Goal: Synthesize a high−level qualitative report on why one model outperforms the other.

Input Data:
You will receive a list of {num_samples} evaluation traces. In each trace, a judge has reasoned about which assistant wrote a better review for a scientific paper.

Instructions:
1. **Identify Patterns**: Do not just list specific examples. Look for recurring themes across the traces (e.g., "Model A consistently hallucinates citations" or "Model B misses novelty discussions").
2. **Cluster & Categorize**: Group similar observations into distinct categories (e.g., "Technical Depth", "Tone", "Constructiveness").
3. **Attribution**:
   - "Gains" are positive traits where the model performed *better* than the competitor.
   - "Losses" are negative traits where the model performed *worse*.

Output Format:
Return a JSON object with the following structure. Limit to the top 5 most significant factors for each category.

```json
{{
  "{model_a} Gains": {{
    "<Category Name (e.g. Conciseness)>": "<Detailed explanation of the advantage observed across multiple papers>"
  }},
  "{model_a} Losses": {{
    "<Category Name (e.g. Hallucination)>": "<Detailed explanation of the failure pattern>"
  }}
}}
```

User Prompt Template:

Here are {num_samples} evaluation traces from the SxS judge. Note: The names have been standardized so that "{model_a}" always refers to the first model and "{model_b}" always refers to the second, regardless of their position in the original prompt.

--- BEGIN TRACES ---

{sxs_evaluations}

--- END TRACES ---

Analyze these traces and generate the summary JSON.

NeurIPS Paper Checklist

1. **Claims**

   Question: Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope?

   Answer: [Yes]

   Justification: The scope and all the contributions are listed in abstract and introduction.

   Guidelines:

   - The answer [N/A] means that the abstract and introduction do not include the claims made in the paper.
   - The abstract and/or introduction should clearly state the claims made, including the contributions made in the paper and important assumptions and limitations. A [No] or [N/A] answer to this question will not be perceived well by the reviewers.
   - The claims made should match theoretical and experimental results, and reflect how much the results can be expected to generalize to other settings.
   - It is fine to include aspirational goals as motivation as long as it is clear that these goals are not attained by the paper.

2. **Limitations**

   Question: Does the paper discuss the limitations of the work performed by the authors?

   Answer: [Yes]

   Justification: The paper addresses limitations and ethical risks in the impact statement and limitations sections.

   Guidelines:

   - The answer [N/A] means that the paper has no limitation while the answer [No] means that the paper has limitations, but those are not discussed in the paper.
   - The authors are encouraged to create a separate "Limitations" section in their paper.
   - The paper should point out any strong assumptions and how robust the results are to violations of these assumptions (e.g., independence assumptions, noiseless settings, model well-specification, asymptotic approximations only holding locally). The authors should reflect on how these assumptions might be violated in practice and what the implications would be.
   - The authors should reflect on the scope of the claims made, e.g., if the approach was only tested on a few datasets or with a few runs. In general, empirical results often depend on implicit assumptions, which should be articulated.
   - The authors should reflect on the factors that influence the performance of the approach. For example, a facial recognition algorithm may perform poorly when image resolution is low or images are taken in low lighting. Or a speech-to-text system might not be used reliably to provide closed captions for online lectures because it fails to handle technical jargon.
   - The authors should discuss the computational efficiency of the proposed algorithms and how they scale with dataset size.
   - If applicable, the authors should discuss possible limitations of their approach to address problems of privacy and fairness.
   - While the authors might fear that complete honesty about limitations might be used by reviewers as grounds for rejection, a worse outcome might be that reviewers discover limitations that aren't acknowledged in the paper. The authors should use their best judgment and recognize that individual actions in favor of transparency play an important role in developing norms that preserve the integrity of the community. Reviewers will be specifically instructed to not penalize honesty concerning limitations.

3. **Theory assumptions and proofs**

   Question: For each theoretical result, does the paper provide the full set of assumptions and a complete (and correct) proof?

   Answer: [N/A]

Justification: The paper proposes an empirical systems architecture and multi-agent framework. It does not contain formal mathematical theorems or proofs.

Guidelines:

- The answer [N/A] means that the paper does not include theoretical results.
- All the theorems, formulas, and proofs in the paper should be numbered and cross-referenced.
- All assumptions should be clearly stated or referenced in the statement of any theorems.
- The proofs can either appear in the main paper or the supplemental material, but if they appear in the supplemental material, the authors are encouraged to provide a short proof sketch to provide intuition.
- Inversely, any informal proof provided in the core of the paper should be complemented by formal proofs provided in appendix or supplemental material.
- Theorems and Lemmas that the proof relies upon should be properly referenced.

4. **Experimental result reproducibility**

Question: Does the paper fully disclose all the information needed to reproduce the main experimental results of the paper to the extent that it affects the main claims and/or conclusions of the paper (regardless of whether the code and data are provided or not)?

Answer: [Yes]

Justification: The paper fully details the algorithmic workflow, evaluation metrics (including the H-Max formulation), and hyperparameters. Furthermore, the exact prompts used for all specialized agents and the LLM judge are provided in the Appendix to ensure the multi-agent orchestration can be reproduced.

Guidelines:

- The answer [N/A] means that the paper does not include experiments.
- If the paper includes experiments, a [No] answer to this question will not be perceived well by the reviewers: Making the paper reproducible is important, regardless of whether the code and data are provided or not.
- If the contribution is a dataset and/or model, the authors should describe the steps taken to make their results reproducible or verifiable.
- Depending on the contribution, reproducibility can be accomplished in various ways. For example, if the contribution is a novel architecture, describing the architecture fully might suffice, or if the contribution is a specific model and empirical evaluation, it may be necessary to either make it possible for others to replicate the model with the same dataset, or provide access to the model. In general. releasing code and data is often one good way to accomplish this, but reproducibility can also be provided via detailed instructions for how to replicate the results, access to a hosted model (e.g., in the case of a large language model), releasing of a model checkpoint, or other means that are appropriate to the research performed.
- While NeurIPS does not require releasing code, the conference does require all submissions to provide some reasonable avenue for reproducibility, which may depend on the nature of the contribution. For example
  (a) If the contribution is primarily a new algorithm, the paper should make it clear how to reproduce that algorithm.
  (b) If the contribution is primarily a new model architecture, the paper should describe the architecture clearly and fully.
  (c) If the contribution is a new model (e.g., a large language model), then there should either be a way to access this model for reproducing the results or a way to reproduce the model (e.g., with an open-source dataset or instructions for how to construct the dataset).
  (d) We recognize that reproducibility may be tricky in some cases, in which case authors are welcome to describe the particular way they provide for reproducibility. In the case of closed-source models, it may be that access to the model is limited in some way (e.g., to registered users), but it should be possible for other researchers to have some path to reproducing or verifying the results.

5. **Open access to data and code**

   Question: Does the paper provide open access to the data and code, with sufficient instructions to faithfully reproduce the main experimental results, as described in supplemental material?

   Answer: [No]

   Justification: The paper does not provide the code but provides all the prompts and details required to reproduce the framework and evaluation.

   Guidelines:

   - The answer [N/A] means that paper does not include experiments requiring code.
   - Please see the NeurIPS code and data submission guidelines (`https://neurips.cc/public/guides/CodeSubmissionPolicy`) for more details.
   - While we encourage the release of code and data, we understand that this might not be possible, so [No] is an acceptable answer. Papers cannot be rejected simply for not including code, unless this is central to the contribution (e.g., for a new open-source benchmark).
   - The instructions should contain the exact command and environment needed to run to reproduce the results. See the NeurIPS code and data submission guidelines (`https://neurips.cc/public/guides/CodeSubmissionPolicy`) for more details.
   - The authors should provide instructions on data access and preparation, including how to access the raw data, preprocessed data, intermediate data, and generated data, etc.
   - The authors should provide scripts to reproduce all experimental results for the new proposed method and baselines. If only a subset of experiments are reproducible, they should state which ones are omitted from the script and why.
   - At submission time, to preserve anonymity, the authors should release anonymized versions (if applicable).
   - Providing as much information as possible in supplemental material (appended to the paper) is recommended, but including URLs to data and code is permitted.

6. **Experimental setting/details**

   Question: Does the paper specify all the training and test details (e.g., data splits, hyperparameters, how they were chosen, type of optimizer) necessary to understand the results?

   Answer: [Yes]

   Justification: The paper explicitly details the test sets used (DeepReview-Bench and AgentReview), the backbone LLMs (Gemini 3.1 Pro, Claude 4.5 Sonnet), and hyperparameter choices such as $k = 3$ search expansion rounds and $N_{QA} = 10$ questions.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - The experimental setting should be presented in the core of the paper to a level of detail that is necessary to appreciate the results and make sense of them.
   - The full details can be provided either with the code, in appendix, or as supplemental material.

7. **Experiment statistical significance**

   Question: Does the paper report error bars suitably and correctly defined or other appropriate information about the statistical significance of the experiments?

   Answer: [Yes]

   Justification: The paper reports the average results across 3 independent runs and explicitly includes standard deviations for both the side-by-side win rates and H-Max scores to demonstrate the stability and low variance of the framework.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - The authors should answer [Yes] if the results are accompanied by error bars, confidence intervals, or statistical significance tests, at least for the experiments that support the main claims of the paper.

- The factors of variability that the error bars are capturing should be clearly stated (for example, train/test split, initialization, random drawing of some parameter, or overall run with given experimental conditions).
- The method for calculating the error bars should be explained (closed form formula, call to a library function, bootstrap, etc.)
- The assumptions made should be given (e.g., Normally distributed errors).
- It should be clear whether the error bar is the standard deviation or the standard error of the mean.
- It is OK to report 1-sigma error bars, but one should state it. The authors should preferably report a 2-sigma error bar than state that they have a 96% CI, if the hypothesis of Normality of errors is not verified.
- For asymmetric distributions, the authors should be careful not to show in tables or figures symmetric error bars that would yield results that are out of range (e.g., negative error rates).
- If error bars are reported in tables or plots, the authors should explain in the text how they were calculated and reference the corresponding figures or tables in the text.

8. **Experiments compute resources**

   Question: For each experiment, does the paper provide sufficient information on the computer resources (type of compute workers, memory, time of execution) needed to reproduce the experiments?

   Answer: [Yes]

   Justification: A dedicated "Computational Complexity and Deployment Analysis" section is included, documenting the empirical token usage (∼443K input/26K output), monetary API costs (∼$1.20 per review), and execution time (∼56s amortized latency).

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - The paper should indicate the type of compute workers CPU or GPU, internal cluster, or cloud provider, including relevant memory and storage.
   - The paper should provide the amount of compute required for each of the individual experimental runs as well as estimate the total compute.
   - The paper should disclose whether the full research project required more compute than the experiments reported in the paper (e.g., preliminary or failed experiments that didn't make it into the paper).

9. **Code of ethics**

   Question: Does the research conducted in the paper conform, in every respect, with the NeurIPS Code of Ethics `https://neurips.cc/public/EthicsGuidelines`?

   Answer: [Yes]

   Justification: The research conforms to the NeurIPS Code of Ethics, and potential ethical risks such as automation bias and data privacy are addressed in the Impact Statement section.

   Guidelines:

   - The answer [N/A] means that the authors have not reviewed the NeurIPS Code of Ethics.
   - If the authors answer [No], they should explain the special circumstances that require a deviation from the Code of Ethics.
   - The authors should make sure to preserve anonymity (e.g., if there is a special consideration due to laws or regulations in their jurisdiction).

10. **Broader impacts**

    Question: Does the paper discuss both potential positive societal impacts and negative societal impacts of the work performed?

    Answer: [Yes]

Justification: The Impact Statement section highlights the positive potential for democratizing feedback and accelerating scientific progress, while explicitly acknowledging the negative risks of automation bias, homogenization of science, and data confidentiality issues.

Guidelines:

- The answer [N/A] means that there is no societal impact of the work performed.
- If the authors answer [N/A] or [No], they should explain why their work has no societal impact or why the paper does not address societal impact.
- Examples of negative societal impacts include potential malicious or unintended uses (e.g., disinformation, generating fake profiles, surveillance), fairness considerations (e.g., deployment of technologies that could make decisions that unfairly impact specific groups), privacy considerations, and security considerations.
- The conference expects that many papers will be foundational research and not tied to particular applications, let alone deployments. However, if there is a direct path to any negative applications, the authors should point it out. For example, it is legitimate to point out that an improvement in the quality of generative models could be used to generate Deepfakes for disinformation. On the other hand, it is not needed to point out that a generic algorithm for optimizing neural networks could enable people to train models that generate Deepfakes faster.
- The authors should consider possible harms that could arise when the technology is being used as intended and functioning correctly, harms that could arise when the technology is being used as intended but gives incorrect results, and harms following from (intentional or unintentional) misuse of the technology.
- If there are negative societal impacts, the authors could also discuss possible mitigation strategies (e.g., gated release of models, providing defenses in addition to attacks, mechanisms for monitoring misuse, mechanisms to monitor how a system learns from feedback over time, improving the efficiency and accessibility of ML).

11. **Safeguards**

    Question: Does the paper describe safeguards that have been put in place for responsible release of data or models that have a high risk for misuse (e.g., pre-trained language models, image generators, or scraped datasets)?

    Answer: [N/A]

    Justification: The paper does not release high-risk pre-trained models, image generators, or scraped datasets. However, the Impact Statement section does emphasize the necessity of privacy safeguards (e.g., ephemeral API processing) for real-world deployment on unpublished manuscripts.

    Guidelines:

    - The answer [N/A] means that the paper poses no such risks.
    - Released models that have a high risk for misuse or dual-use should be released with necessary safeguards to allow for controlled use of the model, for example by requiring that users adhere to usage guidelines or restrictions to access the model or implementing safety filters.
    - Datasets that have been scraped from the Internet could pose safety risks. The authors should describe how they avoided releasing unsafe images.
    - We recognize that providing effective safeguards is challenging, and many papers do not require this, but we encourage authors to take this into account and make a best faith effort.

12. **Licenses for existing assets**

    Question: Are the creators or original owners of assets (e.g., code, data, models), used in the paper, properly credited and are the license and terms of use explicitly mentioned and properly respected?

    Answer: [Yes]

    Justification: The paper credits the creators of the original DeepReviewBench and AgentReview datasets used for evaluation, cites the baselines used and references the commercial LLM APIs (Gemini, Claude) utilized to construct the agents.

Guidelines:

- The answer [N/A] means that the paper does not use existing assets.
- The authors should cite the original paper that produced the code package or dataset.
- The authors should state which version of the asset is used and, if possible, include a URL.
- The name of the license (e.g., CC-BY 4.0) should be included for each asset.
- For scraped data from a particular source (e.g., website), the copyright and terms of service of that source should be provided.
- If assets are released, the license, copyright information, and terms of use in the package should be provided. For popular datasets, `paperswithcode.com/datasets` has curated licenses for some datasets. Their licensing guide can help determine the license of a dataset.
- For existing datasets that are re-packaged, both the original license and the license of the derived asset (if it has changed) should be provided.
- If this information is not available online, the authors are encouraged to reach out to the asset's creators.

13. **New assets**

    Question: Are new assets introduced in the paper well documented and is the documentation provided alongside the assets?

    Answer: [Yes]

    Justification: Such details are thoroughly presented in the Appendix with references in the main paper.

    Guidelines:

    - The answer [N/A] means that the paper does not release new assets.
    - Researchers should communicate the details of the dataset/code/model as part of their submissions via structured templates. This includes details about training, license, limitations, etc.
    - The paper should discuss whether and how consent was obtained from people whose asset is used.
    - At submission time, remember to anonymize your assets (if applicable). You can either create an anonymized URL or include an anonymized zip file.

14. **Crowdsourcing and research with human subjects**

    Question: For crowdsourcing experiments and research with human subjects, does the paper include the full text of instructions given to participants and screenshots, if applicable, as well as details about compensation (if any)?

    Answer: [Yes]

    Justification: The paper includes a human evaluation study involving 20 expert reviewers. Details of the double-blind procedure, the task setup, and the scoring rubrics provided to the experts are documented in the main text and Appendix.

    Guidelines:

    - The answer [N/A] means that the paper does not involve crowdsourcing nor research with human subjects.
    - Including this information in the supplemental material is fine, but if the main contribution of the paper involves human subjects, then as much detail as possible should be included in the main paper.
    - According to the NeurIPS Code of Ethics, workers involved in data collection, curation, or other labor should be paid at least the minimum wage in the country of the data collector.

15. **Institutional review board (IRB) approvals or equivalent for research with human subjects**

Question: Does the paper describe potential risks incurred by study participants, whether such risks were disclosed to the subjects, and whether Institutional Review Board (IRB) approvals (or an equivalent approval/review based on the requirements of your country or institution) were obtained?

Answer: [N/A]

Justification: The paper does not research on human subjects. Human evaluation was performed by a set of 20 senior researchers and the task posed no risk.

Guidelines:

- The answer [N/A] means that the paper does not involve crowdsourcing nor research with human subjects.
- Depending on the country in which research is conducted, IRB approval (or equivalent) may be required for any human subjects research. If you obtained IRB approval, you should clearly state this in the paper.
- We recognize that the procedures for this may vary significantly between institutions and locations, and we expect authors to adhere to the NeurIPS Code of Ethics and the guidelines for their institution.
- For initial submissions, do not include any information that would break anonymity (if applicable), such as the institution conducting the review.

16. **Declaration of LLM usage**

Question: Does the paper describe the usage of LLMs if it is an important, original, or non-standard component of the core methods in this research? Note that if the LLM is used only for writing, editing, or formatting purposes and does *not* impact the core methodology, scientific rigor, or originality of the research, declaration is not required.

Answer: [Yes]

Justification: The entire core methodology centers around a multi-agent framework powered by Large Language Models (LLMs). The specific models used (Gemini 3.1 Pro, Claude 4.5 Sonnet), their roles, and their prompts are exhaustively documented throughout the paper.

Guidelines:

- The answer [N/A] means that the core method development in this research does not involve LLMs as any important, original, or non-standard components.
- Please refer to our LLM policy in the NeurIPS handbook for what should or should not be described.