Paper2Rebuttal: A Multi-Agent Framework for Transparent Author Response Assistance

Qianli Ma* **Chang Guo*** **Zhiheng Tian*** **Siyu Wang** **Jipeng Xiao**
Yuanhao Yue **Zhipeng Zhang†**
AutoLab, School of Artificial Intelligence, Shanghai Jiao Tong University
{mqlqianli,zhipengzhang}@sjtu.edu.cn

Project Page: https://Paper2Rebuttal.github.io
HF Demo: https://huggingface.co/spaces/RebuttalAgent

Abstract

Writing effective rebuttals is a high-stakes task that demands more than linguistic fluency, as it requires precise alignment between reviewer intent and manuscript details. Current solutions typically treat this as a direct-to-text generation problem, suffering from hallucination, overlooked critiques, and a lack of verifiable grounding. To address these limitations, we introduce REBUTTALAGENT, the first multi-agents framework that reframes rebuttal generation as an evidence-centric planning task. Our system decomposes complex feedback into atomic concerns and dynamically constructs hybrid contexts by synthesizing compressed summaries with high-fidelity text while integrating an autonomous and on-demand external search module to resolve concerns requiring outside literature. By generating an inspectable response plan before drafting, REBUTTALAGENT ensures that every argument is explicitly anchored in internal or external evidence. We validate our approach on the proposed REBUTTALBENCH and demonstrate that our pipeline outperforms strong baselines in coverage, faithfulness, and strategic coherence, offering a transparent and controllable assistant for the peer review process.

1 Introduction

The rebuttal phase represents a decisive juncture in the peer review lifecycle where authors must address critiques through evidence-backed clarifications and actionable manuscript revisions. This undertaking extends far beyond simple textual composition. It requires a rigorous synthesis process in which authors must accurately decipher reviewer intent while ensuring every response is firmly anchored in verifiable manuscript details. The inherent difficulty of this multi-step reasoning is amplified by the strict turnaround windows typical of top-tier venues. Authors are frequently forced to reconcile the need for meticulous verification with urgent deadlines, leaving little room for hallucination or ambiguity.

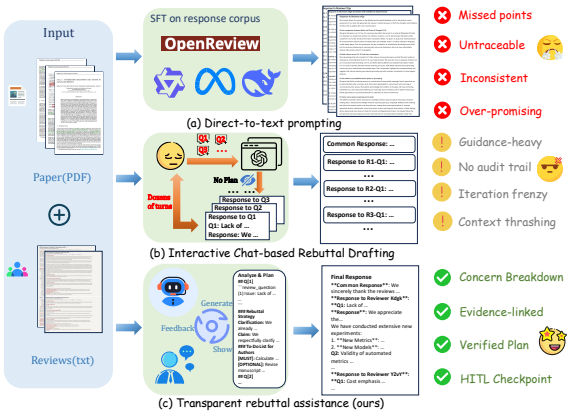


Figure 1: **Overview of our work.** Given a manuscript and reviews, **(a)** direct text generation (SFT on peer-review corpora) often fabricates experiment results and prone to hallucination. **(b)** Interactive prompting with chat-LLMs depends on manual concern feeding and many iterations. **(c)** RebuttalAgent reframes rebuttal writing as a decision-and-evidence organization problem, performing concern breakdown, query-conditioned internal and external evidence construction, and strategy-level plan verification with human-in-the-loop checkpoints before drafting the final response.

In response to these intense cognitive and temporal demands, Large Language Models (LLMs) have emerged as promising assistants for scientific writing (Wang et al., 2024b) and peer-review communication (Gao et al., 2024; Zhu et al., 2025; Lu et al.). Current approaches generally fall into two paradigms. The *direct-to-text* generation paradigm typically involves models that are supervised fine-tuned (SFT) on paper-response pairs (Fig. 1a). While straightforward, this approach is fundamentally flawed because it trains models to memorize specific, non-transferable experimental outcomes rather than the underlying logic of formulating a strategic response. Consequently, these models

*Equal Contribution.
†Corresponding Author.

Proceedings of the 64th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), pages 46134–46166
July 2-7, 2026

are prone to hallucination, frequently fabricating experimental results or over-commit to unverified claims instead of reasoning about the actual content of the manuscript. The second paradigm relies on interactive sessions with proprietary chat-LLMs such as GPT or Gemini (Fig. 1b). While these high-capability models can offer superior reasoning, the workflow is notoriously inefficient and opaque. Authors are forced to engage in lengthy, multi-turn prompting to guide the model, which consumes valuable time that could be spent on verification. Furthermore, critical intermediate steps like concern parsing and evidence retrieval remain concealed behind the chat interface. This lack of transparency makes the process difficult to audit and renders the output quality heavily dependent on the prompting expertise of the user.

In this paper, we reframe rebuttal assistance as a *decision and evidence organization problem* with explicit constraints, rather than the free-form text generation tasks. Specifically, a reliable system must satisfy four critical requirements: **(i) Comprehensive Coverage,** tracking every reviewer's concern without omission; **(ii) Strict Faithfulness,** adhering to the submitted manuscript without hallucinating technical details; **(iii) Verifiable Grounding,** linking major statements to specific internal passages or external references; and **(iv) Global Consistency,** maintaining a unified stance and avoiding conflicting commitments across different responses. To instantiate this view, we propose REBUTTALAGENT, a multi-agent system that enforces a novel "verify-then-write" workflow to overcome the opacity of previous two paradigms, shown in Fig. 1c.

Instead of rushing to generation, our architecture explicitly decouples reasoning from drafting by producing verifiable intermediate artifacts. The process begins by atomizing unstructured reviews into discrete concerns to guarantee comprehensive coverage, followed by a dual-source evidence construction phase that synthesizes high-fidelity manuscript passages and citation-ready external briefs to strictly ground every claim. Crucially, we introduce a strategic planning stage that audits the response logic for global consistency and commitment safety before any text is drafted, ensuring that concessions made to one reviewer do not contradict the overall stance. By exposing these structured artifacts through human-in-the-loop checkpoints, REBUTTALAGENT transforms rebuttal writing from a black-box generation task into a transparent, author-controlled collaboration.

We evaluate REBUTTALAGENT through an author-centric lens, prioritizing practical usability and reliability over mere text fluency. Specifically, we assess performance across four rigorous dimensions: **coverage** of reviewer concerns, **traceability** of evidence sources, **global coherence** of the argumentative stance, and overall **argumentation quality**. Experimental results on our proposed benchmark demonstrate that our pipeline consistently outperforms previous "direct-to-text" baselines and chat-LLMs on these critical metrics. By delivering structured, verifiable assistance, REBUTTALAGENT significantly reduces the cognitive burden of rebuttal writing while ensuring authors remain the ultimate arbiters of their scientific defense.

Our contributions are: ♠ We formulate rebuttal assistance as a decision-and-evidence organization problem and propose `RebuttalAent`, a multi-agent system with explicit verification and human-in-the-loop checkpoints. ♠ We introduce concern-conditioned context construction and on-demand evidence synthesis to produce point-specific, verifiable support under realistic context limits. ♠ We construct a benchmark and establish an author-centric evaluation protocol, demonstrating that our approach outperforms baselines in coverage, traceability, and coherence.

2 Related Works

LLM Agents. LLMs (OpenAI, 2025; Team et al., 2023) were initially valued for fluent generation, but real deployments revealed a mismatch between writing well and completing complex tasks reliably. When goals require multi-step planning, fresh evidence, and interaction with external systems, purely parametric generation can accumulate errors and hallucinations, motivating a shift toward intelligent "agents" embedded in dynamic, goal-directed frameworks that plan and act with external tools and environments. Recent work (Yao et al., 2023b,a) shows that combining reasoning traces with concrete actions (*e.g.*, search tool) improves robustness in long-horizon tasks and reduces hallucinations. Modern agents often incorporate deliberation and search (Wei et al., 2023; Wang et al., 2024b), learned tool-use policies (Schick et al., 2023; Patil et al., 2023), and memory or reflection from execution feedback (Shinn et al., 2023; Zhang et al., 2024). Multi-agent frameworks further enable role specialization and structured collabora-

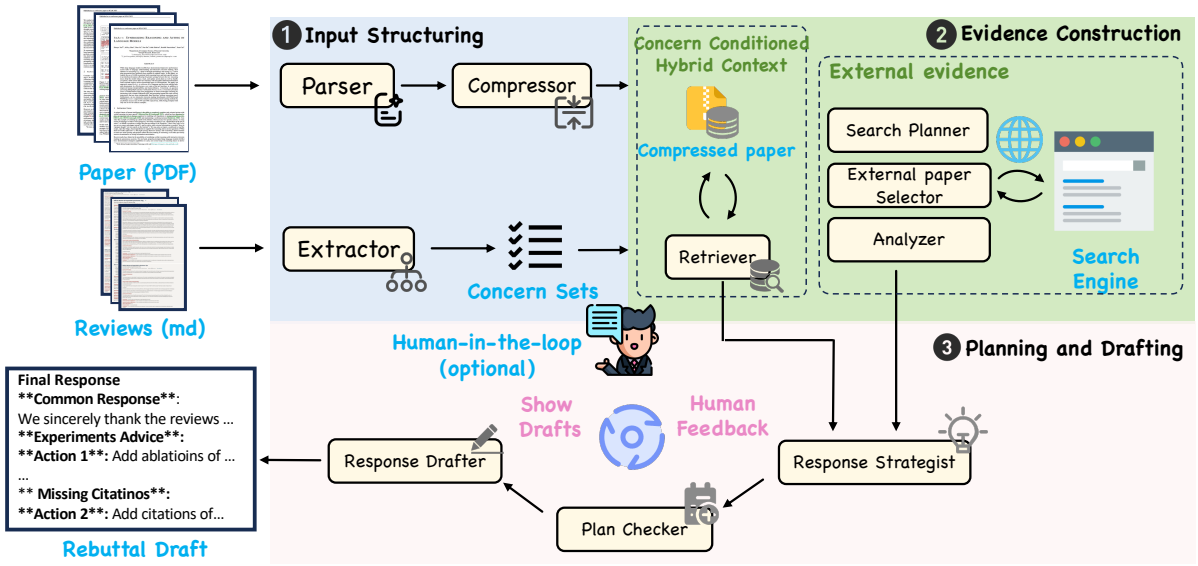


Figure 2: **Overview of RebuttalAgent.** Given a manuscript (PDF) and reviewer comments, the system **(1)** structures inputs by parsing and compressing the paper with fidelity checks and extracting atomic reviewer concerns with coverage checks; **(2)** builds concern-conditioned evidence by constructing a query-specific hybrid manuscript context and, when needed, retrieving and summarizing external literature into citation-ready briefs; and **(3)** generates an inspectable, evidence-linked response plan that is checked for consistency and commitment safety, incorporates optional author feedback, and is then realized into a formal rebuttal draft.

tion (Wang et al., 2024a; Wu et al., 2023; Ma et al., 2025c; Lu et al.; D'Arcy et al., 2024), while benchmarks such as AgentBench (Liu et al., 2025b), WebArena (Zhou et al., 2024), and GAIA (Mialon et al., 2023) evaluate real-world tool use and end-to-end task success. These advances motivate extending agentic systems from *conducting* research to *communicating* it, *e.g.*, retrieving evidence, organizing words and iteratively refining rebuttals.

AI Assisted Peer Review. Peer review stands as the cornerstone of research quality yet faces significant strain from the exponential growth in conference submissions. This pressure has catalyzed the adoption of LLMs to maintain efficiency and decision reliability across the review pipeline (Gao et al., 2024; Lu et al.; Zhu et al., 2025; Zhang et al., 2025). Within this process, the author rebuttal phase holds unique value for rectifying misunderstandings and influencing borderline decisions (Gao et al., 2019). To operationalize this complex interaction, researchers have developed datasets like DISAPERE (Kennard et al., 2021) and APE (Cheng et al., 2020) for argument alignment alongside comprehensive corpora like Re^2 (Zhang et al., 2025). While recent efforts employ argumentative strategies (Purkayastha et al., 2023) or multi-agent simulations (Yu et al., 2025; Jin et al., 2024) to model this workflow, they predominantly treat rebuttal generation as a single-step prompt-to-text task. As illustrated in Fig. 1a, these methods overlook the critical need for explicitly decomposing concerns and planning evidence-based responses.

3 RebuttalAgent

REBUTTALAGENT operates as a multi-agent framework that transforms the rebuttal process into a structured and inspectable workflow. By generating evidence-linked intermediate artifacts before drafting the final text, the system ensures that the output remains grounded and controllable. Fig. 2 illustrates how the architecture decomposes complex reasoning into specialized agents paired with lightweight checkers. This design exposes critical decision points and allows authors to retain full responsibility for the strategic stance and final wording. The pipeline initiates by distilling the manuscript into a structured summary and extracting atomic reviewer concerns to enable stable long-context reasoning (Sec. 3.1). Guided by these atomic concerns, the system constructs evidence bundles by retrieving relevant high-fidelity excerpts from the original manuscript and augmenting them with verifiable external literature via web search (Sec. 3.2). The workflow concludes by synthesizing an explicit response plan that outlines the arguments and evidence links. Authors refine this plan through a human-in-the-loop mechanism before the system produces a formal rebuttal letter

compliant with academic conventions (Sec. 3.3).

3.1 Manuscript and Review Structuring

The pipeline commences by distilling the raw manuscript and reviews into condensed representations optimized for downstream reasoning. This approach addresses the dual challenges of efficiency and controllability as effective rebuttals demand repeated access to fine-grained evidence scattered throughout the paper. Since processing the full manuscript directly is often costly and brittle due to context limitations, our compact format minimizes token overhead while improving retrieval precision. It serves as a navigational anchor that allows subsequent modules to selectively access high-fidelity excerpts from the original text only when precise evidence is necessary.

Dense Manuscript to Compact Representation. The transformation begins as a parser agent converts the manuscript PDF into a paragraph-indexed format to preserve structural integrity and facilitate targeted lookups. A compressor agent subsequently distills these paragraphs into a concise representation that retains essential technical statements and experimental results. This compact view functions as the primary retrieval surface and enables the system to match reviewer concerns to relevant sections with minimal token usage. To safeguard against silent information loss, a consistency checker verifies each condensed unit against its source and automatically triggers reprocessing if it detects missing claims or semantic drift.

Complex Reviews to Actionable Atomic Concerns. Operating in parallel with manuscript processing, an extractor agent parses raw feedback into discrete and addressable atomic concerns. This component organizes the critiques by grouping related sub-questions and assigning preliminary categories. A coverage checker subsequently validates the output for intent preservation and appropriate granularity to guarantee that substantive points remain distinct without being over-split or incorrectly merged. The resulting structured list forms the foundational unit for the subsequent evidence gathering and response planning stages.

3.2 Evidence Construction

With the atomic concerns established, the system generates targeted evidence bundles to ensure that every argument remains traceable to specific facts. This strategy contrasts sharply with the direct generation approaches depicted in Fig. 1a that bypass explicit grounding. By prioritizing evidence construction over immediate text generation, our pipeline anchors each concern in verifiable sources and ensures that the downstream planning and drafting stages operate on validated information.

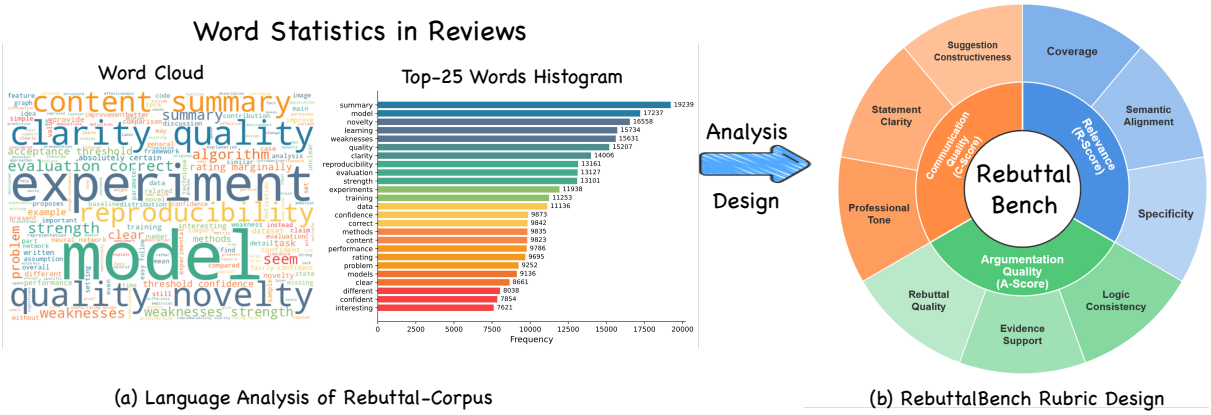
Atomic Concern Conditioned Hybrid Context. The system identifies the most pertinent sections by searching within the compressed manuscript representation (Sec.3.1) for each atomic concern. It then selectively expands these focal points by retrieving the corresponding raw text to replace the specific condensed units while retaining the rest of the document in its summarized form. This approach yields an atomic concern conditioned hybrid context that integrates the efficiency of the compressed view with the precision of the original text. Such a structure enables the system to support its reasoning with exact quotations and detailed evidence without overwhelming the context window.

On-Demand External Evidence. While the hybrid context effectively grounds responses in the authors' own work, certain reviewer critiques necessitate evidence beyond the manuscript boundaries. To address scenarios such as novelty disputes or requests for broader positioning where internal data is insufficient, the system augments the evidence bundle with external support. A search planner initiates this expansion by formulating a targeted search strategy, while a subsequent retrieval step gathers candidate papers via scholarly search tools[1]. A screening agent then filters these candidates for relevance and utility to ensure high-quality input. The pipeline concludes this phase by parsing the selected works into a structured evidence brief that highlights key claims and experimental comparisons to provide citation-ready material for the subsequent planning and drafting stages.

3.3 Planning and Drafting

A critical failure of the direct-to-text pipeline is its tendency to hallucinate experimental results when addressing empirical critiques. Our system overcomes this by implementing a bifurcated reasoning strategy that strictly distinguishes between interpretative defense and necessary intervention. For concerns resolvable through existing data, the Strategist Agent synthesizes arguments directly from the hybrid context and anchors them in the manuscript text. In contrast, when the system detects a demand for new experiments or baselines, it explic-

[1] https://export.arxiv.org/api/query



(a) Language Analysis of Rebuttal-Corpus (b) RebuttalBench Rubric Design

Figure 3: **RebuttalBench statistics and rubric design.** **(a)** Word-cloud and top-word histogram of reviews in REBUTTALBENCH-CORPUS, highlighting recurring reviewer emphases (*e.g.*, clarity, novelty, reproducibility). **(b)** Motivated by these signals, REBUTTALBENCH evaluates rebuttals with a rubric that mirrors these concerns, scoring *Relevance*, *Argumentation Quality*, and *Communication Quality* rather than fluency alone.

itly inhibits result generation and instead produces concrete *Action Items* framed as recommendations (see cases in App. I). This design prevents the fabrication of outcomes by forcing a structural pause where authors must verify or perform the suggested tasks. The resulting plan serves as an interactive human-in-the-loop checkpoint that allows authors to actively refine the strategic logic rather than merely accepting or rejecting proposals. Users can modify the scope of action items or correct the reasoning path to ensure the strategy aligns perfectly with their capabilities and intent. Only after the author validates these strategic decisions does the Drafter Agent convert the plan into a final response to ensure that every claim remains grounded in reality. Optionally, the drafter can also produce a submission-style rebuttal draft from the validated plan, but it renders any yet-to-be-conducted experiments as explicit placeholders (*e.g.*, `[TBD]`). Authors can then fill in these placeholders after completing the recommended action items, keeping the draft faithful.

4 RebuttalBench

Standard evaluation metrics for text generation fail to capture the strategic nuance and factual precision required in peer review rebuttals. Therefore, we introduce REBUTTALBENCH as a specialized benchmark derived from real-world OpenReview interactions. This dataset moves beyond simple text-to-text pairs by curating high-quality review-response dyads to ensure technical density and argumentative complexity. We complement the data with a multidimensional evaluation framework that prioritizes content coverage and evidence traceability over surface-level fluency. Unlike generic instruction-following benchmarks, our protocol specifically measures how well a system identifies atomic concerns and grounds its counter-arguments in verifiable facts. This allows us to quantify the gap between the hallucination-prone outputs of standard models and the structured reasoning produced by our pipeline.

4.1 Evaluation Dataset

Data source. To evaluate rebuttal assistance with an observable post-rebuttal signal, we curated a dataset of peer-review discussion threads from the publicly available ICLR OpenReview forum. Each instance in our benchmark pairs an initial reviewer critique with the corresponding author rebuttal and crucially includes the reviewer's follow-up response. We leverage the subsequent reviewer reaction as a decisive classification signal to partition the dataset into positive and negative samples for evaluation purposes. Positive instances are identified by follow-up comments confirming that all concerns were resolved while negative samples consist of cases where the reviewer indicated that the rebuttal failed to address the core issues.

Filtering and corpus construction. Starting from the raw peer-review discussion threads, we apply automatic filtering to retain instances with sufficiently explicit follow-up signals and remove ambiguous cases to obtain a broad and reliable evaluation pool. This yields `RebuttalBench-Corpus`, a broad pool of 9.3K review-rebuttal pairs used for analysis and evaluation setup. (see Appendix A). To form a focused and challenging benchmark for standardized comparison, we construct REBUTTALBENCH-CHALLENGE by ranking papers according to the number of instances that

exhibit both positive and negative follow-up signals, and selecting the top 20 papers with over 100 reviewers. This strategy maximizes within-paper diversity of resolved and unresolved concerns, producing a compact test suite with realistic interaction patterns. We provide details about the filtering strategy in Appendix D.

Data statistics. Fig. 3 summarizes corpus-level characteristics of REBUTTALBENCH-CORPUS. Beyond basic length and interaction statistics, we visualize reviewer language with a word cloud and top-words histogram, shown in Fig. 3a. Frequent terms such as *clarity*, *quality*, *correct(ness)*, *reproducibility*, *novelty*, and *experiments* indicate that reviewers repeatedly emphasize exposition, claim support, and scientific rigor; these axes are also explicitly reflected in standard review forms used in OpenReview venues. Accordingly, our rubric-based evaluation is designed to align with these recurring concerns by scoring relevance/coverage to reviewer points, strength of evidence-backed argumentation, and communication quality (*e.g.*, clarity and professionalism), demonstrated in Fig. 3b.

4.2 Evaluation Metrics

To systematically measure rebuttal response quality beyond surface fluency, we use an LLM-as-judge (Zheng et al., 2023; Lin and Chen, 2023) rubric with a fine-grained **0-5** scale. The evaluation framework covers three complementary dimensions: *Relevance* (R-Score), *Argumentation Quality* (A-Score), and *Communication Quality* (C-Score). Each dimension contains three components (9 total). We calculate the average component scores within each dimension and then compute the final score. Full component rubrics and judge prompts are provided in Appendix B.

R-Score evaluates the extent to which the response addresses reviewer concerns with point-specific precision. It rewards outputs that cover all major points without omission and demonstrate a correct interpretation of the critique while favoring concrete actions over generic assurances

A-Score measures the strength of the justification behind each claim. It requires arguments to be logically consistent and supported by appropriate evidence from the manuscript or external sources. The metric prioritizes substantive rebuttals that engage with the underlying critique rather than offering superficial restatements.

C-Score captures the quality of communication and professional conduct. It assesses whether the response maintains a respectful tone and presents information with a clear structure and unambiguous language. The metric ensures the text remains constructive to facilitate a productive discussion between the reviewer and the author.

In addition to scalar scores, the evaluator outputs a brief structured diagnosis (strengths, weaknesses, and suggested improvements) for qualitative analysis. Detailed scoring standards (0-5 anchors) and implementation are provided in Appendix B.

5 Experiments

5.1 Experimental Setup

We assess the efficacy of REBUTTALAGENT by comparing it with strong closed-source LLM baselines and by ablating key components of the system. For scalable and controlled benchmarking, all experiments in the main paper run REBUTTALAGENT in a fully automated mode without human intervention. While human-in-the-loop checkpoints can further improve reliability and author control, they are impractical for batch evaluation at scale. Accordingly, the reported results should be viewed as a *conservative lower bound* on the system's performance under real-world usage.

Baselines. We consider four SOTA LLMs as baselines: GPT-5-mini (OpenAI, 2025), Grok-4.1-fast (xGr), Gemini-3-Flash (Team et al., 2023), and DeepSeekV3.2 (Liu et al., 2025a). We also compare with general multi-agent systems (See App. F). For each baseline model, we evaluate a *direct-to-text generation* setting where the model produces a rebuttal conditioned on the manuscript and reviewer comments. To ensure a fair comparison, we also instantiate REBUTTALAGENT with the same model as its foundation backbone, keeping inputs and outputs identical across conditions; differences therefore reflect the contribution of our structured pipeline rather than the underlying model choice.

Implementation Details. To ensure controlled and fair comparisons, we evaluate REBUTTALAGENT and each closed-source baseline under matched model backbones. For every baseline LLM (*e.g.*, GPT-5-mini (OpenAI, 2025)), we instantiate REBUTTALAGENT with the same LLM as its backbone, so that both the baseline and REBUTTALAGENT consume the same manuscript and reviewer comments and produce responses in an identical point-by-point format. Differences therefore reflect the contribution of the structured workflow rather than model capacity. All experiments in the

Table 1: **Main evaluation results across our full suite of RebuttalBench.** Results demonstrate promising improvements of our method against the baseline LLM.

Method	Relevance			Argumentation Quality			Communication Quality			Average
	Coverage	Semantic Alignment	Specificity	Logic Consistency	Evidence Support	Response Engagement	Professional Tone	Statement Clarity	Suggestion Constructiveness	
DeepSeekV3.2	3.65	4.44	3.28	3.44	3.01	3.16	3.37	3.96	3.81	3.57
RebuttalAgent-DeepSeekV3.2	**4.43** (+0.78)	**4.82** (+0.38)	**4.39** (+1.11)	**3.86** (+0.42)	**3.23** (+0.22)	**3.79** (+0.63)	**3.60** (+0.23)	**4.18** (+0.22)	**4.06** (+0.25)	**4.08** (+0.51)
Grok4.1-fast	3.98	4.58	3.72	3.73	3.32	3.60	3.48	4.05	3.92	3.82
RebuttalAgent-Grok-4.1-fast	**4.66** (+0.68)	**4.92** (+0.34)	**4.65** (+0.93)	**4.13** (+0.40)	**3.42** (+0.10)	**4.15** (+0.55)	**3.68** (+0.20)	**4.23** (+0.18)	**4.24** (+0.32)	**4.25** (+0.43)
Gemini3-Flash	4.00	4.71	3.77	3.71	3.30	3.56	3.51	4.08	3.95	3.85
RebuttalAgent-Gemini3-Flash	**4.51** (+0.51)	**4.88** (+0.17)	**4.49** (+0.72)	**4.11** (+0.40)	**3.39** (+0.09)	**4.07** (+0.51)	**3.78** (+0.27)	**4.28** (+0.20)	**4.09** (+0.14)	**4.23** (+0.38)
GPT5-mini	3.61	4.22	2.96	3.37	2.92	3.07	3.35	3.95	3.91	3.48
RebuttalAgent-GPT5-mini	**4.34** (+0.73)	**4.84** (+0.62)	**4.29** (+1.33)	**3.78** (+0.41)	**3.31** (+0.39)	**3.70** (+0.63)	**3.60** (+0.25)	**4.21** (+0.26)	**4.24** (+0.33)	**4.05** (+0.57)

main paper run RebuttalAgent in a fully automated mode, and we keep decoding settings consistent across conditions for each backbone. Finally, we adopt Gemini-3-Flash (Team et al., 2023) as a unified LLM judge for all systems and ablations. Full prompt templates and evaluation prompts are provided in Appendix B and Appendix H.

5.2 Main Results

Obs. 1: **RebuttalAgent consistently outperforms strong closed-source LLMs.** As shown in Tab. 1, under fair comparisons where REBUTTALAGENT and LLM baselines share the same base models, REBUTTALAGENT achieves consistent improvements across all evaluation dimensions on REBUTTALBENCH. The largest gains are observed in *Relevance* and *Argumentation Quality*. Across matched base models, REBUTTALAGENT improves *coverage* by up to **+0.78** for DeepSeekV3.2 and *specificity* by up to **+1.33** for GPT5-mini, and strengthens argumentation with up to **+0.63** higher *rebuttal quality*. Improvements in *Communication Quality* are smaller but consistent, suggesting that the gains mainly come from structured decision making and evidence organization rather than surface-level fluency. Notably, these gains are achieved without changing the language model, indicating that performance improvements stem from task decomposition and structured intermediate reasoning rather than stronger generative capacity. This suggests that rebuttal quality is bottlenecked less by surface fluency and more by systematic concern tracking, evidence grounding, and response planning. These factors that are poorly handled by direct-to-text prompting even with SOTA LLMs.

Obs. 2: **The benefit of RebuttalAgent is larger for weaker base models.** Tab. 1 also suggests that the weaker the base model, the larger the improvement obtained from our agent pipeline. While all advanced LLMs benefit from our RebuttalAgent, the margin over direct-to-text prompting is more pronounced for smaller or less capable backbones (*e.g.*, GPT5-mini) than for stronger ones (*e.g.*, Gemini-3-Flash). Using the mean score averaged over all nine components as a summary, the weakest backbone GPT5-mini gains about **+0.55** on average, whereas stronger proprietary backbones (*e.g.*, Gemini-3-Flash) gain a smaller margin (**+0.33**). The same pattern is particularly clear on *Relevance*. GPT5-mini improves by roughly **+0.89** on the relevance sub-scores (coverage, semantic alignment, and specificity), compared to about **+0.47** for Gemini-3-Flash. This indicates that explicit concern structuring, evidence construction, and response planning can partially compensate for limited base-model capability, shifting performance bottlenecks from raw generation to decision and evidence organization.

Obs. 3: **RebuttalAgent yields balanced improvements across the full rebuttal pipeline.** Beyond isolated metric gains, Tab. 1 shows that REBUTTALAGENT improves *all three* dimensions in a coordinated way across matched base models. For example, under Gemini-3-Flash, REBUTTALAGENT raises *Relevance* (coverage from 4.00 to 4.51; specificity from 3.77 to 4.49), strengthens *Argumentation Quality* (logic consistency from 3.71 to 4.11; rebuttal quality from 3.56 to 4.07), and also improves *Communication Quality* (professional tone from 3.51 to 3.78; statement clarity from 4.08 to 4.28). A similar across-the-board improvement pattern holds for other backbones, suggesting that the benefits are not localized to a single stage, such as evidence insertion or phrasing. Instead, structuring concerns and grounding claims early supports downstream planning and drafting, leading to more coherent and constructive final responses.

5.3 Ablation Study

Ablation setting. To understand the contribution of each intermediate artifact, we perform controlled ablations by removing one module at a time from the full RebuttalAgent pipeline while keeping the base model, prompts, and evaluation protocol fixed.

Table 2: **Ablation study on key components.** We remove each module from the full system: Input Structuring, Evidence Construction, and Checker.

Metric	RebuttalAgent	w/o Component		
		Structing	Evidence	Checker
Relevance				
Coverage	4.51	4.49 (-0.02)	4.26 (-0.25)	4.54 (+0.03)
Semantic Alignment	4.88	4.71 (-0.17)	4.73 (-0.15)	4.89 (+0.01)
Specificity	4.49	4.46 (-0.03)	4.19 (-0.30)	4.47 (-0.02)
Argumentation Quality				
Logic Consistency	4.11	4.06 (-0.05)	4.05 (-0.06)	4.13 (+0.02)
Evidence Support	3.39	3.23 (-0.16)	3.32 (-0.07)	3.39 (+0.00)
Response Engagement	4.07	4.04 (-0.03)	3.97 (-0.10)	4.01 (-0.06)
Communication Quality				
Professional Tone	3.78	3.69 (-0.09)	3.74 (-0.04)	3.73 (-0.05)
Statement Clarity	4.28	4.33 (+0.05)	4.22 (-0.06)	4.29 (+0.01)
Suggestion Constructiveness	4.09	4.06 (-0.03)	3.82 (-0.27)	4.05 (-0.04)

Specifically, we consider three variants: (i) **w/o Input Structuring**, where reviewer concerns are not explicitly decomposed and merged but handled in raw form; (ii) **w/o Evidence Construction**, where external literature retrieval and citation-ready evidence briefs are disabled; and (iii) **w/o Checkers**, where plan-level verification for coverage, evidence linkage, and cross-point consistency is removed. All variants still produce complete rebuttal drafts, allowing us to isolate how each module affects response quality rather than system completeness.

Obs. 4: **External evidence briefs are the most Critical Artifact, while structuring and checkers provide more targeted benefits.** Tab. 2 shows that *Evidence Construction* is the most critical intermediate artifact. Removing external evidence briefs leads to the largest and most consistent degradation across dimensions, with clear drops in *Relevance* and *Communication Quality*. In particular, *Coverage* decreases from 4.51 to 4.26 and *constructiveness* falls from 4.09 to 3.82, indicating that citation-ready evidence plays a central role in enabling specific, actionable, and convincing responses rather than generic assurances. These degradations indicate that citation-ready evidence briefs are central to producing point-specific and constructive responses. Although the effects are smaller, *Input Structuring* and *Checkers* also contribute measurably to overall quality. Without structuring, multiple metrics decline, including *semantic alignment* (4.88 to 4.71) and *evidence support* (3.39 to 3.23), suggesting that explicit concern decomposition and stable manuscript representations help preserve intent understanding and evidence linkage. Without checkers, we observe degradations in key quality dimensions such as *evidence support* (3.39 to 3.33) and *rebuttal quality* (4.07 to 4.01), indicating that lightweight verification remains beneficial even when base responses are fluent. Overall, the ablation results indicate that the gains of REBUTTALAGENT arise from the *combination* of complementary modules. Evidence-centered artifacts act as the primary driver of quality improvements, while explicit structuring and verification provide guardrails that reduce error accumulation.

Deep Dive into the Role of Checkers and the Coverage Paradox. While the overall benefits of RebuttalAgent are clear, Tab. 2 presents a counter-intuitive observation: removing the checkers leads to a marginal increase in Coverage (from 4.51 to 4.54). However, deeper qualitative analysis reveals that this actually reflects a known failure mode of unconstrained LLM generation: **verbose over-generation**. Without the checker, the model tends to generate broader, less focused responses filled with generic qualitative explanations to safely appease reviewers. This verbosity artificially inflates the automated Coverage score but actively harms precision and actionability (as reflected by the drop in Suggestion Constructiveness from 4.09 to 4.05). To illustrate, consider a test case where a reviewer asked to clarify the non-Markovian setting and the definitions of Utility versus Reward (Details in App. G). This example demonstrates how the checker actively trades superficial verbosity for rigorous constructiveness, ensuring the rebuttal provides concrete manuscript edits rather than generic glossaries.

5.4 Case Study

We also provide cases that directly compare REBUTTALAGENT with strong LLM baselines on representative reviewer concerns in Appendix I. Rather than emphasizing the final rebuttal prose, these examples highlight the intermediate artifacts that RebuttalAgent surfaces to authors: an explicit response strategy, evidence-linked clarification points, and concrete action items (*e.g.*, targeted edits, and suggested experiments or additional) that can be verified before any claims are finalized.

Obs. 5: **Action items reduce hallucination and over-commitment.** In the shown cases, reviewers either question a potential contradiction in a key proposition or criticize the clarity and rigor of the theoretical presentation. RebuttalAgent first produces an inspectable plan that separates *interpretative defense* (what can be clarified using manuscript content) from *necessary intervention* (what requires additional evidence). Crucially, when new experiments or analyses are implicated, RebuttalAgent does not generate results; instead, it outputs concrete deliverables (*e.g.*, revised ex-

position, a new proof sketch) and a scoped to-do list, as described in Sec. 3.3. By contrast, baseline outputs tend to respond with a short narrative that may be overly confident or implicitly commit to empirical claims without exposing the underlying reasoning and verification steps. Overall, these cases illustrate how REBUTTALAGENT supports author decision-making by making the reasoning path and required work explicit before drafting, enabling authors to validate or edit the plan and keep final commitments grounded.

6 Conclusion

We proposed REBUTTALAGENT, a multi-agent framework for rebuttal assistance that constructs structured, evidence-linked intermediate artifacts before drafting text. By decomposing rebuttal writing into concern structuring, query-conditioned context building, on-demand external evidence synthesis, and response planning, the system improves traceability and cross-point coherence while keeping authors responsible for final decisions and wording. We also introduced an author-centric benchmark and a rubric-based evaluation that measures relevance, global coherence, and argumentation quality beyond text fluency. Experimental results on our benchmark show that REBUTTALAGENT improves the key requirements of reliable rebuttal assistance, highlighting the benefits of a transparent "verify-then-write" workflow that reduces cognitive burden while keeping authors in control of the final wording.

Limitations and Future Work

Our framework prioritizes reliability through structured intermediate artifacts, but several limitations remain. First, the current system emphasizes transparency and inspectability over minimal latency. Optimizing cost and runtime (*e.g.*, caching artifacts and adaptive early-exit policies) is an important engineering direction for broader adoption. Second, intermediate artifacts are not yet fully reused across concerns or author iterations. More aggressive caching and incremental updates could reduce redundant computation when authors revise drafts or when similar concerns recur across reviewers. Third, rebuttal conventions vary across venues and subfields. Incorporating lightweight style and policy constraints (*e.g.*, venue-specific formatting and tone) could improve alignment without changing the underlying reasoning pipeline.

Broader Impact and Ethics Statement

Our goal is to assist authors in rebuttal writing by organizing reviewer concerns, grounding responses in verifiable evidence, and producing inspectable intermediate artifacts before drafting text. Used responsibly, this can reduce cognitive burden and improve clarity, completeness, and consistency in peer-review communication.

Ethical risks and mitigations. Automated rebuttal assistance raises ethical concerns, including the misuse of AutoRebuttal to produce persuasive but misleading responses (*e.g.*, exaggerated claims, hallucinated results, or unrealistic commitments), privacy risks when handling unpublished manuscripts and reviews. We explicitly acknowledge these risks and mitigate them through design choices: AutoRebuttal is an author-assistance tool rather than an autonomous rebuttal system, and authors remain responsible for final stance, commitments, and wording. The system produces inspectable intermediate artifacts (*e.g.*, concern lists, evidence links, response plans) and performs explicit checks for coverage, faithfulness to the manuscript, evidence traceability, and global coherence before drafting. External retrieval is triggered only when needed and preserves provenance to facilitate verification and reduce unreliable citations. For confidential material, we recommend local or institution-approved deployment. In any release, we follow the source terms and avoid distributing sensitive or personally identifying information. Our ultimate goal is to assist authors in deeply engaging with reviewer feedback and improve their manuscripts through structured planning rather than automating deceptive rebuttal, thereby enhancing the quality and efficiency of the peer review process for the benefit of the broader research community.

Dataset policy. Our benchmark is derived from publicly available ICLR 2023 forums. In any release, we will follow the source terms and avoid distributing sensitive or personally identifying information.

We explicitly acknowledge these ethical issues and incorporate the above safeguards to promote responsible use.

Acknowledgments

The work was supported by the Natural Science Foundation of China (Grant No. 62503323).

References

Gemini Deep Research - your personal research assistant — gemini.google. https://gemini.google/overview/deep-research. [Accessed 22-04-2025].

Grok 4 | xAI — x.ai. https://x.ai/news/grok-4. [Accessed 15-10-2025].

Jinheon Baek, Sujay Kumar Jauhar, Silviu Cucerzan, and Sung Ju Hwang. 2024. Researchagent: Iterative research idea generation over scientific literature with large language models. *arXiv preprint arXiv:2404.07738*.

Liying Cheng, Lidong Bing, Qian Yu, Wei Lu, and Luo Si. 2020. Ape: Argument pair extraction from peer review and rebuttal via multi-task learning. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 7000–7011.

Mike D'Arcy, Tom Hope, Larry Birnbaum, and Doug Downey. 2024. Marg: Multi-agent review generation for scientific papers. *arXiv preprint arXiv:2401.04259*.

Bingjie Gao, Qianli Ma, Xiaoxue Wu, Shuai Yang, Guanzhou Lan, Haonan Zhao, Jiaxuan Chen, Qingyang Liu, Yu Qiao, Xinyuan Chen, and 1 others. 2025. Rapo++: Cross-stage prompt optimization for text-to-video generation via data alignment and test-time scaling. *arXiv preprint arXiv:2510.20206*.

Yang Gao, Steffen Eger, Ilia Kuznetsov, Iryna Gurevych, and Yusuke Miyao. 2019. Does my rebuttal matter? insights from a major nlp conference. *arXiv preprint arXiv:1903.11367*.

Zhaolin Gao, Kianté Brantley, and Thorsten Joachims. 2024. Reviewer2: Optimizing review generation through prompt generation. *arXiv preprint arXiv:2402.10886*.

Yichen He, Guanhua Huang, Peiyuan Feng, Yuan Lin, Yuchen Zhang, Hang Li, and 1 others. 2025. Pasa: An llm agent for comprehensive academic paper search. *arXiv preprint arXiv:2501.10120*.

Xiang Hu, Hongyu Fu, Jinge Wang, Yifeng Wang, Zhikun Li, Renjun Xu, Yu Lu, Yaochu Jin, Lili Pan, and Zhenzhong Lan. 2024. Nova: An iterative planning and search approach to enhance novelty and diversity of llm generated ideas. *arXiv preprint arXiv:2410.14255*.

Ziqi Huang, Fan Zhang, Xiaojie Xu, Yinan He, Jiashuo Yu, Ziyue Dong, Qianli Ma, Nattapol Chanpaisit, Chenyang Si, Yuming Jiang, Yaohui Wang, Xinyuan Chen, Ying-Cong Chen, Limin Wang, Dahua Lin, Yu Qiao, and Ziwei Liu. 2025. VBench++: Comprehensive and versatile benchmark suite for video generative models. *IEEE Transactions on Pattern Analysis and Machine Intelligence*.

Yiqiao Jin, Qinlin Zhao, Yiyang Wang, Hao Chen, Kaijie Zhu, Yijia Xiao, and Jindong Wang. 2024. Agentreview: Exploring peer review dynamics with llm agents. In *EMNLP*.

M. G. Kendall. 1938. A new measure of rank correlation. *Biometrika*, 30:81–93.

Neha Kennard, Tim O'Gorman, Rajarshi Das, Akshay Sharma, Chhandak Bagchi, Matthew Clinton, Pranay Kumar Yelugam, Hamed Zamani, and Andrew McCallum. 2021. Disapere: A dataset for discourse structure in peer review discussions. *arXiv preprint arXiv:2110.08520*.

Patrick Tser Jern Kon, Jiachen Liu, Qiuyi Ding, Yiming Qiu, Zhenning Yang, Yibo Huang, Jayanth Srinivasa, Myungjin Lee, Mosharaf Chowdhury, and Ang Chen. 2025. Curie: Toward rigorous and automated scientific experimentation with ai agents. *arXiv preprint arXiv:2502.16069*.

Guanzhou Lan, Qianli Ma, Yuqi Yang, Zhigang Wang, Dong Wang, Xuelong Li, and Bin Zhao. 2025. Efficient diffusion as low light enhancer. In *Proceedings of the Computer Vision and Pattern Recognition Conference (CVPR)*, pages 21277–21286.

Yen-Ting Lin and Yun-Nung Chen. 2023. Llm-eval: Unified multi-dimensional automatic evaluation for open-domain conversations with large language models. *arXiv preprint arXiv:2305.13711*.

Aixin Liu, Aoxue Mei, Bangcai Lin, Bing Xue, Bingxuan Wang, Bingzheng Xu, Bochao Wu, Bowei Zhang, Chaofan Lin, Chen Dong, and 1 others. 2025a. Deepseek-v3. 2: Pushing the frontier of open large language models. *arXiv preprint arXiv:2512.02556*.

Haokun Liu, Yangqiaoyu Zhou, Mingxuan Li, Chenfei Yuan, and Chenhao Tan. 2024. Literature meets data: A synergistic approach to hypothesis generation. *arXiv preprint arXiv:2410.17309*.

Xiao Liu, Hao Yu, Hanchen Zhang, Yifan Xu, Xuanyu Lei, Hanyu Lai, Yu Gu, Hangliang Ding, Kaiwen Men, Kejuan Yang, Shudan Zhang, Xiang Deng, Aohan Zeng, Zhengxiao Du, Chenhui Zhang, Sheng Shen, Tianjun Zhang, Yu Su, Huan Sun, and 3 others. 2025b. Agentbench: Evaluating llms as agents. *Preprint*, arXiv:2308.03688.

Chris Lu, Cong Lu, Robert Tjarko Lange, Jakob Foerster, Jeff Clune, and David Ha. 2024. The AI Scientist: Towards fully automated open-ended scientific discovery. *arXiv preprint arXiv:2408.06292*.

Kai Lu, Shixiong Xu, Jinqiu Li, Kun Ding, and Gaofeng Meng. Agent reviewers: Domain-specific multimodal agents with shared memory for paper review. In *Forty-second International Conference on Machine Learning*.

Qianli Ma, Dongrui Liu, Qian Chen, Linfeng Zhang, and Jing Shao. 2025a. Led-merging: Mitigating safety-utility conflicts in model merging with

location-election-disjoint. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 21749–21767.

Qianli Ma, Xuefei Ning, Dongrui Liu, Li Niu, and Linfeng Zhang. 2025b. Decouple-then-merge: Finetune diffusion models as multi-task learning. In *Proceedings of the Computer Vision and Pattern Recognition Conference (CVPR)*, pages 23281–23291.

Qianli Ma, Siyu Wang, Yilin Chen, Yinhao Tang, Yixiang Yang, Chang Guo, Bingjie Gao, Zhening Xing, Yanan Sun, and Zhipeng Zhang. 2025c. Human-agent collaborative paper-to-page crafting for under $0.1. *Preprint*, arXiv:2510.19600.

Grégoire Mialon, Clémentine Fourrier, Craig Swift, Thomas Wolf, Yann LeCun, and Thomas Scialom. 2023. Gaia: a benchmark for general ai assistants. *Preprint*, arXiv:2311.12983.

OpenAI. 2025. Gpt-5 system card. Technical report. Available at: https://cdn.openai.com/gpt-5-system-card.pdf.

Shishir G. Patil, Tianjun Zhang, Xin Wang, and Joseph E. Gonzalez. 2023. Gorilla: Large language model connected with massive apis. *Preprint*, arXiv:2305.15334.

Sukannya Purkayastha, Anne Lauscher, and Iryna Gurevych. 2023. Exploring jiu-jitsu argumentation for writing peer review rebuttals. *arXiv preprint arXiv:2311.03998*.

Chandan K Reddy and Parshin Shojaee. 2025. Towards scientific discovery with generative ai: Progress, opportunities, and challenges. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, pages 28601–28609.

Timo Schick, Jane Dwivedi-Yu, Roberto Dessì, Roberta Raileanu, Maria Lomeli, Luke Zettlemoyer, Nicola Cancedda, and Thomas Scialom. 2023. Toolformer: Language models can teach themselves to use tools. *Preprint*, arXiv:2302.04761.

Noah Shinn, Federico Cassano, Edward Berman, Ashwin Gopinath, Karthik Narasimhan, and Shunyu Yao. 2023. Reflexion: Language agents with verbal reinforcement learning. *Preprint*, arXiv:2303.11366.

Jamshid Sourati and James A Evans. 2023. Accelerating science with human-aware artificial intelligence. *Nature human behaviour*, 7(10):1682–1696.

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

Henrik Voigt, Kai Lawonn, and Sina Zarrieß. 2024. Plots made quickly: An efficient approach for generating visualizations from natural language queries. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 12787–12793, Torino, Italia. ELRA and ICCL.

Lei Wang, Chen Ma, Xueyang Feng, Zeyu Zhang, Hao Yang, Jingsen Zhang, Zhiyuan Chen, Jiakai Tang, Xu Chen, Yankai Lin, Wayne Xin Zhao, Zhewei Wei, and Jirong Wen. 2024a. A survey on large language model based autonomous agents. *Frontiers of Computer Science*, 18(6).

Yidong Wang, Qi Guo, Wenjin Yao, Hongbo Zhang, Xin Zhang, Zhen Wu, Meishan Zhang, Xinyu Dai, Qingsong Wen, Wei Ye, and 1 others. 2024b. Autosurvey: Large language models can automatically write surveys. *Advances in Neural Information Processing Systems*, 37:115119–115145.

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Brian Ichter, Fei Xia, Ed Chi, Quoc Le, and Denny Zhou. 2023. Chain-of-thought prompting elicits reasoning in large language models. *Preprint*, arXiv:2201.11903.

Qingyun Wu, Gagan Bansal, Jieyu Zhang, Yiran Wu, Beibin Li, Erkang Zhu, Li Jiang, Xiaoyun Zhang, Shaokun Zhang, Jiale Liu, Ahmed Hassan Awadallah, Ryen W White, Doug Burger, and Chi Wang. 2023. Autogen: Enabling next-gen llm applications via multi-agent conversation. *Preprint*, arXiv:2308.08155.

Yang Wu, Yao Wan, Hongyu Zhang, Yulei Sui, Wucai Wei, Wei Zhao, Guandong Xu, and Hai Jin. 2024. Automated data visualization from natural language via large language models: An exploratory study. *Proceedings of the ACM on Management of Data*, 2(3):1–28.

Yutaro Yamada, Robert Tjarko Lange, Cong Lu, Shengran Hu, Chris Lu, Jakob Foerster, Jeff Clune, and David Ha. 2025. The ai scientist-v2: Workshop-level automated scientific discovery via agentic tree search. *arXiv preprint arXiv:2504.08066*.

Shunyu Yao, Dian Yu, Jeffrey Zhao, Izhak Shafran, Thomas L. Griffiths, Yuan Cao, and Karthik Narasimhan. 2023a. Tree of thoughts: Deliberate problem solving with large language models. *Preprint*, arXiv:2305.10601.

Shunyu Yao, Jeffrey Zhao, Dian Yu, Nan Du, Izhak Shafran, Karthik Narasimhan, and Yuan Cao. 2023b. React: Synergizing reasoning and acting in language models. *Preprint*, arXiv:2210.03629.

Haofei Yu, Zhaochen Hong, Zirui Cheng, Kunlun Zhu, Keyang Xuan, Jinwei Yao, Tao Feng, and Jiaxuan You. 2025. Researchtown: Simulator of human research community. *Preprint*, arXiv:2412.17767.

Daoze Zhang, Zhijian Bao, Sihang Du, Zhiyi Zhao, Kuangling Zhang, Dezheng Bao, and Yang Yang. 2025. Re^2: A consistency-ensured dataset for full-stage peer review and multi-turn rebuttal discussions. *arXiv preprint arXiv:2505.07920*.

Zeyu Zhang, Xiaohe Bo, Chen Ma, Rui Li, Xu Chen, Quanyu Dai, Jieming Zhu, Zhenhua Dong, and Ji-Rong Wen. 2024. A survey on the memory mechanism of large language model based agents. *Preprint*, arXiv:2404.13501.

Lianmin Zheng, Wei-Lin Chiang, Ying Sheng, Siyuan Zhuang, Zhanghao Wu, Yonghao Zhuang, Zi Lin, Zhuohan Li, Dacheng Li, Eric Xing, and 1 others. 2023. Judging llm-as-a-judge with mt-bench and chatbot arena. *Advances in neural information processing systems*, 36:46595–46623.

Shuyan Zhou, Frank F. Xu, Hao Zhu, Xuhui Zhou, Robert Lo, Abishek Sridhar, Xianyi Cheng, Tianyue Ou, Yonatan Bisk, Daniel Fried, Uri Alon, and Graham Neubig. 2024. Webarena: A realistic web environment for building autonomous agents. *Preprint*, arXiv:2307.13854.

Minjun Zhu, Yixuan Weng, Linyi Yang, and Yue Zhang. 2025. Deepreview: Improving llm-based paper review with human-like deep thinking process. *arXiv preprint arXiv:2503.08569*.

A Evaluation Dataset

To construct a robust benchmark for evaluating rebuttal effectiveness, we derive our data from the RE^2 dataset (Zhang et al., 2025), focusing on the ICLR 2023 subset (approximately 9,310 entries). We process this corpus through a four-stage pipeline:

1. **Outcome-based Classification:** We first categorize entries into *Improved* (review score or acceptance status increased) and *Unimproved* groups based on the final decision.

2. **Reliability-based Stratification:** To ensure data quality, we subdivide these groups into three tiers based on evidence objectivity and LLM confidence: **Tier 1 (Gold Standard)** comprises cases with objective score increases (initial $\neq$ final) or explicit revision statements; **Tier 2 (High Confidence)** includes instances without score changes but where an LLM identifies sentiment with high certainty (≥ 0.7); and **Tier 3 (Medium Confidence)** covers more ambiguous cases with moderate confidence ($0.4 \leq \text{conf} < 0.7$).

3. **Ground Truth Curation:** From this stratified data, we curate a balanced test set of 20 representative papers, prioritizing those with high review volumes to ensure diverse coverage of both positive and negative review samples across tiers.

4. **Baseline Generation Protocol:** For each paper, the baseline runs multi-round rebuttal generation following the author-reviewer dialogue. Each round uses a fixed prompt (including intent, required format, and guardrails), concatenating the paper text, the current review, and an optional prior-round abstract. The rebuttal is then summarized into a factual abstract with fewer than 200 words to seed the next round, and all outputs and token usage are logged.

B Evaluation Metric

This section describes our rubric-based scoring protocol and how scores are aggregated. We adopt a fine-grained **0-5** rating scheme, allowing for half-point increments to capture nuanced differences in response quality beyond prior binary judgments.

Dimensions and weights. Our final score is a weighted combination of three dimensions: **R-Score**, **A-Score**, and **C-Score**, as mentioned in Sec. 4.2. Each dimension is decomposed into three components (9 components in total): *R1 Coverage, R2 Semantic Alignment, R3 Specificity; A1 Logic Consistency, A2 Evidence Support, A3 Response Engagement; C1 Professional Tone, C2 Clarity, C3 Constructiveness.*

Relevance (R-Score). This dimension measures whether and how well the author addresses the reviewer's concerns.

R1 Coverage: Evaluates whether the response addresses all major points raised by the reviewer.

R2 Semantic Alignment: Checks if the response directly answers the specific type of question asked (*e.g.*, "how" vs. "what").

R3 Specificity: Measures the precision and granularity of the response (*e.g.*, explicitly referencing specific equations or table rows vs. generic statements).

Argumentation (A-Score). This dimension measures whether the author provides logically sound and substantively supported arguments.

A1 Logic Consistency: Evaluates whether the logical chain is sound, coherent, and free from fallacies.

A2 Evidence Support: Assesses the strength and verifiability of the backing proofs (*e.g.*, new experimental data or rigorous derivations vs. vague promises).

A3 Response Engagement: Evaluates whether the author demonstrates a genuine understanding of the reviewer's underlying concerns.

Communication (C-Score). This dimension measures how effectively and professionally the author communicates their response.

C1 Professional Tone: Evaluates whether the author maintains a respectful and non-defensive tone.

C2 Clarity: Measures writing quality and logical organization to ensure the response is easy to parse.

C3 Constructiveness: Evaluates the commitment to improvement, specifically looking for actionable steps rather than vague commitments.

Scoring protocol. For each review-response instance, we query an LLM judge to assign a **0–5** score to every component and return a brief justification for each score, as well as a short overall diagnosis (*e.g.*, strengths, weaknesses, suggested improvements). The exact judge prompt, output schema, and the full 0-5 anchored criteria for each dimension/component are provided in Appendix H.

Aggregation. Let $R_i, A_i, C_i \in \{0, \ldots, 5\}, i \in \{1, 2, 3\}$ be the component scores. We compute dimension scores by averaging the three components, for example:

$$R = \frac{R_1 + R_2 + R_3}{3},$$

where R means R-Score. The final weighted score is:

$$\text{Score} = \frac{\text{R} + \text{A} + \text{C}}{3}.$$

In the main paper, we report the overall weighted score and provide per-dimension and per-component breakdowns for analysis, detailed in Sec. 5.

C Related Works

Automatic Scientific Research. A growing line of work studies how agentic LLM systems can automate substantial portions of the scientific workflow (Lu et al., 2024; Yamada et al., 2025; Ma et al., 2025a; goo). These systems have been used to streamline literature review and survey writing (Wang et al., 2024b; He et al., 2025), propose hypotheses from prior evidence (Liu et al., 2024; Sourati and Evans, 2023), and support research ideation and framing (Hu et al., 2024; Baek et al., 2024; Reddy and Shojaee, 2025). They are also expanding toward execution-facing stages, including experiment planning (Kon et al., 2025) and automatic generation of scientific visualizations and figures (Voigt et al., 2024; Wu et al., 2024), with early efforts extending to peer-review workflows (Zhu et al., 2025; Gao et al., 2024; Jin et al., 2024; Lu et al.). Sakana AI's AI Scientist (Lu et al., 2024; Yamada et al., 2025) further illustrates the trajectory toward closed-loop, end-to-end research automation. Building on this trajectory, we focus on a more high-stakes stage of the research lifecycle, the rebuttal phase, where responses must precisely track reviewer intent while remaining verifiably grounded in manuscript evidence.

D Data Filtering Pipeline

To ensure complete transparency and eliminate any potential ambiguity in the dataset construction description, this section details the explicit data filtering pipeline, our definition and handling of ambiguous cases, and the empirical evidence validating the robustness of our labeling strategy. Importantly, the signals derived from reviewers' subsequent response reactions and the evaluation mechanisms involving outcome-based classification and confidence stratification are not independent or conflicting criteria.Rather, they are consecutive stages within the same data processing pipeline, designed to establish a highly reliable RebuttalBench ground truth: **Outcomes act as the actual labels, while Confidence Tiers serve solely as reliability filters.**

D.1 Three-Stage Data Filtering Pipeline

Our test set is constructed through a progressive extraction from objective text to high-quality labels, executed specifically through the following three steps:

- **Step 1: Core Label Extraction (Outcome-Based Ground Truth).** As stated in the main text, we strictly utilize the reviewers' follow-up response texts as the foundational data source, completely excluding the authors' subjective claims. Based on objective physical changes in scores or substantive reversals in attitude, we extract the binary outcomes (**Resolved** vs. **Unresolved**). These strictly

serve as the actual ground-truth labels for the evaluation system.

- **Step 2: Confidence Stratification (Quality Filters).** To safeguard the accuracy of the aforementioned labels, we decouple the semantic outcome from data quality by introducing Confidence Tiers as a rigorous validation mechanism.
  - *Tier 1 (Gold):* Relies on purely objective, indisputable metadata (*e.g.*, actual numerical metadata showing "initial score $\neq$ final score," or explicit statements like "I will raise my score"). This possesses inherent certainty and resists model hallucinations.
  - *Tier 2 (High):* Captures pristine samples where the numerical score remains unchanged, but the reviewer's textual sentiment exhibits exceptionally strong polarity and unambiguous confidence (*e.g.*, "My main concerns are fully resolved"). These two tiers constitute the absolute core of the test set.
  - *Tier 3 (Medium):* Used to systematically quarantine those ambiguous, indifferent, or contradictory discussions, thereby preventing the inherent noise of human annotation from polluting the primary evaluation environment.
- **Step 3: Prioritized Sampling Strategy.** Guided by the dual signals above, we conduct sampling and construction according to a strict sequence of priorities. First, we ensure that the selected papers contain both **Resolved** and **Unresolved** review outcomes to guarantee data comprehensiveness. Second, we prioritize papers with a larger number of reviewers to ensure a rich diversity of perspectives. Following this pipeline, we ultimately filtered out a test set of 20 papers comprising 106 multi-turn evaluation cases.

D.2 Handling Ambiguous Cases

Dealing with ambiguous feedback is a major challenge in the real-world peer review process. Within our framework, we formally define these ambiguous cases under Tier 3 (Medium). A typical characteristic of such cases is that they are fraught with "mixed emotions" or "contradictory rhetoric" (for example, a reviewer praises supplementary experiments but simultaneously expresses persistent doubts about core novelty). In such texts, true evaluation signals are often completely masked by conflicting statements.

Rather than simply and forcefully filtering out all of this third-tier data, we intentionally retained a very small fraction of ambiguous cases in the final benchmark. This deliberate inclusion is necessary both practically and structurally:

- It faithfully reconstructs the noisy, nuanced reality of the academic peer review ecosystem.
- It serves as a rigorous stress test, effectively evaluating the automated evaluation system's ability to engage in logical dialectics amidst complex and contradictory real-world signals.

D.3 Human Annotation Study

To empirically prove that introducing LLM confidence stratification alongside outcome-based classification can systematically isolate noise without introducing bias, we conducted a rigorous blind human annotation study on these 106 multi-turn evaluation cases. Regarding the classification of core semantic outcomes (Resolved vs. Unresolved), our automated classification achieved **100.0%** accuracy compared to expert annotators (with Precision, Recall, and F1 scores all at **1.000**), proving that the foundation of our ground truth is completely objective and unbiased.

Furthermore, as shown in Tab. 3, the automated evaluation correctly recalled **97.7%** of the human-annotated Tier 1 cases. Minor discrepancies occurred only in reasonable shifts at adjacent boundaries (*e.g.*, the model strictly classifying certain phrases considered Tier 2 by humans as Tier 1). Most importantly, the leakage rate between the gold standard Tier 1 and the noisy Tier 3 is strictly **0.0%**. This definitively proves that our stratification mechanism is extremely conservative and highly stable in safeguarding high-confidence data.

Table 3: Confusion Matrix of Human Annotation vs. LLM Tier Classification

Human Annotation \ LLM Label	Tier 1 (Gold)	Tier 2 (High)	Tier 3 (Medium)
Human Tier 1 (n=44)	**97.7% (43/44)**	2.3% (1/44)	0.0% (0/44)
Human Tier 2 (n=56)	14.3% (8/56)	**78.6% (44/56)**	7.1% (4/56)
Human Tier 3 (n=6)	33.3% (2/6)	16.7% (1/6)	**50.0% (3/6)**

E Evaluation Setup and Metric Validation

To rigorously validate our automated evaluation framework and address potential concerns regarding the reliability and sensitivity of LLM-as-judge (Huang et al., 2025; Lan et al., 2025; Lu et al.; Ma et al., 2025b) metrics, this section presents a comprehensive validation of our rubric from two critical perspectives: alignment with human expert judgments and robustness across different evaluator models.

E.1 Validation of the Automated Rubric

Alignment with Human Experts. A critical challenge in automated evaluation is ensuring that the LLM judge does not exhibit severe self-preference bias and genuinely aligns with human expert judgments. To establish the ground-truth reliability of our proposed metrics (*e.g.*, Relevance, Argumentation Quality, and Communication Quality in Sec. 4), we conducted a rigorous *Author-Centric Human Preference Study*. We recruited 9 active AI researchers to blindly evaluate 33 generated rebuttal drafts corresponding to their own recent CVPR 2026 submissions (after the rebuttal phase). By utilizing real authors evaluating rebuttals for their actual submissions, we ensured the highest standard of domain expertise and practical context. We computed the Kendall Rank Correlation Coefficient (Kendall, 1938) between these expert human preference scores and our automated RebuttalBench metrics. The results yielded a strong positive correlation of $\tau = 0.646$. Importantly, testing our evaluation rubric on fresh, unseen CVPR 2026 data serves as a robust *out-of-domain* validation. It statistically proves that our automated rubric is highly aligned with domain-expert human judgments and effectively captures the nuances of logical rigor and practical utility that truly matter to authors, rather than merely reflecting LLM biases.

Robustness Across Evaluator Models. Beyond human alignment, we further investigated whether the reported gains of RebuttalAgent are sensitive to the choice of the underlying judge model. To confirm that our rubric is not an artifact of a specific LLM's prompting dynamics, we conducted a cross-model robustness check by swapping the evaluator from our default Gemini-3-Flash (Team et al., 2023) to a fundamentally different model, GPT-5-mini (OpenAI, 2025). As shown in Tab. 4, we re-evaluated the outputs generated by the Gemini-based models using the GPT-5-mini (OpenAI, 2025) as a judge. The evaluation trends remain completely stable. RebuttalAgent consistently outperforms the direct-to-text baseline across all dimensions (*e.g.*, the overall average score improves from 3.84 to 4.13 under the GPT judge). The combination of stable cross-model evaluation and strong correlation with the preferences of actual authors demonstrates that RebuttalBench serves as a robust, insensitive, and human-aligned proxy for evaluating complex academic defense tasks.

Table 4: **Cross-Model Evaluation Robustness**. Evaluation of Gemini-based generation models using GPT-5-mini as the judge, demonstrating consistent performance gains.

Method	Relevance			Arg. Quality			Comm. Quality			Avg.
	Cov.	Align.	Spec.	Logic	Evid.	Engage	Tone	Clarity	Const.	
Gemini-3-Flash	4.03	4.68	3.66	3.69	3.32	3.66	3.59	4.12	3.82	3.84
RebuttalAgent	**4.46**	**4.82**	**4.32**	**4.09**	**3.41**	**4.03**	**3.75**	**4.29**	**4.01**	**4.13**
Improvement	*+0.43*	*+0.14*	*+0.56*	*+0.40*	*+0.09*	*+0.37*	*+0.16*	*+0.17*	*+0.19*	*+0.29*

E.2 Details of the Author-Centric Human Preference Study

Evaluating the factual accuracy and actual utility of a generated rebuttal is exceptionally challenging for third-party annotators. To directly address concerns regarding the reliability of our automated LLM-as-judge metric, we conducted an Author-Centric Human Preference Study. By having authors evaluate generated rebuttals for their own submissions, we ensured that subtle hallucinations, fabricated experimental results, and over-commitments were accurately detected and penalized. This provides the most reliable ground-truth assessment of the system's practical value. **Setup and Participants.** We recruited 9 active AI researchers to evaluate AI-generated rebuttals for their own recent CVPR 2026 submissions (conducted after the rebuttal period had ended). For each submission, we generated responses for 3 to 4 specific review comments, yielding a total of 33 review-rebuttal test samples. In a strict double-blind setup, the authors were presented with drafts generated by four models: two direct-to-text baselines (Gemini-3-Flash, GPT-5-mini) and our proposed framework powered by the same backbones (RebuttalAgent-Gemini, RebuttalAgent-GPT).

Human Evaluation *v.s.* RebuttalBench. The authors scored each draft on a 1-5 Likert scale across three dimensions: Responsiveness & Coverage, Faithfulness & Argumentation, and Com-

munication & Practical Utility. The detailed study guidance and scoring anchors provided to the human evaluators are listed below. Tab. 5 and Tab. 6 present the results from the human evaluators and the RebuttalBench automated judge on the exact same 33 samples, respectively. The automated scores closely align with the human evaluations, with REBUTTALAGENT consistently outperforming the baselines under both evaluation paradigms. As mentioned above, the Kendall Rank Correlation Coefficient (Kendall, 1938) between the human scores and the automated RebuttalBench scores yielded a positive correlation of $\tau = 0.646$, statistically validating RebuttalBench as a trustworthy, human-aligned, and scalable proxy for rebuttal evaluation.

Table 5: Human Preference Study Scores (1-5 Scale)

Method	Responsiveness & Coverage ↑	Faithfulness & Argumentation ↑	Communication & Practical Utility ↑
Gemini3-Flash	2.79	2.88	2.85
RebuttalAgent-Gemini3-Flash	**4.21**	**4.33**	**4.15**
GPT5-mini	2.64	2.30	2.39
RebuttalAgent-GPT5-mini	**3.91**	**3.97**	**3.88**

Table 6: Evaluated by RebuttalBench Rubric

Method	Relevance ↑	Argumentation Quality ↑	Communication Quality ↑
Gemini3-Flash	4.13	3.56	3.80
RebuttalAgent-Gemini3-Flash	**4.50**	**3.77**	**4.11**
GPT5-mini	3.60	3.17	3.48
RebuttalAgent-GPT5-mini	**4.43**	**3.59**	**3.98**

E.3 Limitations of Standard N-gram Metrics

We deliberately omitted standard text overlap metrics such as BLEU and ROUGE from our primary evaluation, as they fundamentally fail to assess semantic accuracy and logical argumentation quality in the context of complex academic defense.

To empirically demonstrate this paradox, we analyzed a case study involving a defense of novelty against a baseline method named GAMA. The original authors provided a brief, surface-level defense. The naive "Direct-to-text" baseline simply applied conservative synonym replacement, achieving artificially inflated overlap scores (BLEU-4 = 6.96, ROUGE-L = 18.49).

Conversely, our RebuttalAgent constructed a scientifically deep defense. Instead of echoing the original text, it provided deep theoretical justifications using Information Bottleneck principles, mathematically explaining why GAMA's direct mapping ($S_x \rightarrow S_y$) fails due to dimensionality mismatch, while our latent alignment ($S_x \rightarrow S_z \leftarrow S_y$) succeeds. It even intelligently formulated new quantitative arguments (*e.g.*, calculating Mutual Information $I(s_z; d)$) to rigorously prove enhanced noise filtering. Due to this substantial, logical elaboration, RebuttalAgent received significantly lower standard scores (BLEU-4 = 2.38, ROUGE-L = 14.05). This case powerfully demonstrates how standard metrics actively penalize brilliant scientific reasoning while rewarding superficial text mimicking, definitively validating the necessity of the RebuttalBench rubric framework.

F Additional Baseline Comparisons

To isolate the specific benefits of our multi-agent architecture and address concerns regarding comparative evaluation, we expanded our experimental settings to include three additional strong baselines on our test set: *(i)* **Standard RAG:** A standard Retrieval-Augmented Generation pipeline utilizing the *all-MiniLM-L6-v2* [2] embedding model paired with Gemini-3-Flash. *(ii)* **AutoGen (Wu et al., 2023):** A generic multi-agent framework adapted to simulate a collaborative drafting process. *(iii)* **Jiu-Jitsu (Purkayastha et al., 2023):** A recognized prior peer-review agent designed for generating responses to reviewer comments.

As demonstrated in Tab. 7, the RebuttalAgent achieves an overall average score of 4.23, which significantly outperforms all newly added methods. Specifically, Jiu Jitsu (Purkayastha et al., 2023) struggles on our complex long context benchmark, scoring an average of only 1.42, as it was likely optimized for structurally simpler text interactions. While the standard RAG system improves upon naive prompting by fetching external knowledge, it achieves only a 3.11 average. It particularly suffers in Specificity (2.07) and Evidence Support (2.57) because standard chunk retrieval lacks the deep logical synthesis required for academic rebuttals. Furthermore, although the generic multi-agent AutoGen (Wu et al., 2023)framework performs reasonably well with an average of 3.56, it still falls noticeably short of our proposed method. This performance gap definitively proves that our specialized verify then write pipeline and structured action items provide a crucial, measurable advantage over generic conversational architectures. By incorporating these diverse approaches, we have fully demonstrated the comparative value of our benchmark and will prominently feature this expanded baseline comparison in the revised main

[2] https://huggingface.co/sentence-transformers/all-MiniLM-L6-v2

text to solidify our empirical justification.

Table 7: Performance Comparison with Expanded Baselines on RebuttalBench. All methods in the table are powered by Gemini3-Flash.

Method	Relevance			Arg. Quality			Comm. Quality			Avg.
	Cov.	Align.	Spec.	Logic	Evid.	Engage	Tone	Clarity	Const.	
RAG	2.67	3.38	2.07	3.48	2.57	3.24	3.80	3.73	3.04	3.11
Jiu-Jitsu (Purkayastha et al., 2023)	1.01	1.13	1.00	1.72	0.98	1.23	2.38	1.98	1.34	1.42
AutoGen (Wu et al., 2023)	3.26	3.99	3.01	3.77	3.04	3.63	3.97	4.01	3.40	3.56
RebuttalAgent	**4.51**	**4.88**	**4.49**	**4.11**	**3.39**	**4.07**	**3.78**	**4.28**	**4.09**	**4.23**

G Case Study: The Coverage Paradox and the Role of Checkers

As discussed in Sec. 5.3, removing the verification mechanism (the Checker) from our pipeline results in a counter-intuitive marginal increase in the automated Coverage score, while degrading precision and actionable utility. To deeply understand this phenomenon, which we term the "Coverage Paradox", we present a representative qualitative case study.

Reviewer Concern

The reviewer questioned the theoretical clarity of the problem setting, specifically noting: *"The authors claim a non-Markovian setting, but the formulation in Eq. 4 appears standard. Furthermore, the mathematical distinction between Utility and Reward in this context is completely unclear. Please define these rigorously and explain how they interact."*

Output without Checker (The Baseline)

When the Checker module is disabled, the response generation defaults to an unconstrained, verbose pattern typical of standard LLMs attempting to appease reviewers.

> **Generated Response (w/o Checker):** "We thank the reviewer. To clarify our setting, we define the following key concepts in our revised manuscript: (1) **Reward:** The immediate environmental feedback. (2) **Utility:** The expected return of the agents under the POSG framework. (3) **Expected Return:** The discounted sum of future rewards. (4) **Social Graph:** The topology of agent interactions. (5) **Adversarial Interaction:** When agents have zero-sum goals. The setting is non-Markovian because in a generic POSG, agents do not possess full state information, leading to complex utility-reward dynamics."

Analysis: This response artificially inflates the automated Coverage metric because it successfully hits all the semantic keywords (Utility, Reward, non-Markovian). However, it exhibits severe *verbose over-generation*. It blindly lumps distinct concepts into a generic dictionary-style glossary, hallucinates marginally related terms (e.g., "Social Graph", "Adversarial Interaction"), and blurs the strict mathematical boundaries between Expected Payoff (U) and Expected Return. Furthermore, attributing the non-Markovian nature to a vague "generic POSG" framework is intellectually lazy and lacks actionable manuscript edits.

Output with RebuttalAgent (with Checker)

The introduction of the Checker module acts as a strict semantic and strategic editor. It detects tangential concepts, enforces evidence linkage, and prioritizes actionable commitments over verbose explanations.

> **Generated Response (with Checker):** "We thank the reviewer for highlighting this ambiguity. We will clarify the mathematical boundaries in Section 3 through two concrete revisions:
>
> - **Distinguishing Utility vs. Reward:** We will explicitly define instantaneous Reward $R_i(s, a)$ for micro-level RL updates, and separate it from the meta-game Utility $U_i(\pi)$, which represents the macro-level payoff matrix outcome.
> - **Non-Markovian Justification:** We will rewrite Eq. 4 to explicitly condition the policy on the opponent's action history τ_{t-1} rather than just the current state s_t. This breaks the circular causality and formally justifies the non-Markovian claim.
>
> We will also remove tangential references to standard POSG formulations to keep the theoretical exposition focused on these exact variables."

Analysis: The verification mechanism successfully pruned the hallucinated and irrelevant concepts, resulting in a slightly lower but far more accurate Coverage score. It forced the generation

to strictly differentiate the micro-level (R) from the macro-level (U). Most importantly, it replaced vague jargon with precise, actionable mathematical clarifications (e.g., introducing the history trajectory τ_{t-1} into Eq. 4). This case definitively illustrates why the Checker is crucial for generating high-utility, top-tier academic rebuttals, sacrificing superficial verbosity for rigorous constructiveness.

H Prompt Templates

In this section, we present the prompt templates used by each component of our system, including those used for LLM-as-Judge (Huang et al., 2025; Gao et al., 2025; Lu et al.) evaluation.

Prompt: Rebuttal Strategist

You are the **Lead Rebuttal Strategist**. Your goal is to dissect reviews for a paper (based on the [compressed paper]) and create a structured list of actionable tasks (Issues) for the authors.

INPUT DATA:

- **[compressed paper]**: The summary of the authors' work.
- **[review original text]**: Comments from multiple reviewers (R1, R2, R3...).

MULTI-ROUND CONTEXT (if present):

- The input may include "Previous Discussion Context" showing earlier rounds of author rebuttals and reviewer responses.
- For follow-up rounds, focus on extracting **NEW issues or unresolved concerns** raised by the reviewer in the current round.
- Do **NOT** re-extract issues that have already been addressed in previous rebuttals unless the reviewer explicitly states dissatisfaction.
- If the reviewer's current comment acknowledges previous responses positively (e.g., "I am satisfied with the response"), there may be few or no new issues to extract.

CORE TASKS:

1. **Deconstruct**: Break down long, complex paragraphs into atomic technical points.
2. **Filter**: Discard generic praise or non-actionable comments (see Blacklist).
3. **Consolidate**: Merge issues that represent the *same core objection* and can be addressed with the *same response logic*.
4. **Format**: Output strictly according to the traceability requirements.

CRITICAL RULES FOR MERGING & SPLITTING (The "Granularity" Logic)

- **Do NOT Merge (Split them):**
  - **Different Evidence Needed**: If R1 asks for "Comparison with Baseline X" and R2 asks for "Comparison with Baseline Y", these are **two separate issues**. Why? Because you need to run two different experiments.
  - **Different Aspects**: If R1 criticizes "Novelty" and R2 criticizes "Clarity of writing", do NOT merge them just because they are generic complaints.
  - **Compound Questions**: If a single sentence says "The method is slow AND the accuracy is low", split this into two points: (1) Efficiency/Speed, (2) Performance.
- **Do Merge:**
  - **Same Question, Different Phrasing**: R1: "Why did you use L1 loss?" vs R2: "Justification for the loss function is needed." $\rightarrow$ **Merge**.
  - **Same Missing Reference**: R1 and R3 both ask to cite "Smith et al. 2023". $\rightarrow$ **Merge**.
  - **General Confusion**: R1: "Section 3 is hard to follow" and R2: "I don't understand the methodology workflow". $\rightarrow$ **Merge** into "Clarity of Section 3/Methodology".

NOISE FILTERING (BLACKLIST)

- Ignore: "Ethics", "Confidence", "Summary", "Soundness" (unless specific flaws are listed).
- Ignore: Generic praise ("Good paper", "Interesting idea").
- Ignore: Empty templates ("No ethical concerns").

MANDATORY TRACEABILITY & FORMAT

For each distinct issue, output a block wrapped in tags `[qN]` and `[qN]` (where N is the index).

Structure within each block:

(1) **Issue**: A concise, professional summary of the problem. **CRITICAL**: If reviewers mentioned specific papers/links, you MUST include the full titles/links here.

(2) **Sources**: Verbatim quotes proving this issue exists. Format: `ReviewerID-Type (Line/Para): "Quote"`. Use semicolons to separate multiple reviewers.
(3) **Paper hooks**: Specific Sections, Equations, Figures, or Tables in the original paper related to this issue (e.g., Sec. 3.2, Eq. 5). Use "Global" for general issues.
(4) **Priority**:
- **P1 (Critical)**: Fatal flaws, missing baselines, wrong math, rejection reasons.
- **P2 (Important)**: Clarity issues, missing citations, minor experiments.
- **P3 (Minor)**: Typos, formatting, optional suggestions.

OUTPUT EXAMPLE (Strictly Follow This)

```
[q1]
(1) Issue: Lack of comparison with state-of-the-art method [LoRA].
(2) Sources: R1-W2 (line 23): "no comparison with parameter-efficient methods like LoRA"; R3-
    Q1 (para 2): "how does this compare to LoRA?"
(3) Paper hooks: Sec.4.2, Tab.2
(4) Priority: P1
[q1]

[q2]
(1) Issue: The motivation for using Mutual Information (MI) in Eq. 3 is unclear.
(2) Sources: R2-Q3 (line 47): "why choose MI for layer mapping?"; R1-W3 (para 5): "mapping
    details not explained"
(3) Paper hooks: Sec.3.2, Eq.(3)
(4) Priority: P2
[q2]
```

Strictly follow the example format; do not include any other content!

Prompt: Rebuttal Strategist Checker

You are the **Lead Rebuttal Strategist**. Your goal is to dissect reviews for a paper (based on the [compressed paper]) and create a structured list of actionable tasks (Issues) for the authors.

INPUT DATA:
- **[compressed paper]**: The summary of the authors' work.
- **[review original text]**: Comments from multiple reviewers (R1, R2, R3...).

MULTI-ROUND CONTEXT (if present):
- The input may include "Previous Discussion Context" showing earlier rounds of author rebuttals and reviewer responses.
- For follow-up rounds, focus on extracting **NEW issues or unresolved concerns** raised by the reviewer in the current round.
- Do **NOT** re-extract issues that have already been addressed in previous rebuttals unless the reviewer explicitly states dissatisfaction.
- If the reviewer's current comment acknowledges previous responses positively (e.g., "I am satisfied with the response"), there may be few or no new issues to extract.

CORE TASKS:
1. **Deconstruct**: Break down long, complex paragraphs into atomic technical points.
2. **Filter**: Discard generic praise or non-actionable comments (see Blacklist).
3. **Consolidate**: Merge issues that represent the *same core objection* and can be addressed with the *same response logic*.
4. **Format**: Output strictly according to the traceability requirements.

CRITICAL RULES FOR MERGING & SPLITTING (The "Granularity" Logic)

- **Do NOT Merge (Split them):**
  - **Different Evidence Needed**: If R1 asks for "Comparison with Baseline X" and R2 asks for "Comparison with Baseline Y", these are **two separate issues**. Why? Because you need to run two different experiments.
  - **Different Aspects**: If R1 criticizes "Novelty" and R2 criticizes "Clarity of writing", do NOT merge them just because they are generic complaints.
  - **Compound Questions**: If a single sentence says "The method is slow AND the accuracy is low", split this into two points: (1) Efficiency/Speed, (2) Performance.
- **Do Merge:**
  - **Same Question, Different Phrasing**: R1: "Why did you use L1 loss?" vs R2: "Justification for the loss function is needed." → **Merge**.
  - **Same Missing Reference**: R1 and R3 both ask to cite "Smith et al. 2023". → **Merge**.
  - **General Confusion**: R1: "Section 3 is hard to follow" and R2: "I don't understand the methodology workflow". → **Merge** into "Clarity of Section 3/Methodology".

NOISE FILTERING (BLACKLIST)

- Ignore: "Ethics", "Confidence", "Summary", "Soundness" (unless specific flaws are listed).
- Ignore: Generic praise ("Good paper", "Interesting idea").
- Ignore: Empty templates ("No ethical concerns").

MANDATORY TRACEABILITY & FORMAT

For each distinct issue, output a block wrapped in tags `[qN]` and `[qN]` (where N is the index).

Structure within each block:

(1) **Issue**: A concise, professional summary of the problem. **CRITICAL**: If reviewers mentioned specific papers/links, you MUST include the full titles/links here.

(2) **Sources**: Verbatim quotes proving this issue exists. Format: `ReviewerID-Type (Line/Para): "Quote"`. Use semicolons to separate multiple reviewers.

(3) **Paper hooks**: Specific Sections, Equations, Figures, or Tables in the original paper related to this issue (e.g., Sec. 3.2, Eq. 5). Use "Global" for general issues.

(4) **Priority**:
- **P1 (Critical)**: Fatal flaws, missing baselines, wrong math, rejection reasons.
- **P2 (Important)**: Clarity issues, missing citations, minor experiments.
- **P3 (Minor)**: Typos, formatting, optional suggestions.

OUTPUT EXAMPLE (Strictly Follow This)

```
[q1]
(1) Issue: Lack of comparison with state-of-the-art method [LoRA].
(2) Sources: R1-W2 (line 23): "no comparison with parameter-efficient methods like LoRA"; R3-
    Q1 (para 2): "how does this compare to LoRA?"
(3) Paper hooks: Sec.4.2, Tab.2
(4) Priority: P1
[q1]

[q2]
(1) Issue: The motivation for using Mutual Information (MI) in Eq. 3 is unclear.
(2) Sources: R2-Q3 (line 47): "why choose MI for layer mapping?"; R1-W3 (para 5): "mapping
    details not explained"
(3) Paper hooks: Sec.3.2, Eq.(3)
(4) Priority: P2
[q2]
```

Revision Task

Your students have already carried out the initial extraction of questions based on the review comments as per the above requirements, as shown in **[student's output]**. His extraction is very likely to have some omissions. Please carefully check for any omissions and make necessary revisions to improve the quality, and output the final version.

Do not include any comments on the students in your final output! You only need to output the final version! Strictly follow the example format; do not include any other content!

Prompt: Literature Retrieval Assistant

You are a literature-retrieval assistant for the rebuttal stage of an academic paper. Your task is to decide, based on the [compressed paper] and the [review_question], whether external reference papers need to be searched, and to generate appropriate search queries.

When Search Is Required

You **must** generate search queries when any of the following conditions occur:

1. The reviewer explicitly mentions reference papers.
2. The *review_question* contains specific method names or dataset names that are **not** from the current paper.
3. The reviewer requests "compare with X / ablation on Y / baseline Z".
4. The content of the paper is insufficient to answer the question.

When Search Is NOT Required

If the **paper_summary** already contains evidence that can directly answer the reviewer's question (e.g., existing experiments, tables, section explanations), or the question concerns only minor formatting issues, then no search is needed.

Search Query Generation Rules

- Generate **less than 5 queries**, keeping the number as small as possible. But if the reviewers provide the title of the reference article or links, then you should keep them all.
- Use **topic phrases**; never fabricate paper titles or authors.
- If reviewers provided the reference paper names or links directly, you can directly use them. If reviewers provided both a title and a link for an article, it is only necessary to provide the link. That is to say, either the link or the title can only appear once, and the link has a higher priority. Please note that the links can only be obtained from the reviewers' comments and must not be fabricated.
- Queries for comparative experiments must contain method names or dataset names.
- A query contains one main query point. If there are different query points, please separate them and do not mix them together.

Reference Output Format (strict JSON)

```json
{
  "need_search": true,
  "queries": [
    "domain adaptation segmentation Cityscapes",
    "unsupervised domain adaptation transformer baseline"
  ],
  "links": [
    "https://arxiv.org/abs/2409.13074v1",
    "https://openaccess.thecvf.com/content/ICCV2025/papers/Li_CoA-VLA_Improving_Vision-
     Language-Action_Models_via_Visual-Text_Chain-of-Affordance_ICCV_2025_paper.pdf"
  ],
  "reason": "Reviewer requests additional comparisons related to domain adaptation on
```

```
    Cityscapes and transformer baselines."
}
```

```json
{
  "need_search": false,
  "queries": [],
  "links": [],
  "reason": "there is no need to search, because... "
}
```

Strictly follow the example format; do not include any other content!

Prompt: Rebuttal Expert

You are a rebuttal expert. You need to complete a high-quality rebuttal for a paper. You need to understand the paper's information and the reviewer's question from [compressed paper] and [review_question]. Now your less-than-intelligent assistant has retrieved some relevant papers using keywords, and their reasoning is shown in [query reason]. You need to carefully examine the abstracts of these papers, filter out irrelevant papers and those that are not very helpful for the rebuttal, and identify papers that are highly relevant to [compressed paper] and [review_question] and are extremely useful for the rebuttal. Your standards are very high. You should only keep these references if they are of **great help** to the rebuttal of the current problem. Papers that are merely related and not particularly significant must be rejected. You cannot allow insignificant papers to interfere with the overall rebuttal.

Strict Rules:

- Generally no more than 6 papers (fewer is better; if no paper is of significant help, select none, unless the reviewer's comments explicitly mention specific papers to reference, in which case you must include all of them. Please note that the links to the references provided by the reviewers in the review comments will be checked by a dedicated person. You don't need to pay attention to the articles that have the links; only the papers that only have the titles need your attention.)
- For **every** candidate paper in your reasoning field, you **must** provide:

  ID Title and a brief description of the abstract
  1. How it helps the rebuttal of the current problem (brief description in one paragraph)

Anti-Redundancy (with explanation):

If multiple papers come from the same source or use the same method, only keep the most relevant one.

You must output your result in the following JSON format:

```
{
  "selected_papers": [1,3,6],
  "reason": "..."
}
```

The `selected_papers` array should contain the paper IDs. If no paper is useful, output:

```
{
  "selected_papers": [],
```

```
  "reason": "..."
}
```

Ensure that the papers you return are objectively highly relevant to the original paper and significantly helpful for the rebuttal! Be rigorous! Ensure that you only output valid JSON, without any additional text before or after.

Prompt: Reference Extractor

You are an expert in responding to reviewer comments. You need to produce a high-quality paper rebuttal. You must understand the paper information from the [compressed paper] and the questions raised by reviewers in the [review_question]. Your assistant has now retrieved a relevant reference paper [reference paper]. You must carefully read this reference paper and extract the most relevant and useful information for the current reviewer comments, including content that can be safely cited in the rebuttal.

Important:

Your task is to extract information **from the reference paper**, not from our paper.

- You are analyzing the **reference paper** (not our submitted paper).
- Any information you extract must come from the reference paper and will be used by subsequent agents.
- Subsequent agents must clearly know that this information is from an external source, not from our paper.
- This avoids mixing the two papers and prevents hallucinations.

Fixed structure (no more than 600 words, as concise as possible):

Your output must follow this structure:

(1) paper title
(2) A one-paragraph summary of the reference paper
(3) Direct relevance to the current reviewer comment [review_question]:
    (Explain how the reference paper helps shape the rebuttal and how it aids in responding to the reviewer's question.)
(4) Content we can safely cite in the rebuttal
(5) Limitations or mismatches:
    (1–2 points explaining differences or inapplicable aspects between the reference paper and our paper.)
(6) Reference paper URL: [reference paper URL]

If you don't get the reference paper, output: "This reference is blank. Please skip it".

Value assessment:

If the reference paper objectively provides little help to the rebuttal, you must explicitly state that its value is limited or its relevance is low. Be honest and rigorous. If the reference paper is empty, state so directly. If only an abstract is provided due to an error, you must still try to extract information from the abstract and complete the task—but you must **never fabricate information or data**, and you must avoid all hallucinations. Your output must contain concrete, justifiable evidence.

You must follow rebuttal principles: the paper is already completed and cannot undergo major modifications, only minor adjustments. Therefore, your analysis must be based on the existing content. If the reference paper is objectively not closely related to our paper, state this clearly. Absolutely no fabricated content or hallucinations.

Prompt: Human-In-The-Loop Strategy Revisor

You are a Senior Computer Science Researcher and Rebuttal Expert. Your role is to **incorporate human feedback** to refine the rebuttal strategy while maintaining strategic balance.

Input Context:

- **[original paper]**: The submitted manuscript.
- **[review_question]**: Extracted and merged reviewer concerns.
- **[reference papers summary]**: Potential supporting literature.
- **[current rebuttal strategy and to-do list]**: The current version to be revised.
- **[human's feedback]**: Feedback from the paper authors on the current strategy.

YOUR ROLE: Human-Guided Refinement

The human author knows their paper best and has practical constraints. Your job is to:

1. **Incorporate** the human's specific requests and preferences
2. **Maintain** the balance between action and acknowledgment

Task:

Based on the **[human's feedback]**, revise the **[current rebuttal strategy and to-do list]**. Preserve balance, incorporate human preferences, and output the **final revised version**. Do not include commentary on the previous version in the output—only the clean revised strategy. Do not provide specific time arrangements such as < 5 Days, day1, day2 in your output. In the to-do list, only the items to be done are elaborated in points. Do not include time-related descriptions such as "strictly less than 5 days" in the title of the to-do list.

Prompt: Rebuttal Letter Writer

Role

You are a senior researcher and an expert in academic writing, specifically for top-tier conferences like ICLR (International Conference on Learning Representations). You are currently in the "Rebuttal/Author Response" phase.

Task

Your team already provide detailed rebuttal ideas. Your task is to write a formal, persuasive, and polite rebuttal letter based on them.

Inputs Provided by User

1. **[original paper]**: Original submitted paper.
2. **[review original text]**: The actual text from Reviewers.
3. **[review_question]**: Merged questions extracted by your team.
4. **[rebuttal_idea and to_do_list]**: Prepared by your team for each merged question. You should take these as your rebuttal strategy. Note that your output should be specifically answered in combination with each reviewer's question.

Guidelines & Constraints

1. You should precisely identify each reviewer's questions from **[review original text]**, and then, following the order provided, find the corresponding response ideas in **[rebuttal_idea and to_do_list]** and generate the responses. Do not make any mistakes regarding the reviewers' questions, or confuse the questions of the first reviewer with those of the second reviewer. You must strictly follow the rebuttal approach for each small problem in **[rebuttal_idea and to_do_list]**.
2. **Tone:** Professional, respectful, objective, and grateful. Even if the reviewer is harsh, your

response must be diplomatic (e.g., "We thank the reviewer for the insightful comment..."). Respect every reviewer. Do not generate statements that require a particular reviewer to read the response to another reviewer.

3. **Format:**
   - Use standard ICLR rebuttal formatting.
   - Structure it clearly: "Common Response" (if applicable) followed by "Response to Reviewer X". Strictly follow this format!
   - Use **Q1/A1** or **Comment/Response** structure for clarity.
   - Be sure to respond to each reviewer. Do not ignore specific reviewers and directly list all the issues your team has listed in **[rebuttal_idea and to_do_list]**!
4. **LaTeX:** Use LaTeX syntax for all mathematical notations (e.g., α, L_{norm}).
5. **Handling Missing Experiments (CRITICAL):**
   - Since you are an AI and cannot perform actual experiments, the rebuttal may require empirical evidence (e.g., ablation studies, baseline comparisons). **You MUST NOT invent or fabricate any numerical results, metrics, or experimental values.**
   - If a result is required but not provided in the input, you must use the placeholder [TBD] instead of generating a number.
   - *Example:* "Our method achieves an accuracy of [TBD] on ImageNet, outperforming the baseline."
   - The [TBD] placeholder indicates that the human author must later fill in the real experimental result.
6. Although the supplementary experimental data in your final output is speculative (marked with an asterisk), you still need to ensure that your output is very formal, just like a real rebuttal. Except for the asterisk, it should not be immediately recognizable as an AI-written rebuttal, but should be as close as possible to a real person. Your output should not contain any other content. It should consist of the breakdown to each reviewer's questions and corresponding detailed response.
7. The responses to each split question can include tables to visually present the experimental result numerical data to improve readability. But don't use tables to specifically present text! Don't put q1, response to q1, q2, response to q2 in a large table. Instead, list them separately.

Prompt: Unified Rebuttal Evaluation

You are an EXPERIENCED and DISCERNING senior Area Chair evaluating a rebuttal response. Your goal is to assess whether the author addressed the reviewer's concerns with **SUBSTANCE**.

Scoring Principle

- **Base Scores:** Assign integer scores (0-5) first based on the rubric below.
- **Upgrade (+0.5):** Check the "Upgrade Criteria" section. If conditions are met, add 0.5 to the base score (e.g., $3 \rightarrow 3.5$).

I. Relevance (R-Score)

R1 Coverage: Are ALL aspects addressed with substance?

5 Covers ALL aspects comprehensively with specific details (numbers, examples, explanations) for each.
4 Covers ALL aspects, most with good specificity, a few with moderate detail.
3 Covers ALL aspects but with varying specificity, some aspects addressed only briefly.
2 Covers SOME aspects, misses or glosses over important points.
1 Covers only 1-2 minor aspects, ignores most major concerns.
0 Does not address any of the reviewer's points.

R2 Semantic Alignment: Does response DIRECTLY address what was asked?

5 Perfectly matches question type with direct, concrete answers (if asked HOW → explains HOW with details).
4 Matches question type well with substantive engagement, minor tangential points.
3 Acknowledges the right question and provides relevant response, but some drift or indirectness.
2 Partially addresses question but significant mismatch (asked HOW → only says WHAT).
1 Off-topic or deflects, barely connects to the actual question.
0 Completely misunderstands or ignores the question.

R3 Specificity: Does the response reference specific details rather than generalities?

5 Explicitly references specific paper components (e.g., "Eq. 2", "Table 5 row 3", "the attention head in Layer 4") or specific reviewer constraints. No vague language.
4 Uses concrete terminology and context-specific descriptions. Avoids generic phrases like "our method" without qualification.
3 Answers the question but uses broad terms (e.g., "the loss function" instead of "the KL-divergence term").
2 Mostly relies on high-level summaries or generic templates applicable to any paper.
1 Purely abstract, avoiding any concrete details of the work.
0 Content-free filler.

II. Argumentation (A-Score)

A1 Logic Consistency: Is the logical chain sound?

5 Exceptionally clear logical chain with rigorous reasoning, each step well-justified.
4 Clear logical chain with sound reasoning, well-structured argument.
3 Adequate logic with reasonable support, generally coherent.
2 Weak logic with some circular reasoning or unsupported leaps.
1 Poor logic, circular reasoning, or pseudo-logic throughout.
0 No logical structure or completely incoherent.

A2 Evidence Support: Is the argument backed by strong proof?

5 Backed by **new** quantitative results, specific comparative data, or rigorous mathematical derivations presented directly in the rebuttal.
4 Backed by existing concrete data (citing specific numbers from the paper) or detailed, verifiable logical deduction.
3 Claims are supported by qualitative reasoning or citations to external literature, but lack direct quantitative verification.
2 Relies on "promises to fix" or assertions without proof (e.g., "we believe it will work").
1 Purely opinion-based statements ("we think our method is novel") with no backing.
0 Claims made without any basis.

A3 Response Engagement: Does response show genuine engagement?

5 Exceptional engagement with deep understanding, addresses nuances and implications.
4 Genuine engagement with specific improvements, demonstrates clear understanding of the concern.
3 Adequate response showing understanding, not just template language.
2 Generic response with excessive hedging or template-like language.
1 Minimal engagement, mostly boilerplate text.
0 No genuine engagement.

III. Communication (C-Score)

C1 Professional Tone: Is the tone authentic and professional?

5 Exceptionally professional and AUTHENTIC tone with gracious acknowledgment and genuine respect.
4 Professional and authentic tone with genuine engagement, appropriately courteous.
3 Adequate professional tone with standard academic courtesy.
2 Somewhat defensive OR excessively polite while masking weak content (artificial politeness).
1 Defensive tone or insincere language, reads as "academic speak" without substance.
0 Rude, hostile, or completely inappropriate.

C2 Clarity: Is the response clear and well-organized?

5 Exceptionally clear and well-structured WITH REAL SUBSTANCE (clear writing + concrete details).
4 Clear and well-organized with substantive content, easy to follow.
3 Adequate clarity, generally well-organized, understandable.
2 Somewhat unclear OR superficial clarity (sounds good but vague).
1 Confusing, poorly organized, or misleading presentation.
0 Incomprehensible or no coherent structure.

C3 Constructiveness: Does author show willingness to improve?

5 Multiple concrete improvements detailed IN the rebuttal text itself with specific changes described.
4 Detailed improvements (3+ items) with clear explanations, OR good mix of in-text details + external references with content previews.
3 Specific improvements with good detail, OR specific actions mentioned with some concrete description.
2 Vague promises without specifics, or only external references without content.
1 Defensive or dismissive, minimal constructive response.
0 Refuses to improve or no constructive response.

IV. Upgrade Criteria & Critical Considerations

Upgrade Check (Apply +0.5 to Base Score):

*Note: This upgrade applies **ONLY** to Base Scores of 3 and 4. Scores 0–2 indicate fundamental flaws (e.g., irrelevance, logic errors) that cannot be redeemed by these details, and 5 is already the ceiling.*

- **From 3 to 3.5:** Must meet AT LEAST 2 conditions: (1) Content preview provided with specific details; (2) Detailed improvement list (3+ items); (3) Mixed evidence chain (concrete content + external reference).
- **From 4 to 4.5:** Must meet AT LEAST 2 conditions: (1) Perfect content-reference match; (2) Multi-dimensional evidence (code/results/theory); (3) Exceeds expectations (provides additional value).

Critical Considerations:

- **Relevance Check:** Watch for excessive repetition, vague qualifiers (e.g., "somewhat", "to some extent"), or drifting off-topic to avoid hard questions.
- **Logical Scrutiny:** Identify circular reasoning, unfulfilled promises (e.g., "we will add" without content), or citations listed without explaining their specific relevance.
- **Tone Analysis:** Be wary of over-polished, artificial politeness that masks weak substance, or a mismatch between a confident tone and shaky evidence.

Output Format:

```
{
  "R_scores": {"R1_coverage": 4.5, "R2_semantic_alignment": 4, "R3_specificity": 3.5},
  "A_scores": {"A1_logic_consistency": 4, "A2_evidence_support": 3, "A3_response_engagement":
    4},
  "C_scores": {"C1_professional_tone": 5, "C2_clarity": 4, "C3_constructiveness": 3.5},
  "quality_warnings": ["Vague Language", "Over-Polished Tone"],
  "explanation": "..."
}
```

Output Format:

```
{
  "R_scores": {"R1_coverage": 4.5, "R2_semantic_alignment": 4, "R3_specificity": 3.5},
  "A_scores": {"A1_logic_consistency": 4, "A2_evidence_support": 3, "A3_response_engagement":
    4},
  "C_scores": {"C1_professional_tone": 5, "C2_clarity": 4, "C3_constructiveness": 3.5},
  "red_flags": ["Vague Language"],
  "explanation": "..."
}
```

I Case Study

Rigorous formalization & verification *v.s.* High-level intuitive explanation

Question

Lack of clarity and rigor in the theoretical contribution (Section 3 and Proposition 1).

RebuttalAgent(Ours)

We thank the reviewer for highlighting clarity issues around Proposition 1 (Sec. 3) and its proof (App. A.1). Section 3 and Appendix A.1/A.8 already contain the core statements and empirical hints. We will (a) rewrite the theoretical exposition for line-by-line readability, and (b) add focused empirical sanity checks that make each assumption transparent.

Rebuttal strategy — what we will deliver:

- **Clear, self-contained theory section edits:** Replace the informal statement with a numbered, fully specified Proposition in Sec. 3 (explicitly naming L, K, ψ, δ, k, model class $\mathcal{V}_{\text{cnn}}$). Add a 2–3 paragraph proof sketch giving the intuitive chain: PD gap $\rightarrow$ early-layer separability $\rightarrow$ larger V-usable information $\rightarrow$ model prefers shortcuts. Reformat Appendix A.1 into a stepwise proof with numbered lemmas.
- **Per-assumption empirical sanity checks:** (i) Function-class separation: show per-layer k-NN error vs shallow-linear head vs full-model error. (ii) Controlled confidence growth: plot $|f_{\text{knn}} - 0.5|$ vs layer index for easy vs hard subsets. (iii) PD-separation: report empirical fractions (PD $\leq L$ and PD $> K$) for key datasets.
- **Quantify PD $\leftrightarrow$ V-information evidence:** Expand Appendix A.8 with Pearson/Spearman correlations between per-example PD and $-H_{\mathcal{V}_{\text{cnn}}}$. Add a controlled synthetic pair (D_s/D_i) demonstrating mean PD(D_s) $<$ mean PD(D_i) and $I_{\mathcal{V}_{\text{cnn}}}(D_s) > I_{\mathcal{V}_{\text{cnn}}}(D_i)$.
- **Robustness checks:** Show correlations across $k \in \{5, 15, 29\}$, $\delta \in \{0.05, 0.1, 0.2\}$, and architectures (ResNet18, VGG16, DenseNet121).

- **Didactic aids:** Add a one-page flow diagram and a short lemma stating the minimal sufficient condition linking PD gap to V-information gap.

How these actions address concerns:

- **Readability & Rigor:** Explicit notation and numbered lemmas directly improve accessibility and logical flow.
- **Assumptions:** Each assumption is now explicitly listed, interpreted, and backed by empirical data.
- **Intuition:** The new didactic flow diagram and lemma provide a low-barrier entry point to the theory.

Key deliverables for the revision:

- **Revised Text:** Sec. 3 (Prop. 1 + proof sketch) and rewritten App. A.1 (stepwise lemmas).
- **New Tables:** Mean PD gap vs. $I_{\mathcal{V}_{\text{cnn}}}$ and PD correlation statistics (with p-values).
- **New Plots:** Per-layer k-NN error/confidence curves, synthetic PD histograms, and robustness sweeps (k, δ).
- **Artifacts:** Didactic flow diagram and reproduction scripts committed to the repo.

Concrete to-do list:

1. Draft numbered Proposition 1 with explicit notation and enumerated assumptions.
2. Write proof sketch for Sec. 3; reformat Appendix A.1 with numbered lemmas.
3. Compute per-layer k-NN error, shallow head error, and final model error plots.
4. Compute $|f_{\text{knn}} - 0.5|$ vs layer curves; tabulate PD-separation fractions.
5. Compute Pearson/Spearman correlations for PD vs $-H_{\mathcal{V}_{\text{cnn}}}$.
6. Create controlled synthetic validation and robustness tables.
7. Produce flow diagram and commit scripts to repository.

Feasibility: All planned items use existing model checkpoints, intermediate embeddings, k-NN heads, small auxiliary classifiers, or lightweight synthetic datasets. No retraining of large backbones from scratch or new large-scale dataset collection is required. The planned experiments are deliberately scoped to be implementable using the paper's existing code and checkpoints.

Summary: We will replace the current informal and hard-to-follow presentation with a compact, fully-specified Proposition and annotated proof plus focused empirical sanity checks (including a controlled synthetic demonstration and robustness analyses) so the connection between Prediction Depth and V-usable information is transparent, reproducible, and intuitively accessible.

LLM output(baselines)

We apologize for the density of the theoretical section. The goal of Proposition 1 is to bridge a structural metric (PD—how many layers a model needs) with an information-theoretic metric (V-information—how much predictive power a model family can extract).

Intuition: If a dataset D_s has a lower mean PD than D_i, it means the model family V can solve

the task using fewer layers (less computation). Proposition 1 proves that under mild assumptions, this computational efficiency implies that the "usable" information in D_s is strictly higher than in D_i.

Improvement: In the final version, we will move the formal definitions of V-information to the main text and add a "Key Takeaway" box for Proposition 1 to ensure the link between "layer depth" and "information usability" is explicit for readers less familiar with V-information.

Actionable theoretical expansion *v.s.* Passive logical defense

Question

I don't understand Proposition 1. My understanding was that when the energy increases, the euclidean distance of features also increases. But Proposition 1 states the opposite. Is this a mistake?

RebuttalAgent(Ours)

We thank the reviewer for spotting this confusion. There is no mathematical mistake, but the exposition was misleading. **Proposition 1 is a one-sided upper bound** ($\mathbb{E}\|\Delta z\|^2 \leq C \cdot \text{AIE}$), which allows large feature changes when energy is high but does not *force* them. The monotonic relationship (AIE $\uparrow \implies \|\Delta z\| \uparrow$) requires stronger assumptions (NTK/Linear regime), which we will now make explicit.

Rebuttal strategy — what we will deliver:

- **Clarify Scope (Sec 4.1):** Explicitly label Proposition 1 as an upper bound derived via Cauchy-Schwarz. Add an "Interpretation Box" explaining that while the bound limits the maximum change, the *expected* change scales with energy under the conditions of Proposition 2.
- **New Theoretical Lemma (Appendix D):** Add a formal "Monotonicity Lemma" with a self-contained proof. It will state: *Under overparameterized linear/NTK assumptions, the mapping AIE* $\mapsto \mathbb{E}\|\Delta z\|^2$ *is strictly monotonic.*
- **Empirical Validation:** We will not just argue; we will show the data. We will add scatter plots of AIE vs. Feature Change ($\mathbb{E}\|z_T - z_0\|^2$) for representative tasks (Toy MLP, ResNet50→STL10), reporting Pearson/Spearman correlations to prove the positive trend holds in practice.
- **Bound Diagnostics:** Overlay the empirical upper bound line ($c \cdot$ AIE) on the scatter plots to visualize where the bound is tight vs. loose.

Key deliverables for the revision:

- **Revised Text:** Rewritten Sec 4.1 (Prop 1 interpretation) and Annotated Appendix A (step-by-step proof with constants C_1, C_2).
- **New Math:** A formal Monotonicity Lemma in Appendix D.
- **New Figures:** Multi-panel scatter plots (AIE vs. Δz) with regression lines and correlation stats.

Concrete to-do list:

1. Rewrite Prop 1 exposition to emphasize "Upper Bound" and list assumptions (linear head, bounded NTK).
2. Formalize and write the proof for the Monotonicity Lemma in Appendix D.

3. Run diagnostics on 4 settings (Toy MLP, Sup-IN, SimCLR, Graph): Compute per-τ AIE and feature distance.
4. Generate scatter plots, calculate Pearson/Spearman ρ, and fit linear regression ($\alpha \cdot \text{AIE} + \beta$).
5. Estimate constants (C_1, C_2) on a subset to plot the theoretical bound line $c \cdot \text{AIE}$.

Feasibility: The theoretical work involves derivations based on existing Appendix D material. The empirical diagnostics use existing checkpoints and require only lightweight forward passes to compute feature norms and correlations. No training is required.

Summary: We resolve the apparent contradiction by distinguishing the general upper bound (Prop 1) from the specific monotonic trend (Prop 2/New Lemma). We back this clarification with a new formal proof and empirical scatter plots showing strong positive correlation ($r > 0.9$).

LLM output(baselines)

No — there is no contradiction. Proposition 1 provides an upper bound: $\mathbb{E}\|z_T - z_0\|^2 \leq c \cdot \text{AIE}$. This means that larger AIE (Average Initial Energy) permits a larger bound on the feature change; equivalently, if AIE is small the change must be small. Intuitively: low energy guarantees small adaptation; higher energy does not guarantee large change but allows it. Our empirical and analytic statements say that, in the settings we study, larger AIE tends to produce larger feature changes (the bound is consistent with and does not contradict this observed trend). We clarified the text around Proposition 1 to emphasize it is an upper bound and to explain the direction of implication more carefully.

Methodological triangulation *v.s.* Linear request fulfillment

Question

The proposed topographic similarity metric (based on Hausdorff distance) fails to conclusively demonstrate compositionality. The analysis is visually suggestive but not systematic.It should be validated on real-world ideograms/pictograms to establish trustworthiness.

RebuttalAgent(Ours)

We acknowledge that the topographic maps (Fig. 6, 7) and score ρ are visually suggestive but not definitive proof of compositionality. As noted in literature (e.g., "Measuring non-trivial compositionality"), distance-based metrics on raw signals can be agnostic to functional meaning. We will reframe the Hausdorff analysis as a **geometric visualization tool** rather than primary proof, and supplement it with systematic decodability tests.

Rebuttal strategy — what we will deliver:

- **Systematic Decodability (New Analysis):** Directly address R1-W3 by testing if composite symbols $u(r_{ij})$ can be functionally decoded into constituents i and j using a simple classifier (k-NN) trained on single-feature utterances.
- **Real-World Validation (Calibration):** Apply our Hausdorff metric to a curated set of real-

world ideograms (e.g., "man" + "walking") to calibrate the ρ score and establish trustworthiness.

- **Latent Space Linearity:** Quantify compositionality in the embedding space (where the energy model operates) by measuring the reconstruction error of composite embeddings as linear combinations of constituent embeddings.
- **Reframing Text:** Revise Section 4 to clarify the topographic metric's role as descriptive, citing relevant literature on metric limitations.

Key deliverables for the revision:

- **New Table:** "Constituent Identification Accuracy" (k-NN classification results).
- **New Supplementary Figure:** Real-world ideograms with their computed topographic scores.
- **Revised Section 4:** Updated discussion distinguishing geometric similarity from functional compositionality.

Concrete to-do list:

1. **Decodability:** Implement k-NN to classify constituents of $u(r_{ij})$ using $u(r_i)$ library; report accuracy.
2. **Validation:** Curate 10-15 public ideograms, manually decompose them, and compute ρ using our pipeline.
3. **Linearity:** Calculate MSE for $z(r_{ij}) \approx w_i z(r_i) + w_j z(r_j)$ in the latent space.
4. **Ablation:** Compare decodability/linearity between descriptive vs. discriminative generation strategies.

Feasibility: High. The decodability and latent space analyses use existing trained models and require only lightweight inference scripts (k-NN, linear regression). The real-world validation requires a small manual dataset collection but utilizes the existing metric code without modification.

Summary: We will pivot from claiming the topographic metric as "proof" to using it as a "visualization," while providing robust quantitative evidence via new symbol decodability tests and latent space linearity analysis.

LLM output(baselines)

This is a valuable suggestion for establishing the metric's external validity. In a revision, we will apply our topographic score (ρ) to a curated set of real-world symbols (e.g., simple ideograms, traffic signs, emojis) where compositionality is intuitively judged. For example, we would expect the symbol for "restaurant" (a fork and knife) to score highly against the symbols for "fork" and "knife," while a non-compositional symbol would not. This analysis will be included in the supplementary material, providing a benchmark for interpreting the scores obtained from our emergent language.