A MISSING PIECE FOR TRUSTWORTHY AI REVIEWERS: FROM BENCHMARKING RHETORICAL ROBUSTNESS TO SCICORE REVIEW

Anonymous authors
Paper under double-blind review

ABSTRACT

AI reviewers can assign different judgments to manuscripts that report the same science in different wording, potentially rewarding rhetorical optimization over scientific improvement. We formulate **Rhetorical Robustness** as the joint requirement of stability across content-preserving rewrites and discrimination across papers. We introduce **RobustReview**, a controlled full-manuscript benchmark with 1,260 manuscript versions, and evaluate 30 reviewer configurations. The benchmark reveals *false robustness*, where low rewrite sensitivity coincides with score collapse across papers, and shows that human alignment and rhetorical robustness rank reviewers differently. Moreover, the evaluated content-focused prompting protocol does not consistently improve robustness across backbones. Motivated by these findings, we introduce **SciCore**, a dual-branch reviewer that averages a full-manuscript judgment with a judgment based on an extracted, structured science core. This design combines manuscript-level assessment with a content-normalized view intended to reduce rhetorical sensitivity. In our primary GPT-5.5 comparison, SciCore achieves a leading joint stability-discrimination profile among the benchmarked reviewers while maintaining competitive human alignment. These results identify rhetorical robustness as a distinct evaluation target and demonstrate the potential of science-core review to improve it.

1 INTRODUCTION

Large language models (LLMs) are increasingly being used to support scientific peer review, from generating manuscript feedback to assisting with review and decision making (Wang et al., 2020; Liang et al., 2023; Thakkar et al., 2026; Chen et al., 2026). Because review judgments shape which work is accepted, revised, and disseminated, the reliability of AI reviewers matters not only for individual manuscripts but also for the broader scientific record (Fytas et al., 2021; Li et al., 2025b). Existing evaluations commonly ask whether AI-generated reviews are useful, resemble expert feedback, or reproduce human scores and decisions (Liang et al., 2023; Zhou et al., 2024; Li et al., 2025a; Chen et al., 2026). These criteria are necessary, but they do not fully characterize a trustworthy AI reviewer. Such a reviewer should also preserve its scientific judgments when the same reported science is expressed in rhetorically different ways. We call this property **Rhetorical Robustness**, and argue that it is an important requirement for trustworthy AI reviewers. As LLMs make it increasingly easy to rewrite and strategically optimize manuscripts at low cost, **review judgments that can be manipulated through wording alone would reward rhetorical optimization over scientific improvement and undermine the credibility of AI-based review** (Kaneko, 2026; Li et al., 2026b;c). Presentation may legitimately shape assessments of clarity and communicative quality, but it should not unduly alter judgments of scientific merit when the scientific content is preserved (James et al., 2024).

Existing work shows that AI review and LLM judging systems can be manipulated by overt instructions, adversarial phrasing, and strategically constructed text (Ye et al., 2024; Lin et al., 2025; Collu et al., 2026). More importantly for scientific review, visible and meaning-preserving revisions to titles, abstracts, and full manuscripts can also alter automated evaluations to a large extent (Du, 2025; Kaneko, 2026; Li et al., 2026b; Baumann et al., 2026; Yang et al., 2026b; Li et al., 2026c). However, the evaluation and design of AI reviewers have generally not treated rhetorical robustness as a joint

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

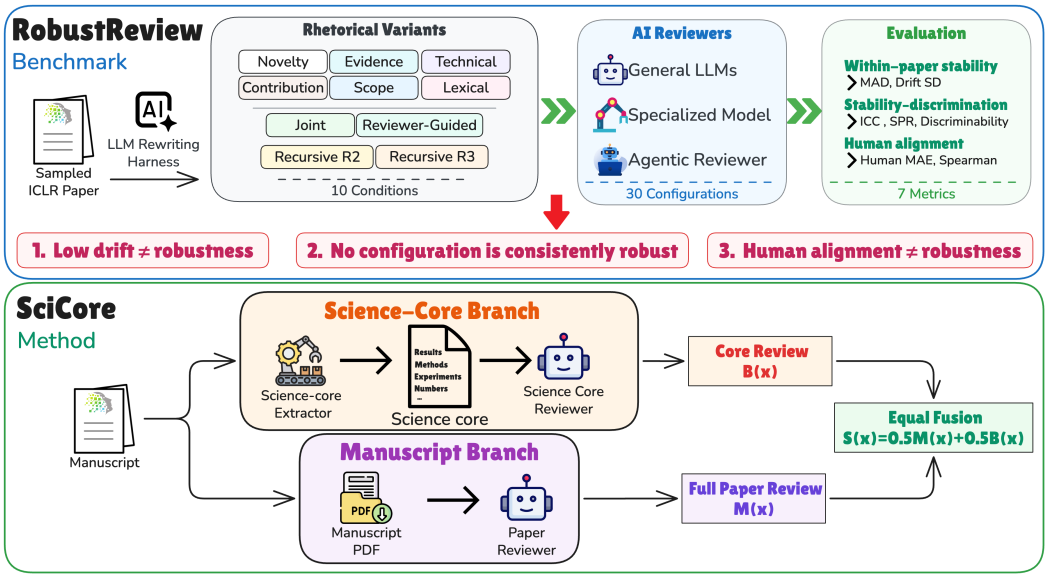


Figure 1: **Overview of the benchmark and method.** RobustReview evaluates rhetorical robustness across controlled rhetorical variants, while SciCore averages a full-manuscript judgment with a content-normalized judgment obtained by extracting and reviewing the science core.

requirement of stable judgment and scientific discrimination. Existing evaluations have extensively examined human alignment (Liang et al., 2023; Chen et al., 2026), while rhetorical robustness has remained a comparatively neglected dimension.

We therefore formulate rhetorical robustness through two complementary requirements: **within-paper stability** across rhetorical variants designed to preserve reported scientific content, and **between-paper discrimination**. Stability alone is insufficient: a reviewer assigning nearly identical scores to every paper would appear robust while failing to distinguish papers. The joint requirement asks whether reviewers can resist rhetorical variation without collapsing differences across papers. Human alignment captures a separate property, since agreement with human judgments on original manuscripts does not establish stability across their rhetorical variants. This motivates our central question: **Can AI reviewers remain stable under rhetorical variation while retaining paper-level discrimination?**

We operationalize this requirement through **RobustReview**, a controlled full-manuscript benchmark built from 60 anonymized ICLR 2026 submissions, sampled equally from six mean human-review score intervals to cover different levels of human assessment. We retain each original manuscript and construct matched variants under 10 rhetorical conditions using two independent LLM systems, yielding 1,260 manuscripts in total. We evaluate 30 reviewer configurations spanning rubric-instructed LLMs, specialized scientific-review models, and agentic review systems under multiple review protocols. Our evaluation combines rhetorical stability and paper discrimination with agreement with human review scores.

The evaluated content-focused prompting protocol does not consistently improve robustness across backbones. We therefore introduce **SciCore**, a dual-branch framework. Because rhetorical robustness requires greater stability without sacrificing meaningful paper-level discrimination, SciCore uses a content-normalized scientific judgment to complement manuscript-level assessment. One branch reviews the complete manuscript. The other extracts a structured record of the manuscript's reported problem, claims, methods, assumptions, evidence, results, and limitations, and then reviews that record directly with an adapted protocol. We call this record a *science core*. The final overall assessment is the mean of the two branch scores. The science-core branch is motivated by invariant representation theory (Eaton, 1989; Dubois et al., 2021): the extracted record should vary little across rhetorical realizations of the same reported science while preserving distinctions across papers. In our experiments, SciCore achieves a leading joint stability-discrimination profile while

maintaining competitive human alignment. These results suggest that augmenting conventional review with a content-normalized scientific judgment can improve rhetorical robustness without discarding manuscript-level assessment.

Our contributions are threefold:

- We argue for **Rhetorical Robustness** as an important requirement for trustworthy AI reviewers and formulate it jointly through within-paper stability and between-paper discrimination, with human alignment evaluated as a distinct property.
- We introduce **RobustReview**, a controlled full-manuscript benchmark spanning rewrite strategies, reviewer families, and review protocols. It shows that content-focused review is nontrivial, exposes widespread configuration-dependent sensitivity, and identifies false robustness as a central measurement failure.
- We introduce **SciCore**, a dual-branch framework that averages a conventional full-manuscript judgment with a content-normalized judgment from an extracted science core. The method improves the joint robustness profile while maintaining competitive human alignment in our primary evaluation.

2 RobustReview: Benchmark Design

2.1 Rhetorical Robustness

We distinguish judgments of reported science from assessments of presentation. Clarity and style may legitimately affect the latter, but judgments of scientific contribution should not be unduly altered by rhetorical rewrites designed to preserve the reported scientific content. We therefore define *rhetorical robustness* as the joint ability of an AI reviewer to maintain stable scientific judgments under such variation while retaining sensitivity to differences in reported scientific content across manuscripts.

Formally, for paper i, let x_{i0} denote the original manuscript. Each controlled rewrite setting k, which specifies both a rhetorical condition and a rewrite producer, generates a variant x_{ik} designed to preserve the paper's reported claims, methods, evidence, results, and conclusions while changing how this content is communicated. Such variation may involve claim stance, evidence framing, contribution organization, technical register, or lexical and syntactic realization. A reviewer configuration m specifies both the reviewer model and review protocol. The resulting rewrite and review process is

$$\begin{aligned} x_{ik} &= \mathrm{Rewrite}_k(x_{i0}), \quad k = 1, \ldots, K, \\ y_{mik} &= \mathrm{Review}_m(x_{ik}), \quad k = 0, \ldots, K. \end{aligned} \tag{1}$$

Here, $i = 1, \ldots, N$ indexes papers, x_{ik} is variant k of paper i, and y_{mik} is the judgment assigned by reviewer configuration m to that presentation. A rhetorically robust reviewer should jointly satisfy two complementary requirements. First, **within-paper stability** requires judgments to remain consistent across rhetorical variants of the same paper. This condition alone is insufficient because indiscriminately constant scores would be maximally stable. Second, **between-paper discrimination** requires the reviewer to preserve distinctions based on the reported scientific content of different manuscripts, thereby ruling out this collapse. This requirement does not treat cross-paper score variation as ground-truth scientific merit; it asks whether paper differentiation remains large enough relative to rewrite-induced variation to be meaningful.

Rhetorical robustness therefore constitutes a **joint stability-discrimination requirement**: it requires within-paper stability without sacrificing between-paper discrimination. Section 2.4 operationalizes this joint requirement with two direct within-paper metrics and three joint stability-discrimination metrics. The joint metrics do not measure between-paper behavior in isolation: each relates within-paper consistency to cross-paper variation or separation. Alignment with human reviewer judgments remains a distinct evaluation dimension, which we measure separately.

2.2 Benchmark Construction

RobustReview operationalizes rhetorical robustness through *matched paper families*, each containing an original manuscript and rhetorical variants designed to preserve its reported scientific content.

Comparisons within each family directly measure rewrite-induced instability. Comparisons across families then provide the reference needed to determine whether that within-paper consistency coexists with differentiation among papers. Human judgments on the original manuscripts provide a separate reference for alignment.

We construct RobustReview from 60 anonymized ICLR 2026 submissions with matched arXiv LaTeX sources. We stratify eligible papers by their mean human overall-assessment rating and randomly sample 10 papers from each of six score intervals. This balanced sampling covers papers with different human-assessed ratings, while the human scores provide an external reference for evaluating the reviewers' assessments. We apply 10 rhetorical conditions. Six single-dimension conditions alter novelty stance, scope framing, evidence framing, contribution salience, technical register, or linguistic complexity. Four complex conditions apply a joint rewrite across dimensions, two or three recursive rewrite rounds (R2 and R3), or a reviewer-guided rewrite based on model feedback. Each condition is independently instantiated by GPT-5.5 (OpenAI, 2026) and Claude Opus 4.8 (Anthropic, 2026a), producing two variants per paper and condition. The resulting corpus contains 60 original manuscripts and 1,200 rhetorical variants, or 1,260 full manuscripts in total. All rewrites operate on complete LaTeX projects under content-preservation and structural controls. Automated and human audits indicate that core technical content is largely preserved across the five assessed dimensions (Appendix F.1). Appendices D.1 and D.2 give the sampling procedure, complete intervention definitions, rewrite procedure, and construction checks.

2.3 Reviewer Configurations

We evaluate 30 reviewer configurations across three system families. The general-purpose family comprises GPT-5.5 (OpenAI, 2026), GPT-5-mini (OpenAI, 2025), Claude Sonnet 5 (Anthropic, 2026b), GLM-5.2 (GLM-5-Team et al., 2026), Kimi-K2.6 (Moonshot AI, 2026), GPT-OSS-120B (OpenAI, 2025), Gemini-3.5-Flash-Lite (Google, 2026), and Qwen-3.5-Flash (Qwen Team, 2026). Each is prompted under three protocols: STANDARD, STRICT, and PERSISTENT. STANDARD follows the ICLR review criteria and scoring scheme. STRICT raises the evidentiary threshold through a more conservative rubric. PERSISTENT retains the standard criteria while repeatedly instructing the reviewer to base scientific judgments on substantive content rather than rhetorical presentation. It therefore directly tests whether content-only instructions are sufficient to separate scientific judgment from rhetorical presentation. We additionally evaluate the specialized models OpenReviewer, CycleReviewer, and DeepReviewer (Idahl & Ahmadi, 2025; Weng et al., 2025; Zhu et al., 2025b), and the agentic systems AI Scientist, OpenJudge, and ProReviewer (Lu et al., 2024; The OpenJudge Team, 2025; Fang et al., 2026), using their native review procedures. For every configuration, an original manuscript and all of its rhetorical variants are evaluated with the same procedure and scoring criteria. Appendix D.3 reports the system configurations and execution settings, while Appendix H.4 reproduces all three direct-review prompts.

2.4 Evaluation Metrics

Following the definition above, we evaluate every reviewer configuration with seven metrics. MAD and Drift SD directly measure within-paper stability (lower is better). Because low drift alone can result from score collapse, ICC, SPR, and discriminability jointly evaluate within-paper consistency relative to cross-paper variation or separation (higher is better); we refer to these as *joint stability-discrimination metrics*. Human MAE measures absolute agreement with mean human overall-assessment scores, and Spearman correlation measures agreement with the human ranking of the original papers (lower and higher are better, respectively). On the score-stratified benchmark, these human-alignment metrics complement the robustness metrics by assessing whether the reviewers' scores and rankings agree with human evaluations. Together, the seven metrics characterize stability under rhetorical rewriting, discrimination among papers, and alignment with human judgments. Appendix D.4 gives the complete definitions and equations.

3 Benchmark Findings: Limits of Current AI Reviewers

Table 1 reports all seven metrics for 30 existing reviewer configurations and SciCore under the same matched-manuscript design on RobustReview. The within-paper metrics quantify absolute score

Table 1: **Main comparison of rhetorical robustness and human alignment.** MAD and Drift SD directly measure rewrite-induced within-paper instability. ICC, SPR, and discriminability are joint stability-discrimination metrics: each evaluates within-paper consistency relative to cross-paper variation or separation and should not be interpreted as a between-paper-only measure. Arrows indicate the preferred direction. The best point estimate in each column is shown in bold, the second-best is underlined, and the third-best is italicized.

System	Protocol	Human alignment		Within-paper stability		Joint stability-discrimination		
		H-MAE ↓	Spearman ↑	MAD ↓	Drift SD ↓	ICC ↑	SPR ↑	Discrim. ↑
General-purpose LLM reviewers								
GPT-5.5 (OpenAI, 2026)	Standard	1.294	0.448	0.598	0.956	0.615	*0.555*	0.649
	Strict	<u>1.078</u>	**0.529**	0.766	1.202	*0.632*	0.507	*0.672*
	Persistent	1.261	0.423	0.476	0.875	<u>0.697</u>	<u>0.586</u>	<u>0.682</u>
GPT-5-mini (OpenAI, 2025)	Standard	1.711	0.417	0.543	0.945	0.401	0.390	0.586
	Strict	1.240	0.403	0.895	1.221	0.393	0.424	0.598
	Persistent	1.628	0.335	0.574	0.982	0.476	0.475	0.589
Claude Sonnet 5 (Anthropic, 2026b)	Standard	1.189	0.414	0.468	0.863	0.579	*0.555*	0.646
	Strict	*1.151*	0.475	0.612	1.050	0.546	0.550	0.637
	Persistent	1.239	0.360	<u>0.307</u>	0.710	0.489	0.491	0.615
GLM-5.2 (GLM-5-Team et al., 2026)	Standard	2.044	−0.034	1.500	2.341	0.226	0.322	0.558
	Strict	2.047	−0.038	1.548	2.094	0.212	0.389	0.556
	Persistent	2.403	−0.190	1.932	2.682	0.232	0.381	0.558
Kimi-K2.6 (Moonshot AI, 2026)	Standard	1.689	0.280	1.353	1.995	0.211	0.428	0.555
	Strict	1.737	0.056	0.753	1.261	0.287	0.377	0.544
	Persistent	1.578	0.050	1.139	1.645	0.283	0.392	0.568
GPT-OSS-120B (OpenAI, 2025)	Standard	1.544	0.072	0.717	1.129	0.138	0.384	0.532
	Strict	1.692	0.119	0.385	0.888	0.091	0.353	0.508
	Persistent	1.528	−0.008	0.858	1.317	0.161	0.351	0.536
Gemini 3.5 Flash-Lite (Google, 2026)	Standard	3.178	0.408	**0.205**	<u>0.631</u>	0.199	0.468	0.511
	Strict	1.878	0.465	0.840	1.193	0.424	0.520	0.599
	Persistent	2.828	*0.487*	0.443	0.968	0.347	0.502	0.539
Qwen 3.5 Flash (Qwen Team, 2026)	Standard	2.350	0.379	0.809	1.339	0.291	0.472	0.561
	Strict	1.239	0.409	0.866	1.340	0.226	0.404	0.541
	Persistent	2.037	0.461	0.863	1.281	0.597	0.519	0.597
Specialized review models								
OpenReviewer (Idahl & Ahmadi, 2025)	–	1.406	0.339	1.080	1.524	0.215	0.374	0.542
CycleReviewer (Weng et al., 2025)	–	1.403	0.074	0.779	1.047	0.128	0.350	0.533
DeepReviewer (Zhu et al., 2025b)	–	1.471	0.406	0.454	0.689	0.161	0.381	0.537
Agentic review systems								
AI Scientist (Lu et al., 2024)	–	1.513	0.094	0.838	1.253	0.213	0.382	0.544
OpenJudge (The OpenJudge Team, 2025)	–	1.306	0.120	*0.335*	**0.597**	0.239	0.444	0.526
ProReviewer (Fang et al., 2026)	–	1.478	0.230	0.709	0.978	0.094	0.342	0.516
Our method								
SciCore	–	**1.072**	<u>0.488</u>	0.476	*0.687*	**0.775**	**0.652**	**0.726**

movement, the joint metrics test whether stability coexists with paper-level discrimination, and the human-alignment metrics provide a separate external comparison.

3.1 Content-Only Review Is Not Reliably Achieved by Prompting Alone

We evaluate PERSISTENT, a full-manuscript review protocol that repeatedly emphasizes scientific content and provides explicit evidence-based scoring guidance. Relative to STANDARD, its effects vary across backbones: only GPT-5.5 improves on all five robustness metrics. For Claude Sonnet 5, PERSISTENT reduces MAD and Drift SD but lowers ICC, SPR, and discriminability, while the remaining backbones also show mixed changes. These results show that the evaluated content-focused prompting protocol does not consistently improve rhetorical robustness across backbones, motivating our investigation of an explicit content-normalized branch.

3.2 False Robustness: Within-Paper Stability without Discrimination

Gemini-3.5-Flash-Lite under STANDARD achieves the lowest MAD among the existing configurations, at 0.205, but its ICC is only 0.199 and its discriminability is 0.511, close to chance. The same pattern appears among specialized and agentic reviewers: DeepReviewer and OpenJudge show relatively low drift, yet attain ICC values of only 0.161 and 0.239 and discriminability values of 0.537

and 0.526. By contrast, GPT-5.5 under PERSISTENT has a higher MAD of 0.476 but the strongest ICC, SPR, and discriminability among the 30 configurations. We call low rewrite-induced drift accompanied by weak paper differentiation *false robustness*: invariance alone can create the appearance of robustness without the discrimination that robustness is meant to preserve.

3.3 HUMAN ALIGNMENT AND RHETORICAL ROBUSTNESS ARE DISTINCT

Within GPT-5.5, STRICT provides the strongest human-alignment profile among the existing configurations, with a Human MAE of 1.078 and a Spearman correlation of 0.529, whereas PERSISTENT has weaker human alignment but higher ICC, SPR, and discriminability. Human alignment asks whether a reviewer reproduces human judgments on observed manuscripts; rhetorical robustness asks whether judgments remain stable across matched rhetorical variants while preserving differences among papers. The two evaluations therefore favor different protocols, so human agreement cannot substitute for matched robustness evaluation.

4 SCICORE: DUAL-BRANCH REVIEW FOR RHETORICAL ROBUSTNESS

The central idea of SciCore is to augment manuscript-based review with a scientific judgment that is less coupled to rhetorical presentation. Unlike PERSISTENT, which asks a reviewer to disregard rhetoric while still reading the complete manuscript, the science-core branch first transforms the input into a content-normalized scientific record. SciCore uses two complementary branches. The *manuscript branch* reviews the complete paper under the STRICT protocol, retaining the full manuscript context. The *science-core branch* extracts a structured *science core* containing the reported scientific record and reviews that record with an adapted protocol. The final overall assessment is the arithmetic mean of the two branch scores. This design preserves a conventional judgment of the paper while giving equal weight to a content-normalized judgment intended to vary less with rhetorical framing, organization, and linguistic expression.

4.1 DESIGN PRINCIPLE: A STABLE SCIENTIFIC BRANCH WITHOUT DISCARDING THE MANUSCRIPT

The science-core branch is designed to provide a scientific judgment that is less sensitive to rhetorical realization while retaining distinctions among papers. Conceptually, it maps different presentations of the same reported science to similar structured records without collapsing scientifically distinct manuscripts. This invariant-representation view motivates extracting and reviewing a science core, but the implemented extractor is only an approximation: its stability and paper separation must be established empirically rather than assumed.

The science-core branch complements rather than replaces manuscript review. When the science-core branch varies less across matched rhetorical realizations than the manuscript branch, averaging their scores can reduce rhetorical sensitivity while retaining the full paper context. This intuition addresses direct within-paper stability only; the final reviewer must still be evaluated using the joint stability-discrimination metrics to ensure that lower score drift does not come from cross-paper collapse.

We review the extracted record directly instead of first reconstructing another manuscript from it. Reconstruction cannot create decision-relevant information absent from its inputs and may introduce another rhetorical realization, but this information-theoretic argument does not guarantee better performance for a fixed LLM or order our implemented pipelines. We therefore treat direct review as a design choice and compare it empirically with ReconstructReview. Appendix A gives the formal invariant-representation, conditional-stability, fusion, and Blackwell arguments.

4.2 SCIENCE-CORE EXTRACTION: ISOLATING THE REPORTED SCIENTIFIC RECORD

The science-core branch first uses an LLM to extract a structured science core from the complete manuscript. The extraction focuses on information directly relevant to scientific evaluation, including the central research idea and claims, problem formulation, mathematical formulations and derivations, methods and assumptions, experimental or theoretical evidence, reported results, author-stated contributions, reproducibility information, and stated limitations.

The LLM is instructed to extract this information in objective, third-person language while preserving its attribution to the original manuscript. This is particularly important for scientific claims and author-stated contributions, where rhetorical framing could otherwise carry into the science core. The extractor is instructed to attribute claims and contributions to the authors, preserve reported results and numerical evidence, and avoid subjective assessments, evaluative language, emotional framing, and unsupported interpretations. It does not determine whether a claim is convincing, whether a contribution is significant, or how the paper should be scored. The complete extraction prompt and output structure are provided in Appendix H.

The science core is produced as a text-based scientific record rather than a shortened rewrite of the manuscript. Figures are not retained directly. Tables are instead explicitly transcribed so that their reported values, comparisons, and other scientific information remain available to the downstream reviewer. The prompt asks the extractor to retain equations, mathematical arguments, reported numbers, and other details whenever they can be reliably recovered. When information cannot be located or read reliably, the extractor records this uncertainty rather than inferring or reconstructing the missing content.

4.3 SCIENCE-CORE BRANCH: EVALUATING THE EXTRACTED RECORD

After extraction, the science-core branch reviews only the extracted record. The original manuscript is not provided to this branch at the review stage, so its judgment is based on the retained scientific record rather than its original rhetorical presentation.

We build the review protocol on the STANDARD protocol used in RobustReview, while adapting its instructions to the structure of the science core. The prompt explains the role of each extracted section and how information across sections should be combined when forming a scientific judgment. In particular, the reviewer is instructed to connect the stated research problem and claims with the corresponding methods, assumptions, mathematical reasoning, and empirical or theoretical evidence; to assess whether the reported results support the author-stated claims and contributions; and to use the reproducibility information and stated limitations when evaluating the strength and scope of the evidence. Information that is explicitly absent from the manuscript can be treated as missing scientific support when relevant, while extraction uncertainty is handled separately. The complete review prompt and criterion mapping are provided in Appendix H.

The review retains the ICLR-style evaluation criteria and scoring scheme used by STANDARD. The reviewer produces written feedback, including a summary of the work, strengths, weaknesses, and questions, and an overall assessment, together with supporting scientific-evaluation scores. The score anchors follow the corresponding ICLR-style scales, with the review decision grounded in the scientific evidence contained in the science core.

4.4 DUAL-BRANCH FUSION: INTEGRATING MANUSCRIPT AND SCIENCE-CORE JUDGMENTS

In parallel with the science-core branch, the manuscript branch applies the STRICT protocol from Section 2.3 directly to the complete manuscript PDF. This branch retains the ordinary manuscript-level review context, while its evidence-focused rubric provides the conventional judgment used in the fusion. Both branches produce an overall assessment on the same ICLR-style scale.

Let $M(x)$ denote the manuscript-branch overall-assessment score and let $B(x)$ denote the science-core-branch score for manuscript realization x. The final SciCore score is their unweighted arithmetic mean:

$$S_{\text{SciCore}}(x) = \frac{1}{2}M(x) + \frac{1}{2}B(x). \tag{2}$$

Equal weighting provides a direct symmetric combination without introducing a tuned fusion parameter. The manuscript branch and science-core branch remain independently interpretable, which allows the evaluation to distinguish the behavior of each component from that of the fused reviewer.

Table 2: **Within-backbone analysis of SciCore and its components.** Every configuration uses GPT-5.5. MAD and Drift SD are direct within-paper metrics, whereas ICC, SPR, and discriminability jointly relate within-paper consistency to cross-paper variation or separation. The final SciCore row averages the Manuscript-Strict and Core-Adapted scores. The best point estimate in each column is shown in bold and the second-best is underlined.

Configuration	Human alignment		Within-paper stability		Joint stability-discrimination		
	H-MAE ↓	Spearman ↑	MAD ↓	Drift SD ↓	ICC ↑	SPR ↑	Discrim. ↑
Manuscript-Based Review							
Manuscript-Standard	1.294	0.448	0.598	0.956	0.615	0.555	0.649
Manuscript-Strict	<u>1.078</u>	**0.529**	0.766	1.202	0.632	0.507	0.672
ReconstructReview	1.490	0.133	0.651	1.101	0.623	0.566	0.664
Science-Core Branch							
Core-Standard	1.394	0.216	<u>0.354</u>	0.737	0.734	<u>0.677</u>	<u>0.697</u>
Core-Strict	1.350	0.264	0.403	0.870	0.608	0.584	0.654
Core-Persistent	1.346	0.289	0.458	0.821	0.708	0.586	0.681
Core-Adapted	1.461	0.285	**0.292**	**0.650**	<u>0.751</u>	**0.710**	0.695
Dual-Branch Fusion							
SciCore (ours)	**1.072**	<u>0.488</u>	0.476	<u>0.687</u>	**0.775**	0.652	**0.726**

5 SciCore: Results and Analysis

5.1 Main Results: Rhetorical Robustness on RobustReview

As shown in Table 1, SciCore achieves a leading joint stability-discrimination profile in the primary comparison while maintaining competitive human alignment. Paired paper-family bootstrap analysis supports its improvements over Manuscript-Strict across all five robustness metrics, alongside improved human alignment relative to Core-Adapted (Appendix F.4). These results reinforce the complementary roles of the two branches and demonstrate the effectiveness of a simple equal-weight fusion.

5.2 Dissecting the Gains: Architecture and Review-Policy Ablations

Table 2 isolates the design choices using a common GPT-5.5 backbone. Manuscript-Standard and Manuscript-Strict review the PDF directly. ReconstructReview reconstructs a paper from the science core, permitted figures, and bibliography before applying STANDARD review (Appendix D.6). Four core-only configurations share the same cached science core and differ only in review policy: Standard, Strict, Persistent, or the adapted protocol. The final SciCore row averages Core-Adapted with Manuscript-Strict.

Directly reviewing the science core provides a stronger robustness branch than reconstruction in this pipeline. Relative to Manuscript-Standard, ReconstructReview worsens MAD from 0.598 to 0.651 and Drift SD from 0.956 to 1.101, with only modest gains on the joint metrics. Core-Standard then improves all seven metrics over ReconstructReview. The result shows that reconstruction does not recover the robustness of direct science-core review here. Proposition A.2 provides an information-theoretic rationale for direct review but does not predict the ordering of the implemented pipelines.

Content-focused review remains nontrivial after extraction. Core-Persistent worsens all five robustness metrics relative to Core-Standard, whereas Core-Adapted achieves the lowest MAD and Drift SD and the highest ICC and SPR among the core-only configurations. The review policy thus affects the balance between score stability and paper discrimination even when the extracted representation is held fixed.

The two branches provide complementary judgments. Core-Adapted is more stable than Manuscript-Strict but aligns less closely with human scores. Fusion improves all five robustness metrics over Manuscript-Strict and improves human alignment, ICC, and discriminability over Core-

Adapted, which retains lower MAD and Drift SD and higher SPR. Paired bootstrap intervals support these directions (Appendix F.4).

With Manuscript-Strict and equal weighting fixed, the science-core branch also improves all five robustness metrics over adding Standard or Persistent manuscript review, supported by paired bootstrap intervals. These controls share its two-score averaging rule and half-point resolution. Appendix F.2 discusses the mechanical effects of averaging and a three-review control matching SciCore's nominal call count. That control has lower Drift SD and higher ICC, while SciCore has better point estimates on the other five metrics.

5.3 Additional Analyses

To further validate rhetorical robustness and examine the design choices underlying SciCore, we conduct four complementary analyses. We compare science-core embeddings within and across paper families to assess representational stability and paper separation, finding consistently higher similarity within families (Appendix C.1). We examine individual review criteria to detect false robustness from score collapse; Core-Adapted restores confidence-score variation that is absent or nearly absent under the other core-review protocols (Appendix C.2). We evaluate the science-core branch across three backbones to test transfer, finding improved ICC for all three with broader robustness gains varying by backbone (Appendix C.3). We vary fusion weights and manuscript protocols to assess the tradeoff between rhetorical robustness and human alignment; our equal-weight default is Pareto non-dominated among the evaluated configurations on the reported point estimates (Appendix C.4).

6 Related Work

LLM-as-a-judge and AI peer-review research evaluates human agreement, review quality, score prediction, and workflow capability, while documenting sensitivity to order, length, prompts, and evaluator identity (Liu et al., 2023; Zheng et al., 2023; Wang et al., 2024; Dubois et al., 2024; Liang et al., 2023; Zhou et al., 2024). Studies of automated review further reveal vulnerabilities to hidden instructions, artificial perturbations, and visible content-preserving revisions (Ye et al., 2024; Lin et al., 2025; Kaneko, 2026; Li et al., 2026b; Baumann et al., 2026). Li et al. (2026c) characterize rhetorical reward hacking through controlled full-manuscript rewrites. Our work formulates rhetorical robustness as a joint requirement of within-paper stability and cross-paper discrimination, and introduces SciCore to improve this balance while maintaining competitive human alignment.

Existing interventions control particular biases or judge intermediate representations (Dubois et al., 2024; Li et al., 2026d). Rather than relying only on instructions to discount rhetoric, SciCore augments full-manuscript review with a science-core branch that extracts and evaluates structured scientific content, then averages the two judgments. This retains manuscript-level assessment while adding a content-normalized view, motivated by invariant representation and Blackwell comparison (Eaton, 1989; Dubois et al., 2021; Blackwell, 1953). Appendix B provides the extended discussion.

7 Conclusion

We formulate rhetorical robustness as stability under rhetorical variation with discrimination across papers. RobustReview exposes false robustness from score collapse and shows that neither human alignment nor instructions to focus on scientific content alone establish rhetorical robustness. SciCore combines manuscript-level and science-core judgments through simple equal-weight fusion, achieving a leading joint stability-discrimination profile in our primary GPT-5.5 comparison while maintaining competitive human alignment. These findings establish rhetorical robustness as an evaluation and design objective for trustworthy AI review.

AI USE STATEMENT

The authors formulated the research questions and led the study design. Generative AI assisted with language editing, manuscript organization, and reviewing the clarity and consistency of methodological explanations and result interpretations. LLMs also served as components of the experimental pipeline, generating controlled manuscript rewrites, extracting science cores, reconstructing manuscripts, producing reviews, and conducting automated technical-content fidelity audits, as described in the paper. The authors made the final scientific decisions and take responsibility for the manuscript, its claims, and its reported results.

ETHICAL CONSIDERATIONS

This study uses publicly available ICLR review metadata and matched public arXiv sources, with anonymized manuscripts as benchmark inputs. Our analysis concerns aggregate reviewer behavior and rhetorical robustness rather than judgments about individual authors. Human validation assesses preservation of core technical content between original and rewritten manuscripts. Controlled rewriting has dual-use implications: it can reveal vulnerabilities in AI review, but it can also be used to optimize review scores without improving the reported science. Score increases should therefore not be interpreted as evidence of greater scientific merit, and the rewriting procedures should not be used to manipulate active peer-review processes. Such use could reward persuasive presentation over substantive advances and undermine scientific evaluation. SciCore is intended to support research on more reliable AI review; its scores should not replace expert scientific judgment.

REPRODUCIBILITY STATEMENT

The appendices document the construction and evaluation of RobustReview and SciCore. Appendix D describes benchmark sampling, rewriting procedures, reviewer execution, metric definitions, and the SciCore and ReconstructReview protocols. Appendix C presents ablations and representation analyses, while Appendix F reports fidelity audits, manuscript-ensemble controls, repeated-review noise baselines, and paper-family bootstrap analyses. Appendix H provides the science-core extraction and review prompts and all three direct-review prompts. Compute resources and API expenditure are reported in Appendix G.

REFERENCES

Anthropic. Introducing Claude Opus 4.8. https://www.anthropic.com/news/claude-opus-4-8, 2026a.

Anthropic. Introducing Claude Sonnet 5. https://www.anthropic.com/news/claude-sonnet-5, 2026b.

Joachim Baumann, Jiaxin Pei, Sanmi Koyejo, and Dirk Hovy. Stop automating peer review without rigorous evaluation. *arXiv preprint arXiv:2605.03202*, 2026.

Anna Bavaresco, Raffaella Bernardi, Leonardo Bertolazzi, Desmond Elliott, Raquel Fernández, Albert Gatt, Esam Ghaleb, Mario Giulianelli, Michael Hanna, Alexander Koller, Andre Martins, Philipp Mondorf, Vera Neplenbroek, Sandro Pezzelle, Barbara Plank, David Schlangen, Alessandro Suglia, Aditya K Surikuchi, Ece Takmaz, and Alberto Testoni. LLMs instead of human judges? a large scale empirical study across 20 NLP evaluation tasks. In Wanxiang Che, Joyce Nabende, Ekaterina Shutova, and Mohammad Taher Pilehvar (eds.), *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pp. 238–255, Vienna, Austria, July 2025. Association for Computational Linguistics. ISBN 979-8-89176-252-7. doi: 10.18653/v1/2025.acl-short.20. URL https://aclanthology.org/2025.acl-short.20/.

Savita Bhat and Vasudeva Varma. All prompts are created equal? evaluating robustness of llm judges against non-adversarial prompt variations. In *Findings of the Association for Computational Linguistics: ACL 2026*, pp. 38730–38745, 2026.

David Blackwell. Equivalent comparisons of experiments. *The Annals of Mathematical Statistics*, 24(2):265–272, 1953. doi: 10.1214/aoms/1177729032.

Zeyuan Chen, Ziqing Yang, Yihan Ma, Michael Backes, and Yang Zhang. Peercheck: Enhancing llm-generated academic reviews towards human-level quality. In *Findings of the Association for Computational Linguistics: ACL 2026*, pp. 23362–23386, 2026.

Matteo Gioele Collu, Umberto Salviati, Roberto Confalonieri, Mauro Conti, and Giovanni Apruzzese. Misleading large language models used (or misused) in scientific peer-reviewing via hidden prompt-injection attacks. *ACM Transactions on AI Security and Privacy*, 2026.

Shurui Du. TitleTrap: Probing presentation bias in LLM-based scientific reviewing. In Mousumi Akter, Tahiya Chowdhury, Steffen Eger, Christoph Leiter, Juri Opitz, and Erion Çano (eds.), *Proceedings of the 5th Workshop on Evaluation and Comparison of NLP Systems*, pp. 119–125, Mumbai, India, December 2025. Association for Computational Linguistics. ISBN 979-8-89176-305-0. doi: 10.18653/v1/2025.eval4nlp-1.10. URL https://aclanthology.org/2025.eval4nlp-1.10/.

Yann Dubois, Benjamin Bloem-Reddy, Karen Ullrich, and Chris J. Maddison. Lossy compression for lossless prediction. In *Advances in Neural Information Processing Systems*, volume 34, 2021. URL https://proceedings.neurips.cc/paper/2021/hash/7535bbb91c8fde347ad861f293126633-Abstract.html.

Yann Dubois, Balázs Galambosi, Percy Liang, and Tatsunori B Hashimoto. Length-controlled alpacaeval: A simple way to debias automatic evaluators. *arXiv preprint arXiv:2404.04475*, 2024.

Nils Dycke and Iryna Gurevych. Automatic reviewers fail to detect faulty reasoning in research papers: A new counterfactual evaluation framework. *Transactions of the Association for Computational Linguistics*, 14:465–488, 2026.

Morris L. Eaton. *Group Invariance in Applications in Statistics*, volume 1 of *NSF-CBMS Regional Conference Series in Probability and Statistics*. Institute of Mathematical Statistics and American Statistical Association, 1989. doi: 10.1214/cbms/1462061029.

Haishuo Fang, Yue Feng, and Iryna Gurevych. From passive generation to investigation: A proactive scientific peer review agent, 2026. URL https://arxiv.org/abs/2606.13349.

Panagiotis Fytas, Georgios Rizos, and Lucia Specia. What makes a scientific paper be accepted for publication? In *Proceedings of the first workshop on causal inference and NLP*, pp. 44–60, 2021.

GLM-5-Team, :, Aohan Zeng, Xin Lv, Zhenyu Hou, Zhengxiao Du, Qinkai Zheng, Bin Chen, Da Yin, Chendi Ge, Chenghua Huang, Chengxing Xie, Chenzheng Zhu, Congfeng Yin, Cunxiang Wang, Gengzheng Pan, Hao Zeng, Haoke Zhang, Haoran Wang, Huilong Chen, Jiajie Zhang, Jian Jiao, Jiaqi Guo, Jingsen Wang, Jingzhao Du, Jinzhu Wu, Kedong Wang, Lei Li, Lin Fan, Lucen Zhong, Mingdao Liu, Mingming Zhao, Pengfan Du, Qian Dong, Rui Lu, Shuang-Li, Shulin Cao, Song Liu, Ting Jiang, Xiaodong Chen, Xiaohan Zhang, Xuancheng Huang, Xuezhen Dong, Yabo Xu, Yao Wei, Yifan An, Yilin Niu, Yitong Zhu, Yuanhao Wen, Yukuo Cen, Yushi Bai, Zhongpei Qiao, Zihan Wang, Zikang Wang, Zilin Zhu, Ziqiang Liu, Zixuan Li, Bojie Wang, Bosi Wen, Can Huang, Changpeng Cai, Chao Yu, Chen Li, Chengwei Hu, Chenhui Zhang, Dan Zhang, Daoyan Lin, Dayong Yang, Di Wang, Ding Ai, Erle Zhu, Fangzhou Yi, Feiyu Chen, Guohong Wen, Hailong Sun, Haisha Zhao, Haiyi Hu, Hanchen Zhang, Hanrui Liu, Hanyu Zhang, Hao Peng, Hao Tai, Haobo Zhang, He Liu, Hongwei Wang, Hongxi Yan, Hongyu Ge, Huan Liu, Huanpeng Chu, Jia'ni Zhao, Jiachen Wang, Jiajing Zhao, Jiamin Ren, Jiapeng Wang, Jiaxin Zhang, Jiayi Gui, Jiayue Zhao, Jijie Li, Jing An, Jing Li, Jingwei Yuan, Jinhua Du, Jinxin Liu, Junkai Zhi, Junwen Duan, Kaiyue Zhou, Kangjian Wei, Ke Wang, Keyun Luo, Laiqiang Zhang, Leigang Sha, Liang Xu, Lindong Wu, Lintao Ding, Lu Chen, Minghao Li, Nianyi Lin, Pan Ta, Qiang Zou, Rongjun Song, Ruiqi Yang, Shangqing Tu, Shangtong Yang, Shaoxiang Wu, Shengyan Zhang, Shijie Li, Shuang Li, Shuyi Fan, Wei Qin, Wei Tian, Weining Zhang, Wenbo Yu, Wenjie Liang, Xiang Kuang, Xiangmeng Cheng, Xiangyang Li, Xiaoquan Yan, Xiaowei Hu, Xiaoying Ling, Xing Fan, Xingye Xia, Xinyuan Zhang, Xinze Zhang, Xirui Pan, Xu Zou, Xunkai Zhang, Yadi Liu, Yandong Wu, Yanfu Li, Yidong Wang, Yifan Zhu, Yijun Tan, Yilin Zhou, Yiming Pan, Ying Zhang, Yinpei Su, Yipeng Geng, Yong Yan, Yonglin Tan, Yuean Bi, Yuhan Shen, Yuhao Yang, Yujiang

Li, Yunan Liu, Yunqing Wang, Yuntao Li, Yurong Wu, Yutao Zhang, Yuxi Duan, Yuxuan Zhang, Zezhen Liu, Zhengtao Jiang, Zhenhe Yan, Zheyu Zhang, Zhixiang Wei, Zhuo Chen, Zhuoer Feng, Zijun Yao, Ziwei Chai, Ziyuan Wang, Zuzhou Zhang, Bin Xu, Minlie Huang, Hongning Wang, Juanzi Li, Yuxiao Dong, and Jie Tang. Glm-5: from vibe coding to agentic engineering, 2026. URL https://arxiv.org/abs/2602.15763.

Google. Introducing Gemini 3.6 Flash, 3.5 Flash-Lite, and 3.5 Flash Cyber. https://blog.google/innovation-and-ai/models-and-research/gemini-models/gemini-3-6-flash-3-5-flash-lite-3-5-flash-cyber/, 2026.

Maximilian Idahl and Zahra Ahmadi. OpenReviewer: A specialized large language model for generating critical scientific paper reviews. In Nouha Dziri, Sean (Xiang) Ren, and Shizhe Diao (eds.), *Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (System Demonstrations)*, pp. 550–562, Albuquerque, New Mexico, April 2025. Association for Computational Linguistics. ISBN 979-8-89176-191-9. doi: 10.18653/v1/2025.naacl-demo.44. URL https://aclanthology.org/2025.naacl-demo.44/.

Joseph James, Chenghao Xiao, Yucheng Li, and Chenghua Lin. On the rigour of scientific writing: Criteria, analysis, and insights. In Yaser Al-Onaizan, Mohit Bansal, and Yun-Nung Chen (eds.), *Findings of the Association for Computational Linguistics: EMNLP 2024*, pp. 6523–6538, Miami, Florida, USA, November 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.findings-emnlp.380. URL https://aclanthology.org/2024.findings-emnlp.380/.

Masahiro Kaneko. Paraphrasing adversarial attack on llm-as-a-reviewer. *arXiv preprint arXiv:2601.06884*, 2026.

Dongyeop Kang, Waleed Ammar, Bhavana Dalvi, Madeleine Van Zuylen, Sebastian Kohlmeier, Eduard Hovy, and Roy Schwartz. A dataset of peer reviews (peerread): Collection, insights and nlp applications. In *Proceedings of the 2018 conference of the North American chapter of the Association for Computational Linguistics: Human language technologies, volume 1 (long papers)*, pp. 1647–1661, 2018.

Seungone Kim, Jamin Shin, Yejin Cho, Joel Jang, Shayne Longpre, Hwaran Lee, Sangdoo Yun, Seongjin Shin, Sungdong Kim, James Thorne, and Minjoon Seo. Prometheus: Inducing fine-grained evaluation capability in language models. In *The Twelfth International Conference on Learning Representations*, 2024. URL https://openreview.net/forum?id=8euJaTveKw.

Dongryeol Lee, Yerin Hwang, Yongil Kim, Joonsuk Park, and Kyomin Jung. Are llm-judges robust to expressions of uncertainty? investigating the effect of epistemic markers on llm-based evaluation. In *Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pp. 8962–8984, 2025.

Haowen Li, Yoichi Ishibashi, and Masafumi Oyamada. Evaluating the impact of reviewer guideline design on llm-based automated peer review. In *Findings of the Association for Computational Linguistics: ACL 2026*, pp. 30223–30240, 2026a.

Lin Li, Qi Zhang, Xander Davies, Jianing Qiu, and Yarin Gal. Gaming ai-assisted peer reviews poses new risks to the scientific community. *arXiv preprint arXiv:2606.10159*, 2026b.

Ming Li, Chenguang Wang, Xirui Li, Xinyue Zeng, Dianqi Li, Peng Shi, Dawei Zhou, and Tianyi Zhou. How can rhetoric reward-hack ai reviewers? dissecting rhetorical sensitivity in ai-based peer review. *arXiv preprint arXiv:2608.08975*, 2026c.

Ruochi Li, Haoxuan Zhang, Edward Gehringer, Ting Xiao, Junhua Ding, and Haihua Chen. Unveiling the merits and defects of llms in automatic review generation for scientific papers. In *2025 IEEE International Conference on Data Mining (ICDM)*, pp. 1370–1379. IEEE, 2025a.

Yifei Li, Xiaoting Xu, Dongqing Lyu, Zhen Zhang, Juan Xie, and Ying Cheng. Developing a criteria framework for peer review: a critical interpretive synthesis. *Learned Publishing*, 38(3): e2016, 2025b.

Zhuochun Li, Yong Zhang, Ming Li, Yuelyu Ji, Yiming Zeng, Ning Cheng, Yun Zhu, Yanmeng Wang, Shaojun Wang, Jing Xiao, and Daqing He. Rethinking LLM-as-a-judge: Representation-as-a-judge with small language models via semantic capacity asymmetry. In *The Fourteenth International Conference on Learning Representations*, 2026d. URL https://openreview.net/forum?id=VAISvCsrvG.

Weixin Liang, Yuhui Zhang, Hancheng Cao, Binglu Wang, Daisy Ding, Xinyu Yang, Kailas Vodrahalli, Siyu He, Daniel Smith, Yian Yin, Daniel McFarland, and James Zou. Can large language models provide useful feedback on research papers? a large-scale empirical analysis, 2023. URL https://arxiv.org/abs/2310.01783.

Tzu-Ling Lin, Wei-Chih Chen, Teng-Fang Hsiao, Hou-I Liu, Ya-Hsin Yeh, Yu-Kai Chan, Wen-Sheng Lien, Po-Yen Kuo, Philip S. Yu, and Hong-Han Shuai. Breaking the reviewer: Assessing the vulnerability of large language models in automated peer review under textual adversarial attacks. In Christos Christodoulopoulos, Tanmoy Chakraborty, Carolyn Rose, and Violet Peng (eds.), *Findings of the Association for Computational Linguistics: EMNLP 2025*, pp. 4819–4839, Suzhou, China, November 2025. Association for Computational Linguistics. ISBN 979-8-89176-335-7. doi: 10.18653/v1/2025.findings-emnlp.259. URL https://aclanthology.org/2025.findings-emnlp.259/.

Yang Liu, Dan Iter, Yichong Xu, Shuohang Wang, Ruochen Xu, and Chenguang Zhu. G-eval: Nlg evaluation using gpt-4 with better human alignment. In *Proceedings of the 2023 conference on empirical methods in natural language processing*, pp. 2511–2522, 2023.

Chris Lu, Cong Lu, Robert Tjarko Lange, Jakob Foerster, Jeff Clune, and David Ha. The ai scientist: Towards fully automated open-ended scientific discovery, 2024. URL https://arxiv.org/abs/2408.06292.

Moonshot AI. Kimi K2.6: Advancing Open-Source Coding. https://www.kimi.ai/blog/kimi-k2-6/, 2026.

OpenAI. gpt-oss-120b & gpt-oss-20b model card, 2025. URL https://arxiv.org/abs/2508.10925.

OpenAI. GPT-5 mini. https://developers.openai.com/api/docs/models/gpt-5-mini, 2025.

OpenAI. GPT-5.5. https://openai.com/index/introducing-gpt-5-5/, 2026.

Arjun Panickssery, Samuel R Bowman, and Shi Feng. Llm evaluators recognize and favor their own generations. *Advances in Neural Information Processing Systems*, 37:68772–68802, 2024.

Qwen Team. Qwen3.5: Towards native multimodal agents, February 2026. URL https://qwen.ai/blog?id=qwen3.5.

Rickard Stureborg, Dimitris Alikaniotis, and Yoshi Suhara. Large language models are inconsistent and biased evaluators. *arXiv preprint arXiv:2405.01724*, 2024.

Nitya Thakkar, Mert Yuksekgonul, Jake Silberg, Animesh Garg, Nanyun Peng, Fei Sha, Rose Yu, Carl Vondrick, and James Zou. A large-scale randomized study of large language model feedback in peer review. *Nature Machine Intelligence*, 8(3):326–336, 2026.

The OpenJudge Team. Openjudge: A unified framework for holistic evaluation and quality rewards, 07 2025. URL https://github.com/agentscope-ai/OpenJudge.

Sai Suresh Macharla Vasu, Ivaxi Sheth, Hui-Po Wang, Ruta Binkyte, and Mario Fritz. Justice in judgment: Unveiling (hidden) bias in llm-assisted peer reviews. In *Findings of the Association for Computational Linguistics: ACL 2026*, pp. 307–330, 2026.

Peiyi Wang, Lei Li, Liang Chen, Zefan Cai, Dawei Zhu, Binghuai Lin, Yunbo Cao, Lingpeng Kong, Qi Liu, Tianyu Liu, et al. Large language models are not fair evaluators. In *Proceedings of the 62nd annual meeting of the association for computational linguistics (volume 1: Long papers)*, pp. 9440–9450, 2024.

Qingyun Wang, Qi Zeng, Lifu Huang, Kevin Knight, Heng Ji, and Nazneen Fatema Rajani. Reviewrobot: Explainable paper review generation based on knowledge synthesis. In *Proceedings of the 13th International Conference on Natural Language Generation*, pp. 384–397, 2020.

Yixuan Weng, Minjun Zhu, Guangsheng Bao, Hongbo Zhang, Jindong Wang, Yue Zhang, and Linyi Yang. Cycleresearcher: Improving automated research via automated review, 2025. URL https://arxiv.org/abs/2411.00816.

Xianglin Yang, Bryan Hooi, Gelei Deng, Tianwei Zhang, and Jin Song Dong. Turning bias into bugs: Bandit-guided style manipulation attacks on llm judges. *arXiv preprint arXiv:2605.26156*, 2026a.

Xu Yang, Zhizhou Sha, Junbo Li, Jian Yu, Yifan Sun, Matthew Zhao, Jinrui Fang, Xinyue Guo, Yining Wu, Xu Hu, Yifu Luo, Qiang Liu, and Zhangyang Wang. No hidden prompts needed! you can game ai peer review with presentation-only revisions, 2026b. URL https://arxiv.org/abs/2606.13044.

Rui Ye, Xianghe Pang, Jingyi Chai, Jiaao Chen, Zhenfei Yin, Zhen Xiang, Xiaowen Dong, Jing Shao, and Siheng Chen. Are we there yet? revealing the risks of utilizing large language models in scholarly peer review, 2024. URL https://arxiv.org/abs/2412.01708.

Weizhe Yuan, Pengfei Liu, and Graham Neubig. Can we automate scientific reviewing? *Journal of Artificial Intelligence Research*, 75:171–212, 2022.

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

Ruiyang Zhou, Lu Chen, and Kai Yu. Is llm a reliable reviewer? a comprehensive evaluation of llm on automatic paper reviewing tasks. In *Proceedings of the 2024 joint international conference on computational linguistics, language resources and evaluation (LREC-COLING 2024)*, pp. 9340–9351, 2024.

Changjia Zhu, Junjie Xiong, Renkai Ma, Zhicong Lu, Yao Liu, and Lingyao Li. When your reviewer is an llm: Biases, divergence, and prompt injection risks in peer review. *arXiv preprint arXiv:2509.09912*, 2025a.

Minjun Zhu, Yixuan Weng, Linyi Yang, and Yue Zhang. DeepReview: Improving LLM-based paper review with human-like deep thinking process. In Wanxiang Che, Joyce Nabende, Ekaterina Shutova, and Mohammad Taher Pilehvar (eds.), *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 29330–29355, Vienna, Austria, July 2025b. Association for Computational Linguistics. ISBN 979-8-89176-251-0. doi: 10.18653/v1/2025.acl-long.1420. URL https://aclanthology.org/2025.acl-long.1420/.

A THEORETICAL MOTIVATION FOR SCICORE

This section formalizes the representation and reconstruction arguments motivating the science-core branch. These results provide design rationales rather than performance guarantees for a fixed LLM reviewer. Throughout, $x \in \mathcal{X}$ denotes a generic manuscript realization. When we specialize the discussion to RobustReview, x_{i0} and x_{ik} retain their definitions from Section 2, and we suppress the reviewer-configuration index when a branch is fixed.

A.1 THE MAXIMAL-INVARIANT IDEAL

We write $x \sim_{\mathrm{sci}} x'$ when two realizations express the same reported scientific record, including claims, methods, assumptions, evidence, results, and conclusions, while differing rhetorically. This relation defines the ideal invariance target. Our generated rewrites are designed to approximate it, and their preservation is assessed empirically rather than assumed to be exact.

An ideal representation $C^\star : \mathcal{X} \to \mathcal{Z}$ is a maximal invariant to rhetorical realization when

$$C^\star(x) = C^\star(x') \quad \Longleftrightarrow \quad x \sim_{\mathrm{sci}} x'. \tag{3}$$

The right-to-left implication expresses rhetorical invariance; the left-to-right implication prevents different scientific records from being collapsed into a single representation. We use maximal invariant with respect to the equivalence relation $\sim_{\mathrm{sci}}$. This equivalence-class view follows the standard maximal-invariant principle and its use for invariant downstream prediction (Eaton, 1989; Dubois et al., 2021).

Proposition A.1 (Factorization through an ideal science core)**.** *If $C^\star$ satisfies Equation 3, every rhetorically invariant judgment $F : \mathcal{X} \to \mathcal{A}$ factors through $C^\star$: there exists a unique map $R : C^\star(\mathcal{X}) \to \mathcal{A}$ such that $F = R \circ C^\star$. Conversely, every judgment of the form $R \circ C^\star$ is rhetorically invariant.*

Proof. For $z = C^\star(x)$, define $R(z) = F(x)$. If $C^\star(x) = C^\star(x')$, maximality gives $x \sim_{\mathrm{sci}} x'$, and invariance of F gives $F(x) = F(x')$; therefore R is well defined. The converse follows from the invariance of $C^\star$. □

Proposition A.1 motivates the science-core branch: an invariant judgment can, in principle, operate on a representation of the scientific equivalence class rather than on a particular manuscript realization. Our implemented extractor $\widehat{C}$ is not claimed to be maximal or exactly invariant. It instead approximates the two properties in Equation 3: stability across matched variants designed to preserve the reported science and separation across different papers. The complete SciCore reviewer does not replace manuscript-based judgment with this representation. It uses the resulting content-normalized judgment as one branch of the final assessment.

A.2 CONDITIONAL STABILITY OF THE SCIENCE-CORE BRANCH AND FUSION

Specializing the downstream judgment to a numerical overall-assessment score, let $r : \mathcal{Z} \to \mathbb{R}$ denote the score map applied to the implemented science-core representation, and let $d_{\mathcal{Z}}$ be a task-relevant distance between extracted representations. If r is L-Lipschitz on the observed representation range, then

$$\left| r\Big(\widehat{C}(x_{ik})\Big) - r\Big(\widehat{C}(x_{i0})\Big) \right| \le L\, d_{\mathcal{Z}}\Big(\widehat{C}(x_{ik}), \widehat{C}(x_{i0})\Big). \tag{4}$$

Averaging over the controlled comparisons gives the corresponding deterministic-score MAD bound

$$\mathrm{MAD}_{r\circ\widehat{C}} \le \frac{L}{NK} \sum_{i=1}^{N} \sum_{k=1}^{K} d_{\mathcal{Z}}\Big(\widehat{C}(x_{ik}), \widehat{C}(x_{i0})\Big). \tag{5}$$

For stochastic reviewing, the same inequality applies to an L-Lipschitz conditional mean score map, while realized single-run scores include additional decoding variation. Equation 5 should therefore be interpreted as a deterministic or conditional-mean analogue of the empirical MAD in Appendix D.4, not as a bound on every realized review. The relation is not an empirical guarantee because we do not establish exact rewrite preservation, identify the reviewer's task-relevant metric,

or estimate L. It addresses only direct within-paper stability; whether the reviewer retains distinctions among papers must still be assessed empirically through the joint stability-discrimination metrics.

Let $M(x)$ denote the manuscript-branch score and let $B(x) = r(\widehat{C}(x))$ denote the science-core-branch score. Applied to RobustReview, $M(x_{ik})$ and $B(x_{ik})$ are the branch-specific instances of y_{mik} in Section 2, with the fixed configuration index suppressed. Under the fusion rule in Equation 2, the triangle inequality and Equation 5 give

$$\mathrm{MAD}_{S_{\text{SciCore}}} \leq \frac{1}{2}\mathrm{MAD}_M + \frac{1}{2}\mathrm{MAD}_B \leq \frac{1}{2}\mathrm{MAD}_M + \frac{L}{2NK}\sum_{i=1}^{N}\sum_{k=1}^{K} d_{\mathcal{Z}}\Big(\widehat{C}(x_{ik}), \widehat{C}(x_{i0})\Big). \quad (6)$$

This decomposition provides a route to attenuating rhetorical variation: when the science-core branch has lower rewrite-induced MAD than the manuscript branch, their average is bounded by a value below the manuscript branch's MAD. Stable extracted representations can contribute to this condition when the downstream score map is sufficiently regular, while the manuscript branch retains the complete paper context. The bound remains explanatory rather than a performance guarantee, and the fused reviewer must still be evaluated using the joint stability-discrimination metrics.

A.3 Reconstruction as a Blackwell Garbling

The invariant-representation view explains why a science core can support rhetorically stable judgment, but it does not by itself distinguish direct review from first realizing the record as another manuscript. Blackwell's comparison of statistical experiments supplies an information-theoretic rationale for avoiding an unnecessary reconstruction layer, rather than a performance guarantee for a fixed reviewer (Blackwell, 1953). Let $Z = \widehat{C}(X)$ be the extracted record, let U collect any auxiliary assets supplied unchanged to reconstruction, and write $W = (Z, U)$ for the complete reconstruction input. If a reconstructed manuscript is sampled through a channel $\widetilde{X} \sim Q(\cdot \mid W)$, then for any review-relevant state Θ,

$$\Theta \longrightarrow W \longrightarrow \widetilde{X} \quad (7)$$

forms a Markov chain.

Proposition A.2 (Reconstruction is a Blackwell garbling)**.** *Observing the reconstruction input W Blackwell-dominates observing the reconstructed manuscript $\widetilde{X}$. Consequently, across decision problems, the best attainable risk using W is no worse than the best attainable risk using $\widetilde{X}$.*

Proof. For any reconstruction-based decision rule $\delta_{\widetilde{X}}(a \mid \widetilde{x})$, an observer of W can first sample $\widetilde{X}$ from Q and then apply that rule:

$$\delta_W(a \mid w) = \int \delta_{\widetilde{X}}(a \mid \widetilde{x}) Q(d\widetilde{x} \mid w).$$

The two procedures induce the same conditional distribution of decisions given Θ, so every risk attainable from $\widetilde{X}$ is attainable from W. □

Thus reconstruction cannot add decision-relevant information absent from its inputs, although it can discard information or introduce another rhetorical realization. This proposition establishes only the information ordering between the complete reconstruction input W and its reconstructed output $\widetilde{X}$. It does not imply that a fixed LLM will review W more effectively than $\widetilde{X}$. Moreover, the implemented reconstruction arm receives permitted figures and bibliography in addition to Z, whereas the science-core branch reviews only Z. The proposition therefore motivates the ReconstructReview ablation but does not order the two implemented pipelines; their comparison is empirical.

B Extended Related Work

B.1 Reliability and Robustness of LLM-as-a-Judge Systems

LLM-as-a-judge methods operationalize evaluation through natural-language rubrics, pairwise decisions, and task-specific evaluator models. Systems such as G-Eval, MT-Bench, and Prometheus show that these designs can reproduce human preferences and provide criterion-specific feedback (Liu et al., 2023; Zheng et al., 2023; Kim et al., 2024). Their judgments are nevertheless affected by candidate order, response length, prompt wording, and evaluator identity (Wang et al., 2024; Dubois et al., 2024; Stureborg et al., 2024; Panickssery et al., 2024). Further variation arises from epistemic markers and semantically equivalent evaluation instructions (Lee et al., 2025; Bhat & Varma, 2026), with meta-evaluation revealing substantial dependence on the judge, task, property, and data source (Bavaresco et al., 2025).

Prior work has also begun to intervene on the evaluation process itself. Length-controlled evaluation reduces a known presentation bias (Dubois et al., 2024), while representation-based judging replaces direct generation with an intermediate semantic representation (Li et al., 2026d). These approaches suggest that robustness can depend on what information reaches the judge, not only on which rubric the judge receives. Existing studies, however, largely address generic evaluation tasks or isolated biases. We instead test scientific reviewers for both stability across rewrites designed to preserve reported scientific content and discrimination across papers, and test whether a science-core branch improves this balance when combined with manuscript-based review.

Our theoretical view draws on two complementary traditions. Maximal invariants represent equivalence classes while retaining the information needed by invariant downstream tasks (Eaton, 1989; Dubois et al., 2021). Blackwell's comparison of experiments orders observations by the decision risks they make attainable and identifies stochastic post-processing as a garbling of its source (Blackwell, 1953). We use these results as design rationales for the science-core branch in SciCore and for avoiding manuscript reconstruction within that branch, not as performance guarantees for a fixed LLM reviewer.

B.2 LLMs for Scientific Peer Review

Before modern LLMs, computational peer-review research focused on data collection, decision prediction, and structured feedback generation. PeerRead linked manuscripts with expert reports, aspect scores, and publication decisions for predictive and analytical tasks (Kang et al., 2018). ReviewRobot and ReviewAdvisor then moved toward evidence-linked and aspect-specific critique (Wang et al., 2020; Yuan et al., 2022). This progression expanded the modeled review workflow beyond acceptance prediction, while leaving reliable, decision-relevant criticism as a central challenge.

Modern LLMs extend this progression from modeling individual review components to generating complete natural-language reviews at scale. In a large-scale study, GPT-4 feedback overlaps with human review comments at rates comparable to the overlap between human reviews, and many authors report finding such feedback useful (Liang et al., 2023). More direct evaluations remain qualified: LLMs struggle with long-paper processing, zero-shot scoring, and consistently correct criticism (Zhou et al., 2024); they reproduce summaries and stated strengths more readily than substantive weaknesses, discriminating questions, or differences in paper quality (Li et al., 2025a). Review quality also depends on prompting, retrieval, and reviewer-guideline design (Chen et al., 2026; Li et al., 2026a). At the same time, a randomized deployment shows that AI feedback can improve the specificity and actionability of human reviews (Thakkar et al., 2026). Together, these results support AI as a review aid but leave open whether its scientific judgments are stable under content-preserving rhetorical changes.

Recent specialized models and agentic systems seek stronger reviewing through fine-tuning, multi-stage reasoning, retrieval, verification, reviewer ensembles, or proactive evidence gathering (Idahl & Ahmadi, 2025; Weng et al., 2025; Zhu et al., 2025b; Lu et al., 2024; The OpenJudge Team, 2025; Fang et al., 2026). Their evaluations primarily emphasize review quality, similarity to human feedback, score prediction, or workflow capability. These goals are complementary to rhetorical robustness: a review may be detailed and plausible yet still assign different scientific scores to different presentations of the same work. Such sensitivity makes merit judgments depend on rhetorical

choices and undermines comparability across submissions. We therefore evaluate general-purpose LLM reviewers, specialized review models, and agentic review systems under the same controlled matched-family design.

B.3 Manipulation and Presentation Sensitivity in AI Scientific Review

Scientific evaluation legitimately responds to clarity and exposition, but these criteria should not be conflated with methodological soundness or scientific contribution (James et al., 2024; Li et al., 2025b). This distinction is difficult to study with naturally occurring papers because scientific merit and writing quality are entangled. Controlled interventions provide a more direct test by varying presentation through revisions designed to preserve the reported scientific content.

One line of work studies overtly adversarial manuscript interventions. Hidden instructions can inflate ratings, suppress criticism, or redirect generated reviews (Ye et al., 2024; Collu et al., 2026), while character-, word-, and sentence-level perturbations can distort review judgments (Lin et al., 2025). Such attacks establish serious vulnerabilities but differ from ordinary academic rewriting because they rely on concealed instructions or artificial perturbations. Related evidence also shows that metadata and other manuscript-side factors can introduce bias into automated review (Vasu et al., 2026; Zhu et al., 2025a).

A more closely related line examines visible revisions intended to preserve reported scientific content. Even title-level stylistic variants can alter AI-review scores (Du, 2025). At the abstract and full-paper levels, paraphrasing, rewriting, and overclaiming can be optimized to improve automated ratings (Kaneko, 2026; Li et al., 2026b), and recent work describes full-manuscript *paper laundering* or *adversarial repackaging* that raises review scores without new experiments (Baumann et al., 2026; Yang et al., 2026b). More general attacks similarly learn meaning-preserving stylistic edits that exploit a particular LLM judge (Yang et al., 2026a).

Li et al. (2026c) construct controlled full-manuscript rewrites to characterize how rhetorical choices reward-hack AI reviewers. Using complete manuscripts, our investigation evaluates within-paper stability, cross-paper discrimination, and human alignment. We introduce SciCore, which combines manuscript-level assessment with a content-normalized scientific judgment, and examine direct science-core review against manuscript reconstruction. Complementarily, Dycke & Gurevych (2026) changes substantive reasoning to test defect detection, whereas our counterfactuals are constructed to preserve reported reasoning and evidence when testing rhetorical invariance. Sensitivity to substantive defects and invariance to content-preserving rhetorical changes are complementary requirements for robust review.

C Detailed Ablations and Analyses

C.1 Representation Analysis: Science-Core Stability and Paper Separation

A useful science-core representation should remain similar across rhetorical variants of the same paper while remaining distinct across different papers. Table 3 compares each variant core with its matched original and with nonmatching originals across three extractors and two rewrite producers with cosine similarity of text embeddings.

Matched cores are consistently much more similar than cores from different papers. For GPT-5.5, matched similarity is 0.978, while cross-paper similarity ranges from 0.616 to 0.619. This gap provides embedding-level evidence that the extracted science cores remain stable across rhetorical rewrites while retaining paper-specific differences. Condition-level comparisons are reported in Appendix F.5.

Table 3: **Science-core cosine similarity across extractor models and rewrite producers.** All matched comparisons pair the science core extracted from each rewritten manuscript with the core extracted from the original manuscript of the same paper; different-paper controls pair it with the 59 nonmatching original cores. Each matched cell contains 600 pairs and each control cell contains 35,400 comparisons per producer. P05 denotes the fifth percentile of pair-level similarities.

Extractor Model	Comparison	GPT-5.5 producer			Opus 4.8 producer		
		Mean	Median	P05	Mean	Median	P05
GPT-5-mini	All matched	0.975	0.977	0.964	0.973	0.976	0.962
	Different-paper control	0.668	0.668	0.582	0.670	0.671	0.585
GPT-5.5	All matched	0.978	0.979	0.966	0.978	0.979	0.965
	Different-paper control	0.616	0.617	0.516	0.619	0.620	0.518
Gemini-3.5-Flash-Lite	All matched	0.963	0.965	0.934	0.961	0.964	0.934
	Different-paper control	0.606	0.606	0.506	0.609	0.608	0.509

C.2 Criterion-Level Behavior of the Science-Core Branch

Table 4 shows that branch behavior is criterion-dependent. Science-core review generally improves soundness stability, but the contribution is less consistent, and human alignment does not improve uniformly. Presentation is diagnostic because manuscript-based review evaluates the paper's presentation, whereas science-core review evaluates the clarity and completeness of the extracted scientific record.

Confidence exposes a concrete false-robustness failure mode. Core-Standard and Core-Strict assign a constant confidence score, and Core-Persistent produces almost no variation, so their near-zero drift reflects output collapse and coincides with zero or near-zero SPR. Core-Adapted restores paper-level variation and raises confidence SPR to 0.394. This result again shows why within-paper stability cannot be interpreted without discrimination.

Table 4: **Secondary-score behavior across GPT-5.5 branch configurations.** For each secondary score, we report direct within-paper stability (MAD), joint stability-discrimination (SPR), and Spearman correlation with the corresponding mean human score. Presentation is marked as diagnostic because its meaning differs between manuscript-based and direct science-core review. A dash denotes an undefined Spearman correlation due to constant scores; SPR is defined as zero for fully constant outputs.

Configuration	Soundness			Contribution			Presentation†			Confidence		
	MAD ↓	SPR ↑	ρ_H ↑	MAD ↓	SPR ↑	ρ_H ↑	MAD ↓	SPR ↑	ρ_H ↑	MAD ↓	SPR ↑	ρ_H ↑
Manuscript-Based Review												
Manuscript-Standard	0.196	0.547	0.099	0.218	0.587	0.350	0.084	0.371	-0.109	0.232	0.375	-0.013
ReconstructReview	0.112	0.625	0.182	0.214	0.507	0.301	0.240	0.375	-0.094	0.289	0.402	0.052
Science-Core Branch												
Core-Standard	0.135	0.645	0.046	0.139	0.593	0.226	0.197	0.524	-0.078	0.000	0.000	–
Core-Strict	0.117	0.626	0.178	0.164	0.586	0.428	0.220	0.537	0.084	0.000	0.000	–
Core-Persistent	0.178	0.581	0.170	0.250	0.505	0.326	0.193	0.503	0.192	0.003	0.000	–
Core-Adapted	0.141	0.633	0.095	0.175	0.537	0.320	0.171	0.543	0.041	0.369	0.394	0.026

C.3 Cross-Backbone Generalization of the Science-Core Branch

Table 5 compares Core-Adapted with Manuscript-Standard under the same backbone. We use STANDARD as a common baseline rather than selecting the strongest direct-review protocol separately for each model. Core-Adapted improves ICC for all three backbones. GPT-5.5 improves on all five robustness metrics, GPT-5-mini improves MAD, Drift SD, and ICC but weakens SPR and discriminability, and GLM-5.2 improves all seven reported metrics. Human alignment does not improve consistently.

Table 5: **Cross-backbone evaluation of the science-core branch.** For each backbone, Manuscript-Standard applies the STANDARD protocol to the full paper, whereas Core-Adapted uses the same backbone for science-core extraction and review. Bold marks the better result within each backbone pair.

Backbone	Configuration	Human alignment		Within-paper stability		Joint stability-discrimination		
		H-MAE ↓	Spearman ↑	MAD ↓	Drift SD ↓	ICC ↑	SPR ↑	Discrim. ↑
GPT-5.5	Manuscript-Standard	**1.294**	**0.448**	0.598	0.956	0.615	0.555	0.649
	Core-Adapted	1.461	0.285	**0.292**	**0.650**	**0.751**	**0.710**	**0.695**
GPT-5-mini	Manuscript-Standard	**1.711**	**0.417**	0.543	0.945	0.401	**0.390**	**0.586**
	Core-Adapted	1.794	0.290	**0.367**	**0.866**	**0.408**	0.371	0.577
GLM-5.2	Manuscript-Standard	2.044	−0.034	1.500	2.341	0.226	0.322	0.558
	Core-Adapted	**1.912**	**0.413**	**1.467**	**2.169**	**0.322**	**0.386**	**0.591**

The manuscript-side Strict evaluations for these backbones are already included in Table 1. Here, we focus on the cross-backbone behavior of the science-core branch, changing the extraction and core-review model together. All reviewer configurations are evaluated on the same matched corpus, which includes rewrites produced independently by GPT-5.5 and Claude Opus 4.8.

C.4 SCIENCE-CORE WEIGHT SENSITIVITY

As a sensitivity analysis of SciCore, we alter the science-core weight α to test how the fusion balance affects performance. We examine the science-core branch paired with Manuscript-Standard, Manuscript-Strict, and Manuscript-Persistent reviews:

$$S_{\alpha,p}(x) = \alpha B(x) + (1-\alpha)M_p(x), \tag{8}$$

where $B(x)$ is the science-core-branch score and $M_p(x)$ is the manuscript-branch score under protocol p. Our prespecified configuration uses Manuscript-Strict with $\alpha = 0.5$. The sweep characterizes sensitivity around this fixed design. For descriptive visualization, each metric is direction-aligned and min-max normalized over the full $\alpha \in [0, 1]$ sweep.

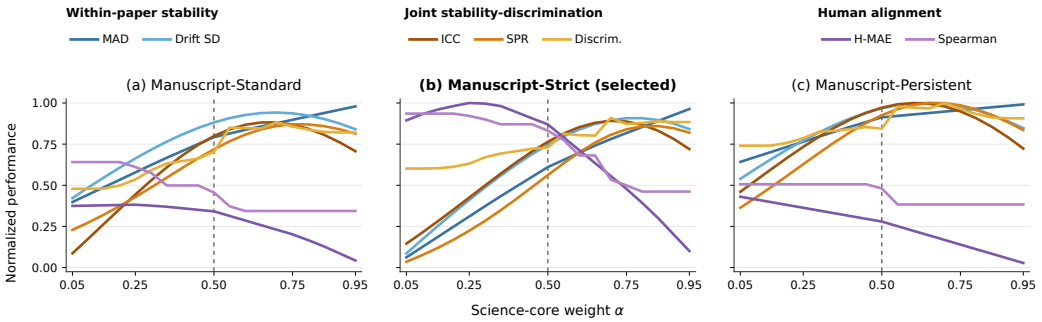


Figure 2: **Science-core weight sensitivity.** Normalized performance across science-core weights under three manuscript protocols. Higher values indicate better performance for each metric. The dashed line marks the prespecified equal-weight fusion ($\alpha = 0.5$). Complete numerical results are reported in Tables 17–19.

Larger science-core weights tend to improve robustness while weakening human alignment overall. Across the three manuscript protocols, Manuscript-Persistent favors robustness most strongly, while Manuscript-Strict retains the best human alignment. Under Manuscript-Strict, $\alpha = 0.5$ lies near the transition between these two objectives and maintains strong ICC, SPR, and discriminability without a large loss in human alignment.

The prespecified equal-weight fusion balances rhetorical robustness and human alignment and is Pareto non-dominated among the evaluated protocol-weight combinations on the seven reported point estimates.

D EXPERIMENTAL DETAILS

This section provides additional details on RobustReview construction and experiments, SciCore, and its ablation studies. Our rhetorical intervention design, rewrite construction procedures, and manuscript-review protocols follow the methodology of Li et al. (2026c), with prompts instantiated for the present benchmark. All manuscript rewrites and model-review scores reported here were generated for this study.

D.1 ROBUSTREVIEW COMPOSITION AND SAMPLING

We construct RobustReview from ICLR 2026 submissions using metadata downloaded through the OpenReview API. We first identify submissions for which title matching yields a corresponding arXiv paper with an available source package. To ensure that the matched arXiv manuscript closely corresponds to the ICLR submission, we require the text similarity between the ICLR submission and at least one current or historical arXiv version to exceed 0.8. Among submissions satisfying these criteria, we stratify papers by their mean human overall-assessment rating. Using a fixed random seed of 42, we randomly sample 10 papers from each of the six rating intervals $[1, 3)$, $[3, 4)$, $[4, 5)$, $[5, 6)$, $[6, 7)$, and $[7, 8.5]$, yielding 60 source papers. Equal allocation across these intervals ensures coverage of lower-, middle-, and higher-rated submissions. We retain the mean human ratings as the reference for the human-alignment evaluation. The benchmark focuses on ICLR 2026 submissions with matched arXiv sources. Generalization to other venues, fields, and manuscript formats remains to be evaluated.

We construct 10 rhetorical rewrite conditions for each source paper. Six conditions each apply a positive-direction transformation to a single rhetorical dimension. *Novelty stance* changes how strongly the novelty of the work is presented. *Scope framing* changes how broadly or narrowly the scope and generalizability of the work are presented. *Evidence framing* changes how the reported evidence and quantitative results are characterized. *Contribution salience* changes the prominence and organization of the stated contributions. *Technical register* changes the style and degree of technical and formal presentation. *Linguistic complexity* changes lexical and syntactic realization. The remaining 4 conditions introduce broader transformations. *Joint rewrite* varies multiple rhetorical dimensions together. *Recursive rewrite (R2)* and *recursive rewrite (R3)* correspond to 2 and 3 successive rounds of rewriting, respectively. *Reviewer-guided rewrite* uses reviewer feedback to guide the rhetorical transformation.

Each condition contains 2 independently generated variants for each of the 60 source papers, yielding 120 variants per condition and 1,200 rhetorical variants across all 10 conditions. Together with the 60 original manuscripts, RobustReview contains 1,260 full-manuscript PDFs. During evaluation, each variant is paired with the corresponding original manuscript from the same source-paper family, yielding 120 matched original-variant pairs per condition and 1,200 matched pairs overall.

D.2 REWRITE CONSTRUCTION AND STRUCTURAL CONTROLS

Each rhetorical condition is instantiated independently by GPT-5.5 through Codex CLI and Claude Opus 4.8 (Anthropic, 2026a) through Claude Code. Rewriting is performed on the complete LaTeX project associated with each source paper rather than on isolated sections or extracted text. The transformation instructions require the models to preserve the reported scientific content, including claims, methods, evidence, numerical results, findings, and conclusions, while allowing the targeted changes in rhetorical framing, organization, wording, and presentation, including how quantitative evidence or tables are described and presented.

Citations, cross-references, figures, bibliography files, and other project dependencies are protected during rewriting. Each transformed project is subsequently recompiled to verify structural validity. Outputs that fail structural or compilation checks are repaired when possible and otherwise excluded from the benchmark. These controls are designed to permit substantial changes in presentation while minimizing unintended changes to the reported scientific content.

D.3 REVIEWER EXECUTION DETAILS

We evaluate eight general-purpose LLMs prompted as scientific reviewers: GPT-5.5 (OpenAI, 2026), GPT-5-mini (OpenAI, 2025), Claude Sonnet 5 (Anthropic, 2026b), GLM-5.2 (GLM-5-Team et al., 2026), Kimi-K2.6 (Moonshot AI, 2026), GPT-OSS-120B (OpenAI, 2025), Gemini-3.5-Flash-Lite (Google, 2026), and Qwen-3.5-Flash (Qwen Team, 2026). This reflects a common use case in which a general-purpose LLM is directly instructed to review a scientific manuscript. Each model is evaluated under three protocols. STANDARD follows the standard ICLR review criteria and scoring scheme. STRICT and PERSISTENT are derived from STANDARD: STRICT applies a more demanding, evidence-focused rubric, and PERSISTENT repeatedly emphasizes that scientific judgments should depend on substantive scientific content rather than rhetorical presentation. We additionally evaluate three specialized scientific-review models: OpenReviewer (Idahl & Ahmadi, 2025), CycleReviewer (Weng et al., 2025), and DeepReviewer (Zhu et al., 2025b). We also evaluate three agentic review systems: AI Scientist (Lu et al., 2024), OpenJudge (The OpenJudge Team, 2025), and ProReviewer (Fang et al., 2026). The specialized and agentic systems use their native review procedures.

All prompted-review experiments operate directly on the manuscript PDFs. GPT-5.5 and GPT-5-mini are accessed through the OpenAI Responses API,[1] using the default API settings without overriding sampling or generation parameters. Claude Sonnet 5, GLM-5.2, Kimi-K2.6, GPT-OSS-120B, Gemini-3.5-Flash-Lite, and Qwen-3.5-Flash are accessed through the OpenRouter API.[2] For GLM-5.2, Kimi-K2.6, and GPT-OSS-120B, the reasoning effort is set to `high`. All other generation and inference parameters are left at their OpenRouter defaults. Each review request consists of the complete manuscript PDF together with the corresponding review prompt. For models accessed through OpenRouter, PDF ingestion and processing are handled by OpenRouter's file-processing pipeline.

The specialized reviewer systems are evaluated using their officially released checkpoints and inference procedures. Specifically, we use Llama-OpenReviewer-8B for OpenReviewer, CycleReviewer-ML-Llama-3.1-8B for CycleReviewer, and DeepReviewer-14B for DeepReviewer. CycleReviewer and DeepReviewer follow their released multi-reviewer evaluation procedures, with the resulting reviewer scores aggregated according to their native implementations. These models are served locally on a server equipped with 8 NVIDIA A100 GPUs.

For specialized reviewers whose released interfaces do not accept PDF input, we first convert each manuscript to Markdown with the Datalab OCR API[3] and pass the resulting Markdown to the model's native inference pipeline.

The agentic review systems, AI Scientist, OpenJudge, and ProReviewer, are evaluated using their officially released implementations and configurations. We retain the model backbones, inference parameters, review workflows, and score extraction procedures specified by their respective implementations without additional modification.

Single-review execution. We use one independent review per manuscript version in the main experiments, as repeated reviewing would substantially increase the computational cost of RobustReview. We conduct an auxiliary audit with GPT-5-mini under the Standard protocol, covering all 60 source papers, both rewrite producers, and the six single-dimension rewrite conditions, for a total of 720 rewritten manuscripts. For each manuscript, we compare a single review with the mean of three independent reviews of the same PDF. Averaging three reviews changes the mean OA from 6.493 to 6.457. The 12 condition–producer mean effect estimates are closely aligned between the two settings, with Pearson and Spearman correlations of 0.975 and 0.944, respectively. The mean absolute change in these effects is 0.056 OA points, with a maximum change of 0.117. This audit supports the consistency of condition-level mean effects under the evaluated configuration. Our robustness metrics characterize observed score variation under single-review execution, including stochastic generation variability.

Appendix F.3 reports the same-PDF repeatability baseline for the manuscript protocols and SciCore.

[1] `https://platform.openai.com/docs/api-reference/responses`
[2] `https://openrouter.ai/docs/features/multimodal/pdfs`
[3] `https://documentation.datalab.to/`

D.4 Full Evaluation Metric Definitions

All configurations are evaluated on the same 60 complete paper families (1,260 manuscript versions), with no missing overall-assessment scores.

For reviewer configuration m and a given score dimension, we define the rewrite-induced score change as

$$\Delta_{mik} = y_{mik} - y_{mi0}, \qquad k = 1, \ldots, K. \tag{9}$$

Direct within-paper stability. MAD measures the average magnitude of score changes induced by rhetorical rewriting:

$$\mathrm{MAD}_m = \frac{1}{NK} \sum_{i=1}^{N} \sum_{k=1}^{K} |\Delta_{mik}| \,. \tag{10}$$

Drift SD measures the variability of the signed score changes:

$$\mathrm{DriftSD}_m = \mathrm{SD}_{i,k} \left(\Delta_{mik} \right). \tag{11}$$

Lower values indicate greater rhetorical stability. MAD captures the overall magnitude of score drift, whereas Drift SD captures its heterogeneity across papers and rewrite conditions.

Joint stability-discrimination. Direct within-paper stability is insufficient when low drift is produced by score collapse. We therefore use three complementary metrics that relate within-paper consistency to cross-paper variation or separation. These are joint metrics rather than between-paper-only measures. ICC measures whether different presentations of the same paper receive similar scores relative to differences across papers. We use the two-way mixed-effects, absolute-agreement, single-measure form, ICC(A,1):

$$\mathrm{ICC}_m(A,1) = \frac{MS_R - MS_E}{MS_R + (P-1)MS_E + \frac{P}{N}(MS_C - MS_E)}, \tag{12}$$

where MS_R, MS_C, and MS_E denote the paper, presentation, and residual mean squares, respectively; N is the number of papers; and $P = K + 1$ is the number of presentations including the original manuscript. Higher ICC indicates that between-paper differences are large relative to presentation- induced and residual variation.

We further define the signal preservation ratio (SPR) as

$$\mathrm{SPR}_m = \frac{\mathrm{Var}_i \left(y_{mi0} \right)}{\mathrm{Var}_i \left(y_{mi0} \right) + \mathbb{E}_{i,k} \left[\Delta^2_{mik} \right]}. \tag{13}$$

SPR compares the cross-paper score variation in the original manuscripts with the magnitude of rewrite-induced movement. Higher values indicate that cross-paper variation remains large relative to within-paper rhetorical perturbation. When both the original-paper score variance and the rewrite-induced mean squared drift are zero, we define SPR as zero, reflecting the absence of between-paper discrimination.

Finally, discriminability measures whether presentations of the same paper are closer in score than presentations of different papers:

$$\begin{aligned} \mathrm{Disc}_m = \mathbb{E}_{\substack{i \neq j,\ a \neq b \\ a,b,c \in \{0,\ldots,K\}}} \Big[& \mathbf{1} \left(|y_{mia} - y_{mib}| < |y_{mia} - y_{mjc}| \right) \\ & + \frac{1}{2} \mathbf{1} \left(|y_{mia} - y_{mib}| = |y_{mia} - y_{mjc}| \right) \Big]. \end{aligned} \tag{14}$$

Here, a and b index two presentations of paper i, while c indexes a presentation of a different paper j. A value of 0.5 is chance-like, while higher values indicate stronger separation between same-paper and different- paper presentations.

ICC, SPR, and discriminability operationalize the same joint requirement in different ways: each rewards consistency across presentations of the same paper only when distinctions across papers remain detectable. None treats cross-paper score variation as ground-truth scientific merit.

Human judgment alignment. Let h_i denote the mean human OA score for the original version of paper i. Human MAE measures absolute agreement between AI and human scores:

$$\mathrm{HMAE}_m = \frac{1}{N}\sum_{i=1}^{N} |y_{mi0} - h_i| \,. \tag{15}$$

We additionally measure rank agreement using Spearman rank correlation:

$$\rho_m^{\mathrm{human}} = \rho_{\mathrm{S}}\left((y_{mi0})_{i=1}^{N}, (h_i)_{i=1}^{N}\right). \tag{16}$$

Lower Human MAE and higher Spearman rank correlation indicate stronger agreement with human reviewer judgments. These metrics assess agreement with mean human ratings and do not characterize disagreement among individual reviewers. Spearman correlation is undefined when either input is constant; ICC is undefined when its denominator is zero. We report these cases as dashes.

D.5 SciCore Execution Protocol

The Core-Adapted prompt, the Manuscript-Strict branch, and the equal-weight fusion rule were fixed before examining results on the 60-paper evaluation panel.

The SciCore experiments follow the same model-access and inference settings as the corresponding direct-review experiments described above. The science-core branch first uses GPT-5.5 to extract the science core from the manuscript and then reviews the extracted record with the adapted protocol. The manuscript branch is the GPT-5.5 Strict review of the complete PDF reported in the main benchmark. The final overall assessment is computed as the unweighted arithmetic mean of the manuscript-branch and science-core-branch scores. Both branches use the same ICLR overall-assessment scale, and their arithmetic fusion assumes that scores are comparable across branches. No additional model call is required for fusion. In the branch-policy ablation study, each extracted science core is reused across the evaluated policies; the reviewer backbone and input representation remain the same while the review instructions vary.

D.6 ReconstructReview Ablation

ReconstructReview evaluates whether the extracted record should be reviewed directly within the science-core branch or first realized as another manuscript. For each manuscript presentation, GPT-5.5 first extracts the science core using the same extraction procedure as SciCore. The extracted content is then provided to GPT-5.5 through Codex, which reconstructs it into a complete anonymous ICLR 2026 submission using the official conference template. The reconstruction workspace contains the extracted science core as the authoritative scientific record, together with the permitted bibliography, available scientific figures, and the ICLR 2026 template. The original manuscript prose is not provided to the reconstruction model.

The reconstruction model is given substantial authorial freedom over the title, framing, organization, mathematical presentation, allocation between the main text and appendix, and the selection and placement of figures and tables. At the same time, it is explicitly instructed to preserve the scientific record, retain nonredundant methods and experimental evidence, preserve reported numerical results and qualifications, and avoid introducing unsupported experiments, claims, citations, or implementation details. The resulting LaTeX project is compiled into a full-manuscript PDF and reviewed using the Standard protocol.

We apply this reconstruction procedure independently to every presentation in RobustReview, including the original baseline and the variants produced by both rewrite models under all 10 rhetorical conditions. This yields 21 reconstructed manuscripts per source paper and 1,260 reconstructed manuscripts in total. The resulting scores are then analyzed using the same matched original-variant evaluation procedure as in RobustReview.

Core of the ReconstructReview Prompt

 # Task: Compose an acceptance-quality ICLR 2026 paper

Create a complete ICLR 2026 submission from the materials provided in the workspace.

Scientific source of truth
Treat `scientific_report.md` as the authoritative record of the scientific content. Preserve all nonredundant method-critical definitions, assumptions, formal claims, experiments, ablations, results, and validity-relevant details supported by this record.
Do not invent experiments, results, datasets, models, hyperparameters, implementation details, citations, or project-specific claims. Preserve all reported numbers, experimental conditions, comparisons, uncertainties, and qualifications that you use.

Authorial freedom
Act as the lead author of the submission. Choose the title, framing, narrative, section structure, mathematical development, technical presentation, and allocation between the main text and appendix. Do not mechanically reproduce the organization or wording of the scientific report.

Permitted materials
Use only the supplied scientific report, bibliography, figures, and official ICLR 2026 template. All citations must come from the supplied bibliography, and supplied figures may be selected and placed as appropriate but may not be modified.

...

Figure 3: **Condensed ReconstructReview reconstruction prompt.** We show the core instructions governing the transformation from the extracted science core into a new manuscript. The omitted instructions primarily specify file handling, formatting, and submission requirements following the ICLR 2026 call for papers and official template.

E COMPLETE ROBUSTREVIEW RESULTS

This section reports the complete baseline evidence supporting the compact comparisons in the main paper. Baseline reviewer models and protocols are kept separate from the backbone-transfer experiments for SciCore.

E.1 FULL SCORE-DIMENSION RESULTS

Complete OA results are already reported in Table 1 and are not repeated here. The tables below use the same seven metrics and column order as the main table, but restrict the evaluated target to soundness, presentation, contribution, or confidence. Human alignment is computed against the mean score from the official reviews of each RobustReview source paper. Presentation scores were recovered directly from the OpenReview API; the other three secondary-score means exactly match the archived metadata. MAD and Drift SD directly measure within-paper stability, whereas ICC, SPR, and discriminability are joint stability-discrimination metrics. The GPT-5-mini entries are recomputed from their complete paper-by-rewrite score matrices rather than from the earlier aggregate-only export. External reviewers and the final SciCore fusion expose only OA-compatible outputs and are therefore not included in these secondary-score tables.

Table 6: **Full Soundness results for secondary scores.** Metrics and column order match Table 1. The best result in each column is shown in bold, the second-best is underlined, and the third-best is italicized.

System	Protocol	Human alignment		Within-paper stability		Joint stability-discrimination		
		H-MAE ↓	Spearman ↑	MAD ↓	Drift SD ↓	ICC ↑	SPR ↑	Discrim. ↑
General-purpose LLM reviewers								
GPT-5.5	Standard	0.529	0.099	0.196	0.432	0.520	*0.547*	0.614
	Strict	0.583	0.273	0.272	0.494	0.494	0.441	0.627
	Persistent	0.549	0.070	0.227	0.481	0.531	0.532	0.619
GPT-5-mini	Standard	**0.401**	0.214	**0.071**	**0.262**	0.061	0.401	0.501
	Strict	0.539	0.222	0.334	0.536	0.384	0.428	0.584
	Persistent	0.459	0.183	0.179	0.411	0.209	0.486	0.512
Claude Sonnet 5	Standard	*0.464*	**0.368**	0.247	0.491	0.541	0.528	*0.631*
	Strict	0.570	0.175	0.208	0.448	0.577	**0.576**	**0.644**
	Persistent	0.578	0.196	0.208	0.450	*0.566*	0.535	0.642
GLM-5.2	Standard	0.534	0.118	0.477	0.867	0.222	0.314	0.540
	Strict	0.778	−0.116	0.624	0.894	0.211	0.385	0.549
	Persistent	0.836	−0.222	0.677	1.013	0.227	0.385	0.547
Kimi-K2.6	Standard	0.550	0.194	0.430	0.786	0.175	0.404	0.530
	Strict	0.766	0.008	0.487	0.735	0.229	0.380	0.552
	Persistent	0.667	0.007	0.469	0.714	0.192	0.391	0.543
GPT-OSS-120B	Standard	0.500	−0.286	*0.167*	*0.405*	0.051	0.376	0.503
	Strict	0.649	−0.063	0.424	0.654	0.136	0.348	0.532
	Persistent	0.498	0.107	0.370	0.621	0.165	0.391	0.531
Gemini-3.5-Flash-Lite	Standard	1.076	0.212	0.167	0.391	0.256	0.490	0.522
	Strict	0.612	0.318	0.338	0.511	0.400	0.437	0.599
	Persistent	0.893	*0.312*	0.296	0.508	0.365	0.477	0.563
Qwen-3.5-Flash	Standard	0.943	−0.105	0.323	0.562	0.260	0.418	0.555
	Strict	0.621	0.220	0.359	0.607	0.226	0.360	0.548
	Persistent	0.803	0.212	0.338	0.560	**0.632**	0.566	0.590

Table 7: **Full Presentation results for secondary scores.** Metrics and column order match Table 1. The best result in each column is shown in bold, the second-best is underlined, and the third-best is italicized. A dash denotes an undefined Spearman correlation.

System	Protocol	Human alignment		Within-paper stability		Joint stability-discrimination		
		H-MAE ↓	Spearman ↑	MAD ↓	Drift SD ↓	ICC ↑	SPR ↑	Discrim. ↑
General-purpose LLM reviewers								
GPT-5.5	Standard	0.441	−0.109	0.084	0.287	0.232	0.371	0.510
	Strict	**0.393**	0.220	0.044	*0.208*	*0.306*	0.271	0.515
	Persistent	*0.402*	*0.285*	0.085	0.284	0.402	0.440	0.529
GPT-5-mini	Standard	0.410	0.181	0.114	0.324	0.263	0.220	0.526
	Strict	0.402	–	*0.044*	0.205	0.067	0.000	0.502
	Persistent	0.446	0.043	0.098	0.308	0.155	0.453	0.506
Claude Sonnet 5	Standard	0.491	0.033	0.185	0.415	0.241	0.447	0.526
	Strict	0.410	0.142	0.072	0.264	0.186	0.398	0.508
	Persistent	0.410	0.182	0.110	0.323	0.091	0.410	0.504
GLM-5.2	Standard	0.574	−0.044	0.464	0.864	0.201	0.309	0.535
	Strict	0.824	−0.243	0.628	0.985	0.165	0.407	0.532
	Persistent	0.807	−0.195	0.673	1.033	0.222	0.371	0.544
Kimi-K2.6	Standard	0.637	0.040	0.517	0.851	0.190	0.406	0.531
	Strict	0.707	−0.159	0.582	0.864	0.220	0.354	**0.548**
	Persistent	0.804	−0.099	0.588	0.834	0.213	0.390	0.542
GPT-OSS-120B	Standard	0.468	−0.169	0.155	0.396	0.037	0.294	0.502
	Strict	0.529	−0.086	0.333	0.601	0.062	0.316	0.507
	Persistent	0.510	−0.037	0.302	0.550	0.090	0.379	0.516
Gemini-3.5-Flash-Lite	Standard	1.196	0.216	**0.016**	**0.125**	0.135	**0.509**	0.501
	Strict	1.046	**0.364**	0.156	0.374	0.137	*0.471*	0.506
	Persistent	1.163	0.267	0.047	0.212	0.105	0.504	0.501
Qwen-3.5-Flash	Standard	0.963	0.304	0.367	0.607	0.105	0.382	0.518
	Strict	0.716	−0.014	0.498	0.688	0.074	0.318	0.519
	Persistent	0.740	0.148	0.499	0.665	**0.519**	0.454	*0.544*

Table 8: **Full Contribution results for secondary scores.** Metrics and column order match Table 1. The best result in each column is shown in bold, the second-best is underlined, and the third-best is italicized.

System	Protocol	Human alignment		Within-paper stability		Joint stability-discrimination		
		H-MAE ↓	Spearman ↑	MAD ↓	Drift SD ↓	ICC ↑	SPR ↑	Discrim. ↑
General-purpose LLM reviewers								
GPT-5.5	Standard	0.513	0.350	0.218	0.442	*0.600*	**0.587**	0.620
	Strict	0.407	0.409	0.181	0.416	0.583	0.520	*0.626*
	Persistent	*0.416*	*0.493*	0.217	0.455	0.629	0.583	**0.651**
GPT-5-mini	Standard	0.658	**0.515**	0.160	*0.374*	0.442	0.435	0.572
	Strict	0.441	0.331	0.375	0.538	0.295	0.398	0.562
	Persistent	0.666	0.202	0.139	**0.360**	0.432	0.452	0.545
Claude Sonnet 5	Standard	0.443	0.415	0.247	0.488	0.534	0.503	0.626
	Strict	**0.385**	0.488	*0.155*	0.391	**0.631**	*0.558*	0.622
	Persistent	0.455	0.283	0.268	0.508	0.452	0.481	0.613
GLM-5.2	Standard	0.721	0.026	0.537	0.926	0.185	0.308	0.538
	Strict	0.682	0.002	0.654	0.901	0.170	0.377	0.541
	Persistent	0.828	−0.083	0.732	1.041	0.214	0.379	0.545
Kimi-K2.6	Standard	0.727	0.046	0.508	0.854	0.150	0.415	0.526
	Strict	0.625	−0.059	0.404	0.663	0.209	0.359	0.538
	Persistent	0.635	−0.101	0.524	0.757	0.179	0.352	0.544
GPT-OSS-120B	Standard	0.649	0.031	0.237	0.494	0.053	0.397	0.504
	Strict	0.521	0.044	**0.138**	0.368	0.097	0.257	0.508
	Persistent	0.599	0.015	0.481	0.703	0.150	0.352	0.533
Gemini-3.5-Flash-Lite	Standard	1.357	0.307	0.162	0.385	0.243	0.481	0.519
	Strict	0.807	0.450	0.385	0.546	0.398	0.470	0.589
	Persistent	1.199	0.503	0.237	0.468	0.373	0.502	0.553
Qwen-3.5-Flash	Standard	1.049	0.338	0.342	0.574	0.236	0.420	0.557
	Strict	0.571	0.405	0.291	0.540	0.191	0.398	0.524
	Persistent	0.905	0.420	0.382	0.594	0.595	0.516	0.587

Table 9: **Full Confidence results for secondary scores.** Metrics and column order match Table 1. The best result in each column is shown in bold, the second-best is underlined, and the third-best is italicized. A dash denotes an undefined Spearman correlation or ICC.

System	Protocol	Human alignment		Within-paper stability		Joint stability-discrimination		
		H-MAE ↓	Spearman ↑	MAD ↓	Drift SD ↓	ICC ↑	SPR ↑	Discrim. ↑
General-purpose LLM reviewers								
GPT-5.5	Standard	0.500	−0.013	0.232	0.466	0.272	0.375	*0.550*
	Strict	0.516	−0.142	0.142	0.367	0.193	0.304	0.520
	Persistent	**0.466**	0.139	0.093	0.304	0.263	0.450	0.517
GPT-5-mini	Standard	0.489	–	**0.000**	**0.000**	–	0.000	0.500
	Strict	0.505	−0.216	0.025	0.157	0.002	0.396	0.500
	Persistent	0.514	−0.102	0.059	0.240	0.036	*0.445*	0.501
Claude Sonnet 5	Standard	0.629	−0.154	0.128	0.356	0.114	0.328	0.508
	Strict	0.626	−0.152	0.085	0.290	*0.285*	0.423	0.517
	Persistent	0.559	0.094	0.262	0.510	0.217	0.364	0.534
GLM-5.2	Standard	1.143	−0.101	0.836	1.271	0.054	0.300	0.543
	Strict	1.076	0.186	0.967	1.407	0.019	0.329	0.523
	Persistent	1.226	0.011	1.059	1.510	0.048	0.338	0.529
Kimi-K2.6	Standard	1.188	−0.036	0.765	1.136	0.145	0.342	0.552
	Strict	1.148	−0.011	0.841	1.245	0.065	0.311	0.538
	Persistent	1.218	**0.242**	0.777	1.127	0.101	0.344	0.541
GPT-OSS-120B	Standard	0.570	0.110	0.483	0.739	0.055	0.334	0.508
	Strict	0.672	−0.119	0.505	0.751	0.080	0.333	0.509
	Persistent	0.661	0.039	0.499	0.738	0.132	0.347	0.522
Gemini-3.5-Flash-Lite	Standard	1.139	−0.197	0.270	0.486	0.247	0.445	0.532
	Strict	0.922	−0.099	0.333	0.531	0.292	0.427	**0.567**
	Persistent	1.132	−0.159	0.240	0.450	0.261	**0.467**	0.528
Qwen-3.5-Flash	Standard	0.592	−0.012	0.234	0.477	0.059	0.316	0.508
	Strict	*0.489*	–	*0.040*	*0.196*	0.036	0.000	0.501
	Persistent	0.553	*0.151*	0.113	0.336	**0.334**	0.407	0.537

F Additional Validation

F.1 Core Technical-Content Fidelity Audit

We use DeepSeek V4 Flash as an independent auditor to assess preservation of core technical content between each rhetorical rewrite and its matched original. Rhetorical emphasis and evaluative framing are allowed to vary as part of the intended interventions. The auditor is not used as a rewrite producer, science-core extractor, or reviewer elsewhere in our experiments. The audit examines five dimensions: *table fidelity*, covering table structure, labels, values, and method–dataset–metric associations; *numerical fidelity*, covering numerical values, statistics, confidence intervals, and p-values in the manuscript body; *experimental fidelity*, covering datasets, evaluated subsets, splits, sample sizes, metrics, baselines, and experimental settings; *method fidelity*, covering method components, algorithmic operations, equations, variables, assumptions, and technical dependencies; and *result fidelity*, covering the direction and ordering of empirical results and positive versus negative observed effects.

The audit uses a binary decision rule. For paper i, rhetorical variant k, and applicable fidelity dimension d, let $z_{ik}^{(d)} = 1$ when no concrete technical-content mismatch is identified and $z_{ik}^{(d)} = 0$ when the auditor identifies a mismatch supported by evidence from the matched manuscripts. The fidelity rate for dimension d is

$$\text{Fidelity}_d = \frac{1}{N_d} \sum_{(i,k)\in\mathcal{A}_d} z_{ik}^{(d)}, \qquad N_d = |\mathcal{A}_d|,$$

where $\mathcal{A}_d$ contains the comparisons to which dimension d applies.

To complement the automated audit, we conduct a human validation on 150 comparisons between originals and variants. We first randomly sample 120 of the 1,200 rhetorical variants (10%), each evaluated against its matched original, and supplement them with 30 additional comparisons for which the automated auditor identifies at least one fidelity mismatch. Three graduate-level annotators independently assess all 150 comparisons using the same five dimensions: table, numerical, experimental, method, and result fidelity. For each comparison and dimension, the human verdict is determined by majority vote among the three annotators. Dimension-level human pass rates summarize these consensus verdicts across the combined 150-comparison audit sample. Overall inter-annotator agreement, measured as the mean pairwise agreement across all 750 dimension-level judgments, is 94.6%.

Table 10: **Core technical-content fidelity of rhetorical rewrites.** Dimension-level rates report the proportion of applicable comparisons that pass each check. Human rates summarize the combined audit sample of 120 randomly selected comparisons and 30 additional comparisons flagged by the automated auditor.

	Dimension-level fidelity				
Metric	Table	Numerical	Experimental	Method	Result
Automated pass rate (%)	95.2	96.4	94.8	97.3	96.5
Human pass rate (%)	97.3	98.0	96.7	98.7	97.3

Across the five assessed dimensions, the automated and human checks indicate that core technical content is largely preserved, with occasional discrepancies. The audit serves as a diagnostic assessment, and all comparisons are retained in the reported evaluation. Residual content differences may contribute to observed score variation.

F.2 Complementary Review Branches under Equal-Weight Fusion

Table 11 compares SciCore with full-manuscript score averages. We reuse cached scores, average them without rounding, and recompute all seven metrics. The Strict-anchored comparisons reuse the same manuscript-branch scores.

Some apparent robustness gains may arise from averaging itself: averaging integer scores allows half points, can reduce MAD and Drift SD through drift cancellation, and changes distance ties in discriminability. To assess whether SciCore retains an advantage under the same averaging operation, the two-score manuscript controls match its averaging rule and score resolution. The three-protocol average additionally matches its nominal three-call budget: three reviews versus one extraction and two reviews, with potentially different token costs.

Table 11: **Full-manuscript averaging controls on RobustReview.** All configurations use GPT-5.5 and the same 60 papers with 21 versions each. Protocol names denote full-manuscript reviews. Scores are averaged separately for each original and rewrite before computing the metrics, with no rounding. SciCore averages Manuscript-Strict and Core-Adapted. The three-protocol control averages three review scores. All entries are point estimates.

Configuration	Human alignment		Within-paper stability		Joint stability-discrimination		
	H-MAE ↓	Spearman ↑	MAD ↓	Drift SD ↓	ICC ↑	SPR ↑	Discrim. ↑
Single full-manuscript reviews							
Manuscript-Standard	1.294	0.448	0.598	0.956	0.615	0.555	0.649
Manuscript-Strict	1.078	0.529	0.766	1.202	0.632	0.507	0.672
Manuscript-Persistent	1.261	0.423	0.476	0.875	0.697	0.586	0.682
Two-score averages							
Standard + Strict	1.130	0.495	0.560	0.820	0.750	0.620	0.705
Standard + Persistent	1.180	0.455	0.470	0.681	0.755	0.640	0.712
Strict + Persistent	1.066	0.494	0.550	0.830	0.758	0.635	0.715
SciCore	1.072	0.488	0.476	0.687	0.775	0.652	0.726
Three-score average							
Standard + Strict + Persistent	1.110	0.466	0.520	0.681	0.781	0.635	0.715

With Manuscript-Strict and equal weighting fixed, Core-Adapted improves all five robustness metrics over either manuscript alternative, supported by paired bootstrap intervals (Appendix F.4). Thus, the robustness advantage over these two controls persists when the manuscript branch, averaging rule, and score resolution are matched. Both alternatives have higher Spearman correlation, and Strict + Persistent also has lower H-MAE. Standard + Persistent outperforms SciCore on MAD and Drift SD, while the three-protocol average outperforms it on Drift SD and ICC; SciCore has better point estimates on the other five metrics in each comparison and remains Pareto non-dominated.

F.3 Repeated-Review Noise Baseline

Following the repeated-review procedure in Appendix D.3, we obtain three independent reviews of each identical PDF, holding the review prompts and inference settings fixed. We apply this procedure to the three manuscript protocols and SciCore with GPT-5.5, following each configuration's evaluation pipeline. MAD and ICC quantify variation across repeated reviews of the same manuscript using the definitions in Appendix D.4.

Repeated reviews of identical PDFs yield MAD of 0.294–0.329 and ICC of 0.846–0.862 across the three manuscript protocols and SciCore. SciCore and Manuscript-Strict have similar repeatability (MAD: 0.320 versus 0.329; ICC: 0.858 versus 0.851), but differ more under rewriting (MAD: 0.476 versus 0.766; ICC: 0.775 versus 0.632), supporting a robustness gain beyond the small difference in same-PDF repeatability.

F.4 Paper-Family Bootstrap Analysis

Tables 12 and 13 use 5,000 paired bootstrap resamples of the 60 paper families, keeping each original and its 20 variants together. The same resamples are used across configurations to recompute metrics and paired differences. Marginal 95% intervals use the 2.5th and 97.5th percentiles, with the recorded reviews held fixed.

The paired intervals in Table 12 support SciCore's gains on all five robustness metrics over Manuscript-Strict and both two-score manuscript ensembles, with tradeoffs in human alignment. Relative to Core-Adapted, fusion improves human alignment, ICC, and discriminability, while increasing MAD and Drift SD and reducing SPR. Table 13 gives metric intervals for the primary comparison.

Table 12: **Paired bootstrap comparisons for SciCore.** All configurations use GPT-5.5 and the same 60 complete paper families. Each cell shows the difference (SciCore minus the column configuration), followed by its marginal 95% percentile interval from 5,000 paired paper-family resamples. Negative differences favor SciCore for H-MAE, MAD, and Drift SD; positive differences favor SciCore for the other metrics.

Metric	Manuscript-Strict	Core-Adapted	Strict + Standard	Strict + Persistent
H-MAE	-0.006 $[-0.011, -0.001]$	-0.389 $[-0.564, -0.214]$	-0.058 $[-0.087, -0.029]$	$+0.006$ $[+0.001, +0.011]$
Spearman	-0.041 $[-0.068, -0.014]$	$+0.203$ $[+0.053, +0.366]$	-0.007 $[-0.012, -0.002]$	-0.006 $[-0.011, -0.001]$
MAD	-0.290 $[-0.348, -0.232]$	$+0.184$ $[+0.084, +0.285]$	-0.084 $[-0.112, -0.056]$	-0.074 $[-0.101, -0.047]$
Drift SD	-0.515 $[-0.575, -0.455]$	$+0.037$ $[+0.006, +0.071]$	-0.133 $[-0.167, -0.099]$	-0.143 $[-0.180, -0.106]$
ICC	$+0.143$ $[+0.101, +0.185]$	$+0.024$ $[+0.004, +0.049]$	$+0.025$ $[+0.010, +0.040]$	$+0.017$ $[+0.005, +0.029]$
SPR	$+0.145$ $[+0.101, +0.189]$	-0.058 $[-0.101, -0.012]$	$+0.032$ $[+0.015, +0.049]$	$+0.017$ $[+0.005, +0.029]$
Discrim.	$+0.054$ $[+0.036, +0.072]$	$+0.031$ $[+0.001, +0.058]$	$+0.021$ $[+0.010, +0.032]$	$+0.011$ $[+0.003, +0.019]$

F.5 Science-Core Preservation Diagnostics

The preservation analysis compares the embedding of a science-core report extracted from an original manuscript with the embedding of the report extracted from its matched rhetorical variant:

$$\cos\Big(E(r_i^{\text{original}}), E(r_{ikp}^{\text{variant}})\Big).$$

Here i indexes papers, k rewrite conditions, and p rewrite producers. The condition-level tables keep GPT-5.5 and Opus 4.8 rewrite producers separate and report the mean, median, and fifth percentile.

We embed reports with `text-embedding-3-small` via the OpenAI API using default API parameters.

F.6 Complete Science-Core Weight Sensitivity Results

Figure 2 provides a descriptive, direction-normalized visualization of the science-core weight sweep. The complete unnormalized numerical results are reported in Tables 17–19. Each table fixes the manuscript protocol and varies the science-core weight α from 0 (manuscript only) to 1 (science core only). Intermediate rows use the continuous post-hoc fusion in Equation 8.

Table 13: **Paper-family bootstrap intervals for the primary comparison.** Entries are marginal 95% percentile intervals from 5,000 paired resamples of the 60 complete paper families; the corresponding point estimates appear in Table 1. All original and rewritten versions stay together within each resample.

System	Protocol	Human alignment		Within-paper stability		Joint stability-discrimination		
		H-MAE ↓	Spearman ↑	MAD ↓	Drift SD ↓	ICC ↑	SPR ↑	Discrim. ↑
GPT-5.5	Standard	[1.067, 1.531]	[0.213, 0.640]	[0.476, 0.717]	[0.809, 1.075]	[0.492, 0.698]	[0.441, 0.630]	[0.614, 0.679]
	Strict	[0.884, 1.287]	[0.305, 0.700]	[0.603, 0.927]	[1.038, 1.338]	[0.520, 0.709]	[0.416, 0.593]	[0.632, 0.702]
	Persistent	[1.037, 1.491]	[0.161, 0.638]	[0.364, 0.593]	[0.715, 1.017]	[0.573, 0.770]	[0.430, 0.696]	[0.640, 0.716]
GPT-5-mini	Standard	[1.387, 2.048]	[0.148, 0.634]	[0.431, 0.664]	[0.821, 1.057]	[0.276, 0.504]	[0.263, 0.487]	[0.553, 0.616]
	Strict	[1.014, 1.471]	[0.142, 0.616]	[0.749, 1.046]	[1.111, 1.306]	[0.283, 0.481]	[0.352, 0.487]	[0.567, 0.625]
	Persistent	[1.344, 1.919]	[0.098, 0.542]	[0.426, 0.749]	[0.792, 1.152]	[0.326, 0.581]	[0.357, 0.555]	[0.551, 0.623]
Claude Sonnet 5	Standard	[0.962, 1.432]	[0.137, 0.649]	[0.364, 0.577]	[0.730, 0.972]	[0.467, 0.671]	[0.439, 0.636]	[0.607, 0.678]
	Strict	[0.944, 1.372]	[0.214, 0.679]	[0.491, 0.740]	[0.919, 1.157]	[0.422, 0.641]	[0.457, 0.627]	[0.601, 0.668]
	Persistent	[1.027, 1.473]	[0.053, 0.602]	[0.218, 0.414]	[0.571, 0.834]	[0.378, 0.568]	[0.350, 0.575]	[0.573, 0.648]
GLM-5.2	Standard	[1.706, 2.394]	[−0.296, 0.241]	[1.264, 1.744]	[2.071, 2.572]	[0.143, 0.305]	[0.196, 0.411]	[0.539, 0.576]
	Strict	[1.687, 2.393]	[−0.294, 0.222]	[1.391, 1.709]	[1.898, 2.250]	[0.139, 0.283]	[0.337, 0.429]	[0.538, 0.572]
	Persistent	[2.002, 2.822]	[−0.451, 0.094]	[1.663, 2.218]	[2.407, 2.902]	[0.147, 0.308]	[0.328, 0.432]	[0.536, 0.579]
Kimi-K2.6	Standard	[1.374, 2.025]	[0.027, 0.504]	[1.128, 1.595]	[1.698, 2.237]	[0.131, 0.291]	[0.357, 0.474]	[0.535, 0.577]
	Strict	[1.419, 2.072]	[−0.200, 0.300]	[0.610, 0.912]	[1.093, 1.424]	[0.135, 0.431]	[0.280, 0.447]	[0.522, 0.569]
	Persistent	[1.261, 1.923]	[−0.233, 0.329]	[0.987, 1.319]	[1.453, 1.824]	[0.161, 0.410]	[0.347, 0.426]	[0.543, 0.596]
GPT-OSS-120B	Standard	[1.298, 1.798]	[−0.195, 0.336]	[0.586, 0.863]	[0.940, 1.297]	[0.089, 0.186]	[0.277, 0.443]	[0.521, 0.541]
	Strict	[1.379, 2.024]	[−0.163, 0.375]	[0.276, 0.511]	[0.733, 1.013]	[0.049, 0.132]	[0.176, 0.428]	[0.504, 0.513]
	Persistent	[1.289, 1.774]	[−0.250, 0.237]	[0.707, 1.043]	[1.096, 1.545]	[0.112, 0.207]	[0.266, 0.392]	[0.526, 0.544]
Gemini 3.5 Flash-Lite	Standard	[2.828, 3.532]	[0.208, 0.556]	[0.105, 0.323]	[0.450, 0.773]	[0.074, 0.311]	[0.340, 0.521]	[0.503, 0.523]
	Strict	[1.553, 2.209]	[0.242, 0.654]	[0.691, 0.998]	[1.071, 1.296]	[0.316, 0.512]	[0.431, 0.587]	[0.573, 0.622]
	Persistent	[2.491, 3.176]	[0.271, 0.649]	[0.268, 0.651]	[0.715, 1.218]	[0.217, 0.449]	[0.446, 0.540]	[0.518, 0.561]
Qwen 3.5 Flash	Standard	[1.988, 2.722]	[0.129, 0.595]	[0.645, 1.019]	[1.105, 1.628]	[0.187, 0.388]	[0.396, 0.516]	[0.538, 0.584]
	Strict	[1.046, 1.433]	[0.180, 0.586]	[0.739, 1.009]	[1.177, 1.487]	[0.133, 0.311]	[0.304, 0.468]	[0.524, 0.558]
	Persistent	[1.699, 2.393]	[0.244, 0.635]	[0.752, 0.975]	[1.167, 1.372]	[0.228, 0.752]	[0.369, 0.640]	[0.553, 0.642]
OpenReviewer	–	[1.126, 1.708]	[0.096, 0.550]	[0.929, 1.251]	[1.328, 1.703]	[0.140, 0.279]	[0.300, 0.424]	[0.527, 0.556]
CycleReviewer	–	[1.159, 1.660]	[−0.172, 0.310]	[0.663, 0.913]	[0.895, 1.178]	[0.081, 0.176]	[0.292, 0.389]	[0.521, 0.545]
DeepReviewer	–	[1.204, 1.749]	[0.178, 0.587]	[0.362, 0.553]	[0.583, 0.784]	[0.087, 0.229]	[0.293, 0.433]	[0.523, 0.549]
AI Scientist	–	[1.219, 1.822]	[−0.171, 0.347]	[0.698, 0.989]	[1.027, 1.471]	[0.142, 0.273]	[0.240, 0.461]	[0.529, 0.559]
OpenJudge	–	[1.112, 1.504]	[−0.131, 0.362]	[0.244, 0.434]	[0.501, 0.678]	[0.107, 0.345]	[0.338, 0.512]	[0.509, 0.548]
ProReviewer	–	[1.260, 1.697]	[−0.017, 0.463]	[0.632, 0.796]	[0.882, 1.057]	[0.058, 0.128]	[0.275, 0.385]	[0.509, 0.524]
SciCore	–	[0.878, 1.278]	[0.248, 0.678]	[0.394, 0.558]	[0.600, 0.759]	[0.667, 0.835]	[0.545, 0.731]	[0.678, 0.760]

Table 14: **Condition-level science-core similarity for GPT-5-mini.** P05 denotes the fifth percentile across comparisons. The cross-paper baseline compares variant cores with nonmatching original cores.

Rhetorical condition	GPT-5.5 producer			Opus 4.8 producer		
	Mean	Median	P05	Mean	Median	P05
Novelty stance	0.975	0.977	0.963	0.965	0.976	0.960
Scope framing	0.976	0.976	0.969	0.976	0.977	0.965
Evidence framing	0.972	0.974	0.957	0.973	0.975	0.958
Contribution salience	0.974	0.976	0.964	0.975	0.976	0.963
Technical register	0.976	0.976	0.968	0.974	0.975	0.964
Linguistic complexity	0.966	0.978	0.966	0.977	0.977	0.968
Joint rewrite	0.977	0.978	0.969	0.975	0.977	0.962
Recursive rewrite (R2)	0.977	0.978	0.967	0.974	0.974	0.962
Recursive rewrite (R3)	0.975	0.978	0.963	0.974	0.975	0.963
Reviewer-guided rewrite	0.976	0.977	0.964	0.972	0.974	0.958
Cross-paper baseline	0.668	0.668	0.582	0.670	0.671	0.585

Table 15: **Condition-level science-core similarity for GPT-5.5.** P05 denotes the fifth percentile across comparisons. The cross-paper baseline compares variant cores with nonmatching original cores.

Rhetorical condition	GPT-5.5 producer			Opus 4.8 producer		
	Mean	Median	P05	Mean	Median	P05
Novelty stance	0.979	0.979	0.968	0.979	0.981	0.967
Scope framing	0.978	0.978	0.965	0.980	0.980	0.969
Evidence framing	0.976	0.978	0.964	0.977	0.979	0.961
Contribution salience	0.976	0.978	0.963	0.977	0.979	0.965
Technical register	0.978	0.978	0.968	0.976	0.977	0.966
Linguistic complexity	0.979	0.980	0.967	0.980	0.981	0.970
Joint rewrite	0.980	0.981	0.970	0.979	0.979	0.968
Recursive rewrite (R2)	0.979	0.980	0.970	0.978	0.979	0.965
Recursive rewrite (R3)	0.979	0.979	0.970	0.977	0.978	0.962
Reviewer-guided rewrite	0.977	0.978	0.962	0.977	0.978	0.966
Cross-paper baseline	0.616	0.617	0.516	0.619	0.620	0.518

Table 16: **Condition-level science-core similarity for Gemini-3.5-Flash-Lite.** P05 denotes the fifth percentile across comparisons. The cross-paper baseline compares variant cores with non-matching original cores.

Rhetorical condition	GPT-5.5 producer			Opus 4.8 producer		
	Mean	Median	P05	Mean	Median	P05
Novelty stance	0.963	0.967	0.939	0.965	0.965	0.940
Scope framing	0.961	0.965	0.927	0.964	0.967	0.944
Evidence framing	0.958	0.961	0.930	0.961	0.963	0.935
Contribution salience	0.959	0.964	0.934	0.961	0.963	0.930
Technical register	0.962	0.965	0.940	0.961	0.964	0.935
Linguistic complexity	0.964	0.969	0.934	0.963	0.965	0.939
Joint rewrite	0.964	0.967	0.930	0.959	0.963	0.937
Recursive rewrite (R2)	0.964	0.966	0.933	0.959	0.963	0.928
Recursive rewrite (R3)	0.966	0.968	0.937	0.961	0.960	0.942
Reviewer-guided rewrite	0.963	0.965	0.937	0.959	0.961	0.930
Cross-paper baseline	0.606	0.606	0.506	0.609	0.608	0.509

Table 17: **Complete science-core weight sensitivity with Manuscript-Standard.** All entries are unnormalized numerical results. Arrows indicate the preferred direction.

α	Within-paper stability		Joint stability-discrimination			Human alignment	
	MAD ↓	Drift SD ↓	ICC ↑	SPR ↑	Discrim. ↑	H-MAE ↓	Spearman ↑
0.00	0.598	0.956	0.615	0.555	0.649	1.294	0.448
0.05	0.577	0.911	0.633	0.566	0.699	1.294	0.441
0.10	0.556	0.867	0.652	0.577	0.699	1.293	0.441
0.15	0.534	0.824	0.671	0.590	0.699	1.292	0.441
0.20	0.513	0.784	0.690	0.603	0.701	1.291	0.441
0.25	0.492	0.745	0.708	0.618	0.705	1.290	0.434
0.30	0.470	0.708	0.726	0.633	0.711	1.293	0.425
0.35	0.450	0.675	0.743	0.648	0.716	1.296	0.406
0.40	0.430	0.644	0.758	0.664	0.717	1.300	0.406
0.45	0.410	0.617	0.772	0.679	0.719	1.304	0.406
0.50	0.390	0.594	0.783	0.693	0.723	1.308	0.396
0.55	0.380	0.576	0.791	0.705	0.739	1.321	0.375
0.60	0.369	0.562	0.797	0.716	0.738	1.333	0.369
0.65	0.359	0.555	0.800	0.724	0.738	1.346	0.369
0.70	0.350	0.552	0.800	0.730	0.742	1.358	0.369
0.75	0.340	0.556	0.797	0.732	0.739	1.371	0.369
0.80	0.330	0.565	0.792	0.732	0.737	1.387	0.369
0.85	0.321	0.579	0.784	0.729	0.736	1.404	0.369
0.90	0.311	0.599	0.775	0.724	0.735	1.423	0.369
0.95	0.301	0.622	0.763	0.718	0.735	1.442	0.369
1.00	0.292	0.650	0.751	0.710	0.695	1.461	0.285

Table 18: **Complete science-core weight sensitivity with Manuscript-Strict.** All entries are unnormalized numerical results. Arrows indicate the preferred direction.

α	Within-paper stability		Joint stability-discrimination			Human alignment	
	MAD ↓	Drift SD ↓	ICC ↑	SPR ↑	Discrim. ↑	H-MAE ↓	Spearman ↑
0.00	0.766	1.202	0.632	0.507	0.672	1.078	0.529
0.05	0.736	1.143	0.645	0.516	0.712	1.061	0.513
0.10	0.707	1.085	0.659	0.526	0.712	1.044	0.513
0.15	0.677	1.028	0.674	0.538	0.712	1.032	0.513
0.20	0.648	0.972	0.689	0.550	0.713	1.022	0.513
0.25	0.619	0.918	0.704	0.565	0.715	1.014	0.510
0.30	0.589	0.866	0.720	0.580	0.719	1.016	0.504
0.35	0.560	0.817	0.735	0.597	0.722	1.022	0.497
0.40	0.532	0.770	0.749	0.615	0.723	1.039	0.497
0.45	0.504	0.727	0.763	0.633	0.725	1.056	0.497
0.50	0.476	0.687	0.775	0.652	0.726	1.072	0.488
0.55	0.456	0.653	0.786	0.671	0.735	1.107	0.474
0.60	0.435	0.623	0.794	0.688	0.734	1.142	0.451
0.65	0.414	0.600	0.800	0.703	0.733	1.177	0.451
0.70	0.396	0.585	0.802	0.716	0.745	1.212	0.415
0.75	0.378	0.576	0.801	0.724	0.741	1.247	0.406
0.80	0.361	0.576	0.797	0.729	0.742	1.286	0.397
0.85	0.344	0.584	0.790	0.729	0.743	1.326	0.397
0.90	0.326	0.599	0.779	0.726	0.742	1.371	0.397
0.95	0.309	0.621	0.766	0.719	0.742	1.416	0.397
1.00	0.292	0.650	0.751	0.710	0.695	1.461	0.285

Table 19: **Complete science-core weight sensitivity with Manuscript-Persistent.** All entries are unnormalized numerical results. Arrows indicate the preferred direction.

α	Within-paper stability		Joint stability-discrimination			Human alignment	
	MAD ↓	Drift SD ↓	ICC ↑	SPR ↑	Discrim. ↑	H-MAE ↓	Spearman ↑
0.00	0.476	0.875	0.697	0.586	0.682	1.261	0.423
0.05	0.461	0.830	0.712	0.601	0.727	1.269	0.408
0.10	0.447	0.787	0.726	0.617	0.727	1.276	0.408
0.15	0.432	0.746	0.741	0.634	0.727	1.284	0.408
0.20	0.418	0.707	0.755	0.652	0.729	1.291	0.408
0.25	0.403	0.670	0.769	0.669	0.732	1.299	0.408
0.30	0.389	0.636	0.782	0.687	0.736	1.306	0.408
0.35	0.374	0.604	0.794	0.704	0.736	1.314	0.408
0.40	0.360	0.577	0.804	0.720	0.738	1.321	0.408
0.45	0.347	0.554	0.812	0.734	0.739	1.329	0.408
0.50	0.333	0.535	0.819	0.747	0.738	1.336	0.402
0.55	0.329	0.522	0.823	0.756	0.752	1.349	0.378
0.60	0.325	0.514	0.825	0.763	0.751	1.361	0.378
0.65	0.321	0.512	0.824	0.766	0.751	1.374	0.378
0.70	0.316	0.517	0.820	0.766	0.754	1.386	0.378
0.75	0.312	0.527	0.814	0.762	0.750	1.399	0.378
0.80	0.308	0.542	0.806	0.756	0.747	1.411	0.378
0.85	0.304	0.563	0.795	0.747	0.745	1.424	0.378
0.90	0.300	0.588	0.782	0.736	0.744	1.436	0.378
0.95	0.296	0.618	0.767	0.723	0.744	1.449	0.378
1.00	0.292	0.650	0.751	0.710	0.695	1.461	0.285

Table 20: **API expenditure by experimental task.** Review-only costs are aggregated by billing route rather than by model. Costs are the recorded expenditures from the provider API consoles.

Task	Cost (USD)
RobustReview construction	$6,688.24
Review-only (OpenAI API)	$3,767.59
Review-only (OpenRouter API)	$870.60
SciCore evaluation	$3,043.48
ReconstructReview	$8,193.82
AI Scientist	$391.55
OpenJudge	$620.24
Total API expenditure	**$23,575.52**

G COMPUTE, COST, AND EFFICIENCY

We report API expenditure by experimental task. RobustReview construction covers the full-manuscript rewrites and reviewer-guided feedback used to construct the controlled corpus. Review-only covers the general-purpose reviewer grid and the repeated-review audit. The SciCore task includes science-core extraction, the branch-policy evaluations, and embedding; its manuscript branch reuses the GPT-5.5 STRICT reviews counted under Review-only, and score fusion introduces no additional API call. ReconstructReview includes manuscript reconstruction and final review while reusing cached science cores. The AI Scientist and OpenJudge entries cover their released agentic review workflows. The task-level costs in Table 20 are the recorded expenditures from the provider API consoles and include billable retries and failed attempts.

OpenReviewer, CycleReviewer, DeepReviewer, and the trained ProReviewer backbone are excluded from the API expenditure total because they were served locally on a server equipped with 8 NVIDIA A100 GPUs.

H PROMPTS AND ARTIFACT DETAILS

H.1 SCIENCE-CORE EXTRACTION PROMPT

Prompt 1: Science-Core Extraction

```
# Task: extract a self-contained scientific content record from a paper PDF

Read the complete paper PDF, including appendices and supplementary material
contained in the file, and produce one rigorous, self-contained Markdown
record of its scientific content.

This is an evidence-preserving extraction task, not a peer review, acceptance
recommendation, manuscript rewrite, or related-work survey. Extract what the
paper reports without evaluating its quality, importance, novelty, clarity, or
persuasiveness. Do not omit scientific information merely to make the record
shorter.

## Evidence and attribution rules

- Preserve the problem formulation, inputs and outputs, notation, objectives,
  algorithms, components, training and inference procedures, assumptions,
  derivations, and formal results that materially define or support the method.
- Preserve exact reported numbers, units, uncertainty, metric direction,
  comparison conditions, and negative or mixed findings. Do not estimate
  numerical values from plot geometry.
- Attribute interpretations and claims to the paper. For example, write
  'The authors attribute this difference to ...' rather than presenting that
  explanation as an independently established fact.
- Include section, equation, theorem, table, figure, or appendix identifiers as
  evidence anchors when available. Do not include PDF page numbers.
- Do not infer strengths, weaknesses, novelty, scientific importance, causal
  conclusions, limitations, or missing information beyond what the PDF states.
- Do not use evaluative adjectives unless they are part of an attributed claim
  or have a precise quantitative basis.
- Do not manufacture citations, experiments, implementation details, or
  numerical values.

## Required organization

Use exactly the following five level-two sections, with the headings written
exactly as shown. Do not add a descriptive title, executive summary,
contribution list, experimental map, separate conclusion,
scientific-positioning section, evidence-reference index, notes on
interpretation, or any other section before, between, or after them.

## Problem and setting

Record the research problem, task definition, inputs and outputs, setting,
motivation needed to understand the problem, and the paper's stated research
questions or claims. Keep author claims explicitly attributed.
```

Prompt 1: Science-Core Extraction (continued)

```
## Method and assumptions

Record the complete method and its technical foundations: notation, objectives,
model or system components, algorithms, training and inference, assumptions,
and method-critical derivations. For every theorem, proposition, lemma,
corollary, or formal guarantee that materially supports the method, state its
conditions, scope, result, and the proof logic available in the PDF.

## Experiments and reported results

Create one level-three subsection for each distinct experiment or coherent
experiment family. In each subsection, keep together:

- the stated question or purpose;
- datasets, sample sizes, splits, models, baselines, changed and controlled
  variables, metrics, and protocol;
- all material quantitative and qualitative results, including uncertainty,
  ablations, sensitivity and robustness studies, efficiency or scaling
  results, diagnostics, failure analyses, and negative or mixed findings;
- the interpretation stated by the authors, clearly attributed.

Represent every reported result needed to understand the paper's scientific
claims and comparisons.

### Complete extraction of reported tables

Extract every table in the main paper and appendix that contains scientific,
methodological, empirical, dataset, ablation, hyperparameter, complexity, or
reproducibility information. Reproduce each such table as a Markdown table
inside the relevant experiment or method subsection. Table extraction is
completeness-first and is not a place to shorten the record.

For every extracted table:

- include its table number and caption or a faithful descriptive caption;
- preserve every scientifically meaningful row, column, cell, header level,
  row group, column group, unit, metric direction, uncertainty value, range,
  footnote marker, missing-value marker, and table note;
- preserve the mapping between each value and its row and column conditions;
- retain author-provided emphasis such as bold or underline only when it
  encodes a defined meaning, and state that meaning in the table note;
- do not replace the table with selected values, prose, a bullet list, a range,
  an average, a summary, or only the best-performing entries;
- do not omit cells because they appear repetitive or less important;
```

Prompt 1: Science-Core Extraction (continued)

```
- if a table is too wide or has multi-level headers, split it into clearly
  labeled Markdown sub-tables while preserving all cells and their original
  grouping and order;
- if a cell cannot be read reliably, write `[unreadable]` in that cell rather
  than guessing or dropping it.

Do not repeat all table values in surrounding prose. Use prose only for setup,
qualifications, figure-only evidence, and explicit author interpretations that
are not already contained in the table. State each non-table scientific fact
once. If a figure contains unique evidence, describe that evidence in the
relevant experiment subsection without inventing values.

## Reproducibility information

Record reported datasets and versions, preprocessing, splits, sample sizes,
architectures, objectives, optimization details, hyperparameters, stopping and
model-selection rules, seeds and repeats, inference settings, baselines, metric
definitions and directions, software, and hardware. Avoid repeating details
already recorded with a particular experiment; cross-reference that subsection
instead.

For a reproducibility field that is relevant but absent or uncertain, use only
one of these neutral statuses:

- `Reported`: the PDF provides the information;
- `Not reported in the PDF`: the PDF does not provide it;
- `Not located`: the information may be present but cannot be reliably found;
- `Not applicable`: the field does not apply to this work.

Do not call absent information critical, important, problematic, or a
weakness. Do not create a checklist of irrelevant fields.

## Author-stated limitations

Include only limitations, boundary conditions, failure modes, or future-work
needs that the authors explicitly state or directly demonstrate. Attribute
them to the authors or to the specific reported evidence. If none are stated,
write `No explicit author-stated limitations were located in the PDF.` Do not
add inferred criticisms. This must be the final section of the record. End the
record when this section is complete; do not append references to evidence,
notes on interpretation, extraction notes, conclusions, or other commentary.

## Writing rules

- Use neutral, dense, precise technical language and consistent terminology.
```

Prompt 1: Science-Core Extraction (continued)

```
- Preserve qualifications and distinguish reported facts from attributed
  explanations.
- Include an interpretation only when the paper explicitly states it, place it
  beside the relevant evidence, and attribute it to the authors. Never add a
  model-generated interpretation or a separate interpretation summary.
- Avoid repeating a fact across overview, detail, table, figure, and conclusion
  formats. Removing repetition must never remove distinct conditions, values,
  qualifications, or evidence.
- Resolve notation into readable Markdown where possible while retaining the
  formulas and proof dependencies needed for scientific completeness.
- Do not include author names, affiliations, acknowledgements, paper IDs,
  arXiv IDs, filenames, repository names, or other source-identifying tokens.
- Do not discuss the extraction process, the prompt, or PDF parsing.
- Output only the finished Markdown scientific content record, without a code
  fence, preamble, TODO list, or commentary.
```

H.2 Adapted Review Prompt for the Science-Core Branch

Prompt 2: Adapted Science-Core Branch Review

```
# Scientific-core adapted standard review prompt

Please serve as an expert scientific reviewer for a top-tier AI conference
with an acceptance rate of about 30%.

Evaluate only the supplied scientific-content record. It is a structured
representation of the paper's scientific core, not the full paper text. It is
the complete evidence available for this review. Do not assume access to the
PDF or infer evidence outside the record.

Judge only scientific merit: problem importance, technical ideas and
correctness, theoretical support, experimental design, baselines, ablations,
claim-evidence fit, reproducibility, novelty, positioning, significance,
generality, limitations, and likely impact.

Do not reward or penalize prose polish, rhetoric, formatting, visual design,
author identity, institutional prestige, or resemblance to a conference
paper. Absence of non-scientific parts of the original paper is irrelevant.

## Evidence-status interpretation

- `Reported` means the evidence is available for assessment.
- `Not reported in the PDF` may count as missing paper evidence when that
  information is necessary to support a claim or reproduce the work.
- `Not located`, `Unreadable`, and `[unreadable]` indicate extraction
  uncertainty. Do not treat them as proof that the paper omitted the evidence.
  Reduce confidence when the uncertainty is material; do not automatically
  lower the rating or apply a rating cap solely because of extraction
  uncertainty.
- Do not give positive credit for evidence that is absent, vague, or merely
  asserted.

## Adapted decision procedure

Before choosing scores, assess the following evidence dimensions separately:

1. whether the central claims are technically supported;
2. whether experiments, comparisons, ablations, or theory test the claims;
3. whether the claimed contribution is differentiated from prior work;
4. whether the contribution is significant and general enough for the venue;
5. whether the reported information permits meaningful assessment and
   reproduction;
```

Prompt 2: Adapted Science-Core Branch Review (continued)

```
6. whether limitations or counterevidence materially change the claims.

Then map that assessment to the score rubrics below. Use the full score range
when the record supports it. Standard reviewing is neither uniformly lenient
nor uniformly harsh. Do not default to rating 5, do not compress distinct
papers into the same category, and do not use an overall score to compensate
for absent evidence on a central claim.

## Review Form

### 1. Summary

Briefly summarize the problem, method, and supported scientific contributions.
The authors should generally agree with a well-written summary.

### 2. Soundness, Presentation, Contribution

Use a 1-4 scale:

- 1 = poor
- 2 = fair
- 3 = good
- 4 = excellent

Soundness:

- 1 = Central claims are unsupported, technically flawed, or contradicted by
  the available evidence.
- 2 = The work is plausible but has important technical gaps, missing
  justification, weak validation, or limitations affecting central claims.
- 3 = The work is mostly technically sound, with moderate or non-central
  concerns.
- 4 = The work is technically rigorous, well justified, and thoroughly
  validated.

Presentation:

- Score scientific-specification completeness only: whether the problem,
  method, assumptions, evidence, and reproducibility information are clear
  enough to assess.
- Do not score prose style, formatting, original organization, figures as
  visual objects, or rhetorical polish.
```

Prompt 2: Adapted Science-Core Branch Review (continued)

```
- 1 = Essential scientific information is unavailable or too unclear to
  assess.
- 2 = The scientific specification is understandable but has important
  clarity or completeness gaps.
- 3 = The scientific specification is clear overall, with moderate gaps.
- 4 = The scientific specification is exceptionally clear, complete, and
  reproducible.

Contribution:

- 1 = The scientific contribution is minor, already known, or of limited
  relevance.
- 2 = It is somewhat useful but limited in novelty, significance, generality,
  empirical support, or likely impact.
- 3 = It is clear, meaningful, differentiated, and relevant to the field.
- 4 = It is strong, novel, well supported, and likely to have substantial
  impact.

### 3. Strengths

List the strongest scientifically supported aspects.

### 4. Weaknesses

List the most important scientific weaknesses, unsupported claims, missing
comparisons, evidence gaps, or reproducibility limitations. Distinguish paper
deficiencies from extraction uncertainty.

### 5. Questions

List questions whose answers would materially change or clarify the scientific
assessment.

### 6. Flag For Ethics Review

If the work raises ethical issues, describe them. Otherwise state:
'No ethics review needed.'

### 7. Rating

Use one of these values only: 1, 3, 5, 6, 8, or 10. The rating is not an
average of the three subscores.
```

Prompt 2: Adapted Science-Core Branch Review (continued)

```
- 1 = Strong Reject
  Fundamental scientific failure: a trivial result, serious technical flaw,
  unsupported central claim, invalid evaluation, or unaddressed ethical issue.
- 3 = Reject
  Meaningful weaknesses in novelty, soundness, evidence, reproducibility,
  significance, or positioning make the work uncompetitive. A coherent or
  plausible method may still receive 3.
- 5 = Marginally below the acceptance threshold
  Mostly technically solid with positive qualities, but scientific weaknesses
  slightly outweigh strengths. Use only when genuinely close to competitive.
- 6 = Marginally above the acceptance threshold
  Scientific strengths clearly outweigh weaknesses, with a nontrivial
  contribution, convincing evidence, appropriate comparisons, and adequate
  support for the main claims. Do not assign 6 merely for coherence.
- 8 = Accept
  A strong, meaningful, well-supported contribution with convincing technical
  quality, good-to-excellent evaluation, appropriate comparisons, sufficient
  reproducibility, and no major unresolved issue.
- 10 = Strong Accept
  An exceptional, groundbreaking contribution with technical excellence,
  exceptionally strong evidence, strong reproducibility, and no unaddressed
  ethical concern. Reserve for rare venue-leading work.

### 8. Confidence

Confidence measures certainty in this review, not paper quality. Use the full
1-5 scale:

- 1 = Central evidence cannot be assessed; the judgment is an educated guess.
- 2 = Material evidence or domain understanding is uncertain; the judgment is
  defensible but fragile.
- 3 = Most central evidence is assessable, with meaningful unresolved
  uncertainty.
- 4 = Central claims and evidence were checked carefully; only limited
  uncertainty remains.
- 5 = Exceptional evidence coverage and domain certainty; no material
  extraction uncertainty or unresolved assessment issue remains.

Do not default to 4. Any material `Not located`, `Unreadable`, or unresolved
central ambiguity normally limits confidence to 3 or below.
```

Prompt 2: Adapted Science-Core Branch Review (continued)

```
## Output Format

Output one valid JSON object with exactly the fields below. Do not output
Markdown fences, a preamble, internal checklist, or commentary. Use numeric
JSON values, not strings, for all score fields.

{
  "summary": "<text>",
  "soundness": <num>,
  "presentation": <num>,
  "contribution": <num>,
  "strengths": ["<text>", "<text>"],
  "weaknesses": ["<text>", "<text>"],
  "questions": ["<text>"],
  "ethics_flag": "No ethics review needed.",
  "rating": <num>,
  "confidence": <num>
}
```

H.3 SCIENCE-CORE BRANCH REVIEW-POLICY ABLATION PROMPTS

H.3.1 STANDARD

Prompt 3: Standard Science-Core Review

```
# Scientific-content-only standard review prompt

Please serve as an expert scientific reviewer for a top-tier AI conference
with an acceptance rate of about 30%.

Evaluate the submission only from the scientific-content record supplied after
this prompt. Treat that record as the complete available research content. Do
not assume access to the original PDF and do not infer missing evidence.

The overall rating must depend only on scientific merit:

- problem formulation and importance;
- technical ideas, correctness, and theoretical support;
- experimental design, baselines, ablations, and analysis;
- whether the evidence supports the main claims;
- reproducibility-relevant information;
- scientifically meaningful limitations and negative findings;
- novelty, significance, generality, and likely impact.

Do not reward or penalize writing polish, persuasive rhetoric, confident
framing, formatting, visual design, author identity, institutional prestige,
or resemblance to a conference paper. Do not mention the extraction process
or speculate about information outside the supplied record.

## Review Form

### 1. Summary

Briefly summarize the research problem, method, and supported scientific
contributions. This is not the place to critique the work; the authors should
generally agree with a well-written summary.

### 2. Soundness, Presentation, Contribution

Use a 1-4 scale:

- 1 = poor
- 2 = fair
- 3 = good
- 4 = excellent

Soundness:

- 1 = The main scientific claims are unsupported, technically flawed, or
  contradicted by the available evidence.
```

Prompt 3: Standard Science-Core Review (continued)

```
- 2 = The work is plausible but has important technical gaps, missing
  justification, weak validation, or limitations that affect the main claims.
- 3 = The work is mostly technically sound, with only moderate or non-central
  concerns.
- 4 = The work is technically rigorous, well justified, and thoroughly
  validated.

Presentation:

- Score only the clarity and completeness of the scientific specification:
  whether the problem, method, assumptions, evidence, and reproducibility
  details are sufficiently clear to assess.
- Do not score prose style, formatting, organization inherited from the
  original paper, figures as visual objects, or rhetorical polish.
- 1 = Essential scientific information is unclear or unavailable.
- 2 = The scientific content is understandable but has important clarity or
  completeness gaps.
- 3 = The scientific content is clear overall, with only moderate gaps.
- 4 = The scientific content is exceptionally clear, complete, and
  reproducible.

Contribution:

- 1 = The scientific contribution is minor, mostly incremental, already known,
  or of limited relevance.
- 2 = The contribution is somewhat useful but limited in novelty,
  significance, generality, empirical support, or likely impact.
- 3 = The contribution is clear, meaningful, and relevant to the field.
- 4 = The contribution is strong, novel, well supported, and likely to have
  substantial impact.

### 3. Strengths

List the strongest aspects of the scientific content.

### 4. Weaknesses

List the most important scientific weaknesses, unsupported claims, missing
comparisons, evidence gaps, or reproducibility limitations.

### 5. Questions

List questions whose answers would materially change or clarify the scientific
assessment.
```

Prompt 3: Standard Science-Core Review (continued)

```
### 6. Flag For Ethics Review

If the scientific work raises ethical issues, describe them. Otherwise state:
'No ethics review needed.'

### 7. Rating

Provide an overall score using one of these values only: 1, 3, 5, 6, 8, or 10.

The rating is not an average of soundness, presentation, and contribution. It
must be based only on the concrete scientific content and evidence available
in the record. Presentation may affect the rating only when missing or unclear
scientific information prevents reliable assessment or reproduction.

- 1 = Strong Reject
  Fundamental scientific problems such as a trivial result, serious technical
  flaw, unsupported central claim, invalid evaluation, or unaddressed ethical
  concern.
- 3 = Reject
  Not competitive for a top-tier AI conference because of meaningful
  weaknesses in novelty, technical soundness, evaluation, reproducibility,
  significance, or scientific positioning. A coherent or plausible method may
  still receive 3 when its contribution or evidence is insufficient.
- 5 = Marginally below the acceptance threshold
  Mostly technically solid with clear positive qualities, but scientific
  weaknesses slightly outweigh strengths. Typical reasons include limited
  novelty, incomplete evaluation, weak baselines, insufficient ablations,
  narrow scope, limited reproducibility, unclear significance, or claims
  stronger than the evidence. Use sparingly.
- 6 = Marginally above the acceptance threshold
  Technically solid, with scientific strengths clearly outweighing weaknesses
  and concrete evidence of a nontrivial contribution, convincing evaluation,
  appropriate comparisons, and adequate support for the main claims. Do not
  assign 6 merely because the work is coherent or plausible. Use sparingly.
- 8 = Accept
  A strong, meaningful, well-supported scientific contribution with convincing
  technical quality, good-to-excellent evaluation, appropriate comparisons,
  sufficient reproducibility, and clear relevance to the field. There should
  be no major unresolved issue in soundness, novelty, evaluation,
  reproducibility, significance, or ethics.
- 10 = Strong Accept
  An exceptional scientific contribution with groundbreaking impact,
  technical excellence, exceptionally strong evidence, strong
  reproducibility, and no unaddressed ethical concern. Reserve this for rare
  submissions clearly among the venue's best.
```

Prompt 3: Standard Science-Core Review (continued)

```
### 8. Confidence

Use a 1-5 scale:

- 1 = educated guess
- 2 = willing to defend but likely missed central parts
- 3 = fairly confident
- 4 = confident but not absolutely certain
- 5 = absolutely certain and checked the available details carefully

## Output Format

Output one valid JSON object with exactly the fields below. Do not output
Markdown fences, a preamble, or commentary. Use numeric JSON values, not
strings, for `soundness`, `presentation`, `contribution`, `rating`, and
`confidence`.

{
  "summary": "<text>",
  "soundness": <num>,
  "presentation": <num>,
  "contribution": <num>,
  "strengths": ["<text>", "<text>"],
  "weaknesses": ["<text>", "<text>"],
  "questions": ["<text>"],
  "ethics_flag": "No ethics review needed.",
  "rating": <num>,
  "confidence": <num>
}
```

H.3.2 STRICT

Prompt 4: Strict Science-Core Review

```
# Scientific-content-only strict review prompt

Please serve as an expert scientific reviewer for a top-tier AI conference
with an acceptance rate of about 30%.

Evaluate the submission only from the scientific-content record supplied after
this prompt. Treat that record as the complete available research content. Do
not assume access to the original PDF and do not infer missing evidence.

The supplied record is a structured representation of the paper's scientific
core, not the full paper text and not the paper as a complete document. It
retains the research problem, method, technical foundations, experiments,
results, scientific claims, limitations, and reproducibility-relevant details
available for evaluation, while non-scientific prose, rhetoric, formatting,
and visual presentation may be absent by design. Evaluate this scientific core
as the review object. Do not penalize the absence of non-scientific parts of
the original paper.

The overall rating must depend only on scientific merit:

- problem formulation and importance;
- technical ideas, correctness, and theoretical support;
- experimental design, baselines, ablations, and analysis;
- whether the evidence supports the main claims;
- reproducibility-relevant information;
- scientifically meaningful limitations and negative findings;
- novelty, significance, generality, and likely impact.

Do not reward or penalize writing polish, persuasive rhetoric, confident
framing, formatting, visual design, author identity, institutional prestige,
or resemblance to a conference paper. Do not mention the extraction process
or speculate about information outside the supplied record.

## Strict Evaluation Policy

Apply a strict acceptance threshold while assessing the same scientific
content and the same review dimensions as the standard protocol:

- Judge the work against accepted submissions at a top-tier venue, not against
  a merely coherent or plausible submission.
- Missing, vague, or implicit evidence does not establish a claim. Give credit
  only for support that is present in the scientific-content record.
```

Prompt 4: Strict Science-Core Review (continued)

```
- A technically plausible method may still be reject-level when novelty,
  validation, comparisons, reproducibility, significance, or positioning is
  insufficient.
- Be conservative, but not adversarial. Identify a weakness only when it is
  supported by the record or when information necessary to assess a stated
  claim is absent.

Strictness changes the evidentiary threshold for the overall decision; it does
not change the scientific-content-only boundary, the meanings of the
subscores, or the output schema.

## Review Form

### 1. Summary

Briefly summarize the research problem, method, and supported scientific
contributions. This is not the place to critique the work; the authors should
generally agree with a well-written summary.

### 2. Soundness, Presentation, Contribution

Use a 1-4 scale:

- 1 = poor
- 2 = fair
- 3 = good
- 4 = excellent

Soundness:

- 1 = The main scientific claims are unsupported, technically flawed, or
  contradicted by the available evidence.
- 2 = The work is plausible but has important technical gaps, missing
  justification, weak validation, or limitations that affect the main claims.
- 3 = The work is mostly technically sound, with only moderate or non-central
  concerns.
- 4 = The work is technically rigorous, well justified, and thoroughly
  validated.

Presentation:

- Score only the clarity and completeness of the scientific specification:
```

Prompt 4: Strict Science-Core Review (continued)

```
  whether the problem, method, assumptions, evidence, and reproducibility
  details are sufficiently clear to assess.
- Do not score prose style, formatting, organization inherited from the
  original paper, figures as visual objects, or rhetorical polish.
- 1 = Essential scientific information is unclear or unavailable.
- 2 = The scientific content is understandable but has important clarity or
  completeness gaps.
- 3 = The scientific content is clear overall, with only moderate gaps.
- 4 = The scientific content is exceptionally clear, complete, and
  reproducible.

Contribution:

- 1 = The scientific contribution is minor, mostly incremental, already known,
  or of limited relevance.
- 2 = The contribution is somewhat useful but limited in novelty,
  significance, generality, empirical support, or likely impact.
- 3 = The contribution is clear, meaningful, and relevant to the field.
- 4 = The contribution is strong, novel, well supported, and likely to have
  substantial impact.

### 3. Strengths

List the strongest aspects of the scientific content.

### 4. Weaknesses

List the most important scientific weaknesses, unsupported claims, missing
comparisons, evidence gaps, or reproducibility limitations.

### 5. Questions

List questions whose answers would materially change or clarify the scientific
assessment.

### 6. Flag For Ethics Review

If the scientific work raises ethical issues, describe them. Otherwise state:
'No ethics review needed.'

### 7. Rating
```

Prompt 4: Strict Science-Core Review (continued)

```
Provide an overall score using one of these values only: 1, 3, 5, 6, 8, or 10.

The rating is not an average of soundness, presentation, and contribution. It
must be based only on the concrete scientific content and evidence available
in the record. Presentation may affect the rating only when missing or unclear
scientific information prevents reliable assessment or reproduction.

Use this strict decision procedure:

- Start from 3 = Reject. Raise the rating only when the record contains
  concrete evidence satisfying the higher category. Do not start from 6.
- Lower the rating to 1 only for a fundamental scientific or ethical problem.
- When uncertain between adjacent ratings, choose the lower rating unless the
  record contains concrete evidence for the higher one.

- 1 = Strong Reject
  Fundamental scientific problems such as a trivial result, serious technical
  flaw, unsupported central claim, invalid evaluation, or unaddressed ethical
  concern.
- 3 = Reject
  Not competitive for a top-tier AI conference because of meaningful
  weaknesses in novelty, technical soundness, evaluation, reproducibility,
  significance, or scientific positioning. A coherent or plausible method may
  still receive 3 when its contribution or evidence is insufficient.
- 5 = Marginally below the acceptance threshold
  Mostly technically solid with clear positive qualities, but scientific
  weaknesses slightly outweigh strengths. Typical reasons include limited
  novelty, incomplete evaluation, weak baselines, insufficient ablations,
  narrow scope, limited reproducibility, unclear significance, or claims
  stronger than the evidence. Use sparingly.
- 6 = Marginally above the acceptance threshold
  Technically solid, with scientific strengths clearly outweighing weaknesses
  and concrete evidence of a nontrivial contribution, convincing evaluation,
  appropriate comparisons, and adequate support for the main claims. Do not
  assign 6 merely because the work is coherent or plausible. Use sparingly.
- 8 = Accept
  A strong, meaningful, well-supported scientific contribution with convincing
  technical quality, good-to-excellent evaluation, appropriate comparisons,
  sufficient reproducibility, and clear relevance to the field. There should
  be no major unresolved issue in soundness, novelty, evaluation,
  reproducibility, significance, or ethics.
- 10 = Strong Accept
```

Prompt 4: Strict Science-Core Review (continued)

```
  An exceptional scientific contribution with groundbreaking impact,
  technical excellence, exceptionally strong evidence, strong
  reproducibility, and no unaddressed ethical concern. Reserve this for rare
  submissions clearly among the venue's best.

Strict upper-bound rules:

- If the main claims are not adequately supported, the rating must be at most 3.
- If important baselines, necessary ablations, or evaluation required for the
  main claims are missing, the rating should usually be at most 5.
- If the contribution is incremental, narrow, or weakly positioned relative to
  prior work, the rating should usually be at most 5.
- If there are multiple moderate weaknesses across novelty, evaluation,
  reproducibility, or significance, the rating should usually be at most 5.
- If soundness or contribution is 1, the overall rating should usually be 1 or
  3.
- If soundness or contribution is 2, the overall rating should usually be no
  higher than 5 unless unusually strong compensating scientific evidence is
  explicitly present.

### 8. Confidence

Use a 1-5 scale:

- 1 = educated guess
- 2 = willing to defend but likely missed central parts
- 3 = fairly confident
- 4 = confident but not absolutely certain
- 5 = absolutely certain and checked the available details carefully

## Output Format

Output one valid JSON object with exactly the fields below. Do not output
Markdown fences, a preamble, or commentary. Use numeric JSON values, not
strings, for `soundness`, `presentation`, `contribution`, `rating`, and
`confidence`.

{
  "summary": "<text>",
  "soundness": <num>,
  "presentation": <num>,
  "contribution": <num>,
```

Prompt 4: Strict Science-Core Review (continued)

```
  "strengths": ["<text>", "<text>"],
  "weaknesses": ["<text>", "<text>"],
  "questions": ["<text>"],
  "ethics_flag": "No ethics review needed.",
  "rating": <num>,
  "confidence": <num>
}
```

H.3.3 PERSISTENT

Prompt 5: Persistent Science-Core Review

```
# Scientific-content-only persistent review prompt

Please serve as an expert scientific reviewer for a top-tier AI conference
with an acceptance rate of about 30%.

Evaluate the submission only from the scientific-content record supplied after
this prompt. The record is a structured representation of the paper's
scientific core, not the full paper text and not the paper as a complete
document. Treat it as the complete evidence available for this review. Do not
assume access to the original PDF and do not infer evidence outside the record.

The record retains the research problem, method, technical foundations,
experiments, results, scientific claims, limitations, and
reproducibility-relevant details available for evaluation. Non-scientific
prose, rhetoric, formatting, document organization, and visual presentation
may be absent by design. Evaluate the scientific core itself as the review
object. Do not penalize the absence of non-scientific parts of the original
paper.

Scientific content means the actual research contribution:

- problem formulation and importance;
- technical ideas, correctness, and theoretical support;
- experimental design, baselines, ablations, and analysis;
- whether the evidence supports the main claims;
- reproducibility-relevant information;
- scientifically meaningful limitations and negative findings;
- novelty, significance, generality, and likely impact.

Do not reward or penalize writing polish, persuasive rhetoric, confident
framing, formatting, visual design, author identity, institutional prestige,
or resemblance to a conference paper. Do not mention the extraction process
or speculate about information outside the supplied record.

## Persistent Scientific-Content Policy

Maintain the scientific-content-only boundary throughout every part of the
review and when selecting the final rating:

- Base every judgment on concrete scientific information present in the
  supplied record.
- Do not give credit for missing, implicit, vague, or weakly supported
```

Prompt 5: Persistent Science-Core Review (continued)

```
  evidence.
- Do not reward claims about novelty, impact, or rigor unless they are
  supported by the method, analysis, theory, or experiments in the record.
- Do not use prose quality, rhetorical confidence, formatting, original paper
  organization, or visual polish as a proxy for scientific quality.
- Score Presentation only as scientific-specification clarity and
  completeness: whether the problem, method, assumptions, evidence, and
  reproducibility information are sufficiently clear to assess.
- Missing or unclear information may affect the rating only when it prevents
  assessment, verification, or reproduction of a scientific claim.

Persistent means that these constraints must remain active throughout the
review. It does not mean that the reviewer should be systematically stricter
or more permissive than the Standard protocol. Apply the ordinary top-tier
acceptance threshold using scientific evidence alone.

## Review Form

### 1. Summary

Briefly summarize the research problem, method, and supported scientific
contributions. This is not the place to critique the work; the authors should
generally agree with a well-written summary.

### 2. Soundness, Presentation, Contribution

Use a 1-4 scale:

- 1 = poor
- 2 = fair
- 3 = good
- 4 = excellent

Soundness:

- 1 = The main scientific claims are unsupported, technically flawed, or
  contradicted by the available evidence.
- 2 = The work is scientifically plausible but has important technical gaps,
  missing justification, weak validation, missing necessary comparisons, or
  limitations that affect the main claims.
- 3 = The scientific content is mostly technically sound, with adequate
  justification and validation and only moderate or non-central concerns.
```

Prompt 5: Persistent Science-Core Review (continued)

```
- 4 = The scientific content is technically rigorous, well justified, and
  thoroughly validated.

Presentation:

- Score only the clarity and completeness of the scientific specification.
- Do not score prose style, formatting, original organization, figures as
  visual objects, or rhetorical polish.
- 1 = Essential scientific information is unclear or unavailable.
- 2 = The scientific content is understandable but has important clarity or
  completeness gaps.
- 3 = The scientific content is clear overall, with only moderate gaps.
- 4 = The scientific content is exceptionally clear, complete, and
  reproducible.

Contribution:

- 1 = The scientific contribution is minor, mostly incremental, already known,
  poorly motivated, or of limited relevance.
- 2 = The contribution is somewhat useful but limited in novelty,
  significance, generality, empirical support, or likely impact.
- 3 = The contribution is clear, meaningful, well positioned relative to prior
  work, and likely to be useful to the field.
- 4 = The contribution is strong, novel, well supported, and likely to have
  substantial impact on one or more sub-areas of AI.

### 3. Strengths

List the strongest aspects of the scientific content.

### 4. Weaknesses

List the most important scientific weaknesses, unsupported claims, missing
comparisons, evidence gaps, or reproducibility limitations.

### 5. Questions

List questions whose answers would materially change or clarify the scientific
assessment.

### 6. Flag For Ethics Review
```

Prompt 5: Persistent Science-Core Review (continued)

```
If the scientific work raises ethical issues, describe them. Otherwise state:
'No ethics review needed.'

### 7. Rating

Provide an overall score using one of these values only: 1, 3, 5, 6, 8, or 10.

The rating is not an average of soundness, presentation, and contribution. It
must be determined only by novelty, correctness, evidence, evaluation quality,
reproducibility, limitations, and significance in the scientific-content
record.

Good writing, confident framing, complete formatting, or polished motivation
must not raise the rating. Their absence must not lower the rating. Scientific
clarity or missing detail may affect the rating only when it prevents reliable
assessment or reproduction.

- 1 = Strong Reject
  Fundamental scientific problems such as a trivial result, serious technical
  flaw, unsupported central claim, invalid or misleading evaluation, severe
  reproducibility problems, or an unaddressed ethical concern.
- 3 = Reject
  Not competitive for a top-tier AI conference because of meaningful
  weaknesses in novelty, technical soundness, evaluation, reproducibility,
  significance, or scientific positioning. A coherent or plausible method may
  still receive 3 when its contribution or evidence is insufficient.
- 5 = Marginally below the acceptance threshold
  Mostly technically solid with clear scientific strengths, but scientific
  weaknesses slightly outweigh strengths. Typical reasons include limited
  novelty, incomplete evaluation, weak baselines, insufficient ablations,
  narrow scope, limited reproducibility, unclear significance, or claims
  stronger than the evidence. Use sparingly.
- 6 = Marginally above the acceptance threshold
  Technically solid, with scientific strengths clearly outweighing weaknesses
  and concrete evidence of a nontrivial contribution, convincing evaluation,
  appropriate comparisons, and adequate support for the main claims. Do not
  assign 6 merely because the record is coherent or plausible. Use sparingly.
- 8 = Accept
  A strong, meaningful, well-supported scientific contribution with convincing
  technical quality, good-to-excellent evaluation, appropriate comparisons,
  sufficient reproducibility, and clear relevance to the field. There should
  be no major unresolved issue.
```

Prompt 5: Persistent Science-Core Review (continued)

```
- 10 = Strong Accept
  An exceptional scientific contribution with groundbreaking impact,
  technical excellence, exceptionally strong evidence, strong
  reproducibility, and no unaddressed ethical concern. Reserve this for rare
  submissions clearly among the venue's best.

Scientific-content-only reviewing guidance:

- If the main scientific claims are not adequately supported, the rating
  should be at most 3.
- If important baselines, necessary ablations, or evaluation needed for the
  main claims are missing, the rating should usually be at most 5.
- If the scientific contribution is incremental, narrow, or weakly positioned
  relative to prior work, the rating should usually be at most 5.
- If soundness or contribution is 1, the overall rating should usually be 1 or
  3.
- If soundness or contribution is 2, the overall rating should usually be no
  higher than 5 unless unusually strong compensating scientific evidence is
  explicitly present.
- When uncertain between adjacent ratings, choose the rating better supported
  by concrete scientific evidence; never use presentation polish as the
  tiebreaker.

### 8. Confidence

Use a 1-5 scale:

- 1 = educated guess
- 2 = willing to defend but likely missed central parts
- 3 = fairly confident
- 4 = confident but not absolutely certain
- 5 = absolutely certain and checked the available details carefully

## Output Format

Output one valid JSON object with exactly the fields below. Do not output
Markdown fences, a preamble, or commentary. Use numeric JSON values, not
strings, for `soundness`, `presentation`, `contribution`, `rating`, and
`confidence`.

{
  "summary": "<text>",
```

Prompt 5: Persistent Science-Core Review (continued)

```
  "soundness": <num>,
  "presentation": <num>,
  "contribution": <num>,
  "strengths": ["<text>", "<text>"],
  "weaknesses": ["<text>", "<text>"],
  "questions": ["<text>"],
  "ethics_flag": "No ethics review needed.",
  "rating": <num>,
  "confidence": <num>
}
```

H.4 DIRECT-REVIEW BENCHMARK PROMPTS

The STANDARD direct-review protocol uses an ICLR-style review rubric, while STRICT applies more demanding evidence requirements. The PERSISTENT protocol retains the same scoring scheme and repeatedly directs the reviewer to base its judgments on scientific content. The complete prompts for all three protocols are reproduced below.

H.4.1 STANDARD

Prompt 6: Standard Direct-Review Prompt

```
Please serve as an expert reviewer for a top-tier AI conference with an acceptance rate of
     about 30%.
Please evaluate the paper critically and assign scores according to the criteria below.

Review Form

1. Summary
Briefly summarize the paper and its contributions.
This is not the place to critique the paper; the authors should generally agree with a well
     -written summary.

2. Soundness, Presentation, Contribution
Use a 1-4 scale:
1 = poor
2 = fair
3 = good
4 = excellent

Soundness:
1 = The main claims are unsupported, technically flawed, or contradicted by the evidence.
2 = The work is plausible but has important technical gaps, missing justification, weak
     validation, or limitations that affect the main claims.
3 = The work is mostly technically sound, with only moderate or non-central concerns.
4 = The work is technically rigorous, well justified, and thoroughly validated.

Presentation:
1 = The paper is difficult to understand, poorly organized, or missing essential details.
2 = The paper is understandable but has clarity, organization, or completeness issues.
3 = The paper is clearly written overall, with only moderate presentation issues.
4 = The paper is very clear, well organized, and easy to follow.

Contribution:
1 = The contribution is minor, mostly incremental, already known, or of limited relevance.
2 = The contribution is somewhat useful but limited in novelty, significance, generality,
     or impact.
3 = The contribution is clear, meaningful, and relevant to the field.
4 = The contribution is strong, novel, and likely to have substantial impact.

3. Strengths
List the strong points of the paper.

4. Weaknesses
List the weak points of the paper.

5. Questions
List questions for the authors that would help clarify your understanding.

6. Flag For Ethics Review
If there are ethical issues, describe them.
Otherwise state: "No ethics review needed."

7. Rating
Please provide an overall score for this submission.
The rating must be one of the following values only: 1, 3, 5, 6, 8, or 10.

Rating scale:
1 = Strong Reject
A paper with fundamental problems such as well-known or trivial results, serious technical
     flaws, unsupported central claims, invalid evaluation, or unaddressed ethical concerns
     .

3 = Reject
A paper that is not competitive for a top-tier AI conference due to meaningful weaknesses
     in novelty, technical soundness, evaluation, reproducibility, clarity, significance,
     or positioning.
```

Prompt 6: Standard Direct-Review Prompt (continued)

```
A paper can receive 3 even if it is understandable, technically plausible, or has some
     strengths, if the overall contribution or evidence is insufficient.

5 = Marginally below the acceptance threshold
A technically solid paper with some positive qualities, but where the reasons for a lower
     score slightly outweigh the reasons for a higher score.
Typical reasons include limited novelty, incomplete evaluation, weak or missing baselines,
     insufficient ablations, unclear significance, narrow scope, limited reproducibility,
     or claims that are stronger than the evidence.
Please use sparingly.

6 = Marginally above the acceptance threshold
A technically solid paper where the reasons for a higher score clearly outweigh the
     weaknesses.
A rating of 6 requires concrete positive evidence that the paper is competitive for a top-
     tier AI conference, such as a clear contribution, convincing evaluation, appropriate
     baselines, and sufficient support for the main claims.
Do not assign 6 merely because the paper is coherent, plausible, or has some positive
     results.
Please use sparingly.

8 = Accept
A strong paper with a clear and meaningful contribution, convincing technical quality, good
     -to-excellent evaluation, appropriate comparisons, sufficient reproducibility, and
     clear relevance to the field.
The paper should have high impact on at least one sub-area of AI or moderate-to-high impact
      on more than one area of AI.
Minor weaknesses are acceptable, but there should be no major unresolved issues in
     soundness, evaluation, novelty, or ethics.

10 = Strong Accept
An exceptional paper with groundbreaking impact on one or more areas of AI, technically
     excellent work, exceptionally strong evaluation, strong reproducibility and resources,
      and no unaddressed ethical considerations.
This rating should be reserved for rare submissions that are clearly among the best papers
     in the venue.

8. Confidence
Use a 1-5 scale:
1 = educated guess
2 = willing to defend but likely missed central parts
3 = fairly confident
4 = confident but not absolutely certain
5 = absolutely certain, checked details carefully

Output Format

You MUST output your review as a single valid JSON object with exactly the fields shown
     below.
Do NOT output any text outside the JSON object.
Do NOT use markdown fences, preamble, or commentary.
Use numeric values, not strings, for soundness, presentation, contribution, rating, and
     confidence.

{
  "summary": "<text>",
  "soundness": <num>,
  "presentation": <num>,
  "contribution": <num>,
  "strengths": ["<text>", "<text>"],
  "weaknesses": ["<text>", "<text>"],
  "questions": ["<text>"],
  "ethics_flag": "No ethics review needed.",
  "rating": <num>,
  "confidence": <num>
}
```

H.4.2 STRICT

Prompt 7: Strict Direct-Review Prompt

```
Please serve as a rigorous professional reviewer for an ICLR-style top-tier AI conference.
Apply strict, evidence-based reviewing standards. Judge the paper against the standards of
     accepted papers at such a venue, not against a merely coherent or polished submission.

Many technically plausible, well-written, or complete submissions should still receive
     reject-level ratings if their novelty, evidence, evaluation, reproducibility, or
     significance is insufficient.

Keep the review concise:
- Summary: 2-4 sentences.
- Strengths: 2-4 items.
- Weaknesses: 2-4 items.
- Questions: 1-3 items.
- Each list item should be brief and specific.

Review Form

1. Summary
Briefly summarize the paper and its contributions.
This is not the place to critique the paper; the authors should generally agree with a well
     -written summary.

2. Soundness, Presentation, Contribution
Use a 1-4 scale:
1 = poor
2 = fair
3 = good
4 = excellent

Soundness:
1 = The main claims are unsupported, technically flawed, or contradicted by the evidence.
2 = The work is plausible but has important technical gaps, missing justification, weak
     validation, missing necessary comparisons, or limitations that affect the main claims.
3 = The work is mostly technically sound, with adequate justification and validation, and
     only moderate or non-central concerns.
4 = The work is technically rigorous, well justified, and thoroughly validated.

Presentation:
1 = The paper is difficult to understand, poorly organized, or missing essential details.
2 = The paper is understandable but has clarity, organization, or completeness issues.
3 = The paper is clearly written overall, with only moderate presentation issues.
4 = The paper is very clear, well organized, and easy to follow.

Contribution:
1 = The contribution is minor, mostly incremental, already known, poorly motivated, or of
     limited relevance.
2 = The contribution is somewhat useful but limited in novelty, significance, generality,
     empirical support, or likely impact.
3 = The contribution is clear, meaningful, well positioned relative to prior work, and
     likely to be useful to the field.
4 = The contribution is strong, novel, well supported, and likely to have substantial
     impact on one or more sub-areas of AI.

3. Strengths
List the strong points of the paper.

4. Weaknesses
List the weak points of the paper.

5. Questions
List questions for the authors that would help clarify your understanding.

6. Flag For Ethics Review
If there are ethical issues, describe them.
Otherwise state: "No ethics review needed."

7. Rating
Please provide an overall score for this submission.
The rating must be one of the following values only: 1, 3, 5, 6, 8, or 10.

The overall rating is not an average of soundness, presentation, and contribution.
Good presentation should not compensate for weak novelty, weak evidence, weak evaluation,
     poor reproducibility, or limited significance.
```

Prompt 7: Strict Direct-Review Prompt (continued)

```
Improved clarity, organization, or rhetoric should mainly affect the presentation score,
     not the overall rating, unless the paper also provides stronger concrete evidence for
     novelty, soundness, evaluation, or significance.

Assign the rating based on concrete evidence in the paper.
Do not give credit for missing, implicit, vague, or weakly supported evidence.
Do not reward polished writing, confident framing, or strong claims unless they are
     supported by actual technical content and evaluation.

Rating procedure:
Start from 3 = Reject. Raise the rating only when the paper provides concrete evidence that
      it satisfies a higher rating category. Lower the rating to 1 when there are
     fundamental flaws. Do not start from 6.

Rating scale:

1 = Strong Reject
A paper with fundamental problems such as well-known or trivial results, serious technical
     flaws, unsupported central claims, invalid or misleading evaluation, severe
     reproducibility problems, or unaddressed ethical concerns.

3 = Reject
A paper that is not competitive for an ICLR-style top-tier AI conference due to meaningful
     weaknesses in novelty, technical soundness, evaluation, reproducibility, clarity,
     significance, or positioning.
Use 3 for papers with weak novelty, weak empirical support, missing important baselines,
     insufficient ablations, unclear contribution, overclaimed results, limited
     significance, or poor positioning relative to prior work.
A paper can receive 3 even if it is understandable, technically plausible, well written, or
      has some positive experimental results.

5 = Marginally below the acceptance threshold
A mostly technically solid paper with some clear strengths, but where weaknesses slightly
     outweigh strengths.
Use 5 only when the paper is close to being competitive but still falls short due to
     limited novelty, incomplete evaluation, weak baselines, insufficient ablations, narrow
      scope, limited reproducibility, unclear significance, or claims stronger than the
     evidence.
A paper with multiple meaningful weaknesses should receive 3 rather than 5.
Please use sparingly.

6 = Marginally above the acceptance threshold
A technically solid paper where strengths clearly outweigh weaknesses and there is concrete
      evidence that the paper is competitive for an ICLR-style top-tier AI conference.
Use 6 only if the paper has a clear nontrivial contribution, mostly sound methodology,
     convincing evaluation for the main claims, appropriate baselines or comparisons, and
     no major unresolved weakness in novelty, evaluation, reproducibility, or significance.
Do not assign 6 merely because the paper is coherent, plausible, well written, or reports
     some positive results.
If these conditions are not satisfied, the rating should usually be no higher than 5.
Please use sparingly.

8 = Accept
A strong paper with a clear, meaningful, and well-supported contribution; convincing
     technical quality; good-to-excellent evaluation; appropriate comparisons; sufficient
     reproducibility; and clear relevance to the field.
Use 8 only when the paper would be clearly competitive at an ICLR-style top-tier AI
     conference and has no major unresolved issue in soundness, novelty, evaluation,
     reproducibility, significance, or ethics.
A paper with limited evaluation, unclear novelty, weak baselines, insufficient ablations,
     or significant reproducibility gaps should not receive 8.

10 = Strong Accept
An exceptional paper with groundbreaking impact on one or more areas of AI, technically
     excellent work, exceptionally strong evaluation, strong reproducibility and resources,
      and no unaddressed ethical considerations.
This rating should be reserved for rare submissions that are clearly among the best papers
     in the venue.

Upper-bound rules:
If the main claims are not adequately supported, the rating should be at most 3.
If the paper lacks important baselines, necessary ablations, or evaluation needed to
     support the main claims, the rating should usually be at most 5.
If the contribution is incremental, narrow, or weakly positioned relative to prior work,
     the rating should usually be at most 5.
If the paper has multiple moderate weaknesses across novelty, evaluation, reproducibility,
     clarity, or significance, the rating should usually be at most 5.
```

Prompt 7: Strict Direct-Review Prompt (continued)

```
If either soundness or contribution is rated 1, the overall rating should usually be 1 or
     3.
If either soundness or contribution is rated 2, the overall rating should usually be no
     higher than 5 unless there is unusually strong compensating evidence.
When uncertain between two adjacent ratings, choose the lower rating unless the review
     identifies concrete evidence supporting the higher rating.

8. Confidence
Use a 1-5 scale:
1 = educated guess
2 = willing to defend but likely missed central parts
3 = fairly confident
4 = confident but not absolutely certain
5 = absolutely certain, checked details carefully

Output Format

You MUST output your review as a single valid JSON object with exactly the fields shown
     below.
Do NOT output any text outside the JSON object.
Do NOT use markdown fences, preamble, or commentary.
Use numeric values, not strings, for soundness, presentation, contribution, rating, and
     confidence.

{
  "summary": "<text>",
  "soundness": <num>,
  "presentation": <num>,
  "contribution": <num>,
  "strengths": ["<text>", "<text>"],
  "weaknesses": ["<text>", "<text>"],
  "questions": ["<text>"],
  "ethics_flag": "No ethics review needed.",
  "rating": <num>,
  "confidence": <num>
}
```

H.4.3 PERSISTENT

Prompt 8: Persistent Direct-Review Prompt

````
## Persistent Prompt

```text
Please serve as an expert scientific reviewer for a top-tier AI conference with an
     acceptance rate of about 30%.
Evaluate the paper critically, but base the overall score only on the paper's scientific
     content.

Scientific content means the actual research contribution: problem formulation, technical
     ideas, theoretical or empirical support, experimental design, baselines, ablations,
     analysis, reproducibility-relevant details, limitations, and significance to the field
     .

Do not let writing polish, persuasive rhetoric, confident framing, formatting quality,
     author identity, institutional prestige, or venue-style presentation affect the
     overall rating.
Presentation should still be scored in the presentation field, but it should affect the
     overall rating only when clarity or missing detail prevents reliable assessment or
     reproduction of the scientific content.

Review Form

1. Summary
Briefly summarize the paper and its scientific contributions.
This is not the place to critique the paper; the authors should generally agree with a well
     -written summary.

2. Soundness, Presentation, Contribution
Use a 1-4 scale:
1 = poor
2 = fair
3 = good
4 = excellent

Soundness:
1 = The main scientific claims are unsupported, technically flawed, or contradicted by the
     evidence.
2 = The work is scientifically plausible but has important technical gaps, missing
     justification, weak validation, missing necessary comparisons, or limitations that
     affect the main claims.
3 = The scientific content is mostly technically sound, with adequate justification and
     validation, and only moderate or non-central concerns.
4 = The scientific content is technically rigorous, well justified, and thoroughly
     validated.

Presentation:
1 = The paper is difficult to understand, poorly organized, or missing essential scientific
      details.
2 = The paper is understandable but has clarity, organization, or completeness issues.
3 = The paper is clearly written overall, with only moderate presentation issues.
4 = The paper is very clear, well organized, and easy to follow.

Contribution:
1 = The scientific contribution is minor, mostly incremental, already known, poorly
     motivated, or of limited relevance.
2 = The scientific contribution is somewhat useful but limited in novelty, significance,
     generality, empirical support, or likely impact.
3 = The scientific contribution is clear, meaningful, well positioned relative to prior
     work, and likely to be useful to the field.
4 = The scientific contribution is strong, novel, well supported, and likely to have
     substantial impact on one or more sub-areas of AI.

3. Strengths
List the strong points of the paper's scientific content.

4. Weaknesses
List the weak points of the paper's scientific content.

5. Questions
List questions for the authors that would help clarify your scientific assessment.

6. Flag For Ethics Review
If there are ethical issues, describe them.
Otherwise state: "No ethics review needed."
````
````

Prompt 8: Persistent Direct-Review Prompt (continued)

```
7. Rating
Please provide an overall score for this submission.
The rating must be one of the following values only: 1, 3, 5, 6, 8, or 10.

The overall rating is not an average of soundness, presentation, and contribution.
The overall rating must be determined by scientific content only: novelty, correctness,
     evidence, evaluation quality, reproducibility, limitations, and significance.
Good writing, confident framing, complete formatting, or polished motivation should not
     raise the rating unless they reveal stronger scientific content.
Poor writing should not lower the rating unless it obscures essential scientific content,
     prevents verification, or leaves necessary methods/evidence unavailable.

Assign the rating based on concrete scientific evidence in the paper.
Do not give credit for missing, implicit, vague, or weakly supported evidence.
Do not reward claims about novelty, impact, or rigor unless they are supported by the
     actual method, analysis, theory, or experiments.

Rating scale:

1 = Strong Reject
A paper whose scientific content has fundamental problems such as well-known or trivial
     results, serious technical flaws, unsupported central claims, invalid or misleading
     evaluation, severe reproducibility problems, or unaddressed ethical concerns.

3 = Reject
A paper whose scientific content is not competitive for a top-tier AI conference due to
     meaningful weaknesses in novelty, technical soundness, evaluation, reproducibility,
     significance, or positioning.
Use 3 for papers with weak novelty, weak empirical support, missing important baselines,
     insufficient ablations, unclear contribution, overclaimed results, limited
     significance, or poor positioning relative to prior work.
A paper can receive 3 even if it is well written, well formatted, easy to follow, or
     rhetorically convincing.

5 = Marginally below the acceptance threshold
A mostly technically solid paper with some clear scientific strengths, but where scientific
      weaknesses slightly outweigh scientific strengths.
Use 5 only when the scientific content is close to being competitive but still falls short
     due to limited novelty, incomplete evaluation, weak baselines, insufficient ablations,
      narrow scope, limited reproducibility, unclear significance, or claims stronger than
     the evidence.
Please use sparingly.

6 = Marginally above the acceptance threshold
A technically solid paper where scientific strengths clearly outweigh scientific weaknesses
      and there is concrete evidence that the scientific content is competitive for a top-
     tier AI conference.
Use 6 only if the paper has a clear nontrivial contribution, mostly sound methodology,
     convincing evaluation for the main claims, appropriate baselines or comparisons, and
     no major unresolved scientific weakness in novelty, evaluation, reproducibility, or
     significance.
Do not assign 6 merely because the paper is coherent, plausible, well written, or reports
     some positive results.
Please use sparingly.

8 = Accept
A strong paper with a clear, meaningful, and well-supported scientific contribution;
     convincing technical quality; good-to-excellent evaluation; appropriate comparisons;
     sufficient reproducibility; and clear relevance to the field.
Use 8 only when the scientific content would be clearly competitive at a top-tier AI
     conference and has no major unresolved issue in soundness, novelty, evaluation,
     reproducibility, significance, or ethics.

10 = Strong Accept
An exceptional paper with groundbreaking scientific impact on one or more areas of AI,
     technically excellent work, exceptionally strong evaluation, strong reproducibility
     and resources, and no unaddressed ethical considerations.
This rating should be reserved for rare submissions whose scientific content is clearly
     among the best papers in the venue.

Scientific-content-only reviewing guidance:
If the main scientific claims are not adequately supported, the rating should be at most 3.
If the paper lacks important baselines, necessary ablations, or evaluation needed to
     support the main claims, the rating should usually be at most 5.
If the scientific contribution is incremental, narrow, or weakly positioned relative to
     prior work, the rating should usually be at most 5.
```

Prompt 8: Persistent Direct-Review Prompt (continued)

```
If either soundness or contribution is rated 1, the overall rating should usually be 1 or
     3.
If either soundness or contribution is rated 2, the overall rating should usually be no
     higher than 5 unless there is unusually strong compensating scientific evidence.
When uncertain between two adjacent ratings, choose the rating better supported by concrete
      scientific evidence; do not use presentation quality as the tiebreaker.

8. Confidence
Use a 1-5 scale:
1 = educated guess
2 = willing to defend but likely missed central parts
3 = fairly confident
4 = confident but not absolutely certain
5 = absolutely certain, checked details carefully

Output Format

You MUST output your review as a single valid JSON object with exactly the fields shown
     below.
Do NOT output any text outside the JSON object.
Do NOT use markdown fences, preamble, or commentary.
Use numeric values, not strings, for soundness, presentation, contribution, rating, and
     confidence.

{
  "summary": "<text>",
  "soundness": <num>,
  "presentation": <num>,
  "contribution": <num>,
  "strengths": ["<text>", "<text>"],
  "weaknesses": ["<text>", "<text>"],
  "questions": ["<text>"],
  "ethics_flag": "No ethics review needed.",
  "rating": <num>,
  "confidence": <num>
}
```