# Evaluating LLM Judgment via Surgical Edits

Khuong T. G. Hieu[1], I. Guyon[2], B. Kent Rachmat[1], I. Ullah[2], and Z. Xu[3*]

1- Université Paris-Saclay, France.

2- ChaLearn, USA.

3- University of Chicago, USA.

**Abstract**. The growing volume of scientific paper submissions raises interest in LLMs as review assistants, but imitation of historical reviews risks reproducing spurious correlations and biases. We propose a controlled-intervention protocol to test whether an agent is causally sensitive to real quality improvements. Each instance uses a triplet of papers (original flawed, substantive human revision, and superficial LLM revision) to assess whether the agent detects genuine improvements while ignoring superficial edits. The strongest models successfully detect gains in flaw categories requiring internal logical consistency, while exhibiting near-zero or negative sensitivity to those requiring external empirical verification.

Supplemental material including data, code, and appendices is provided.

## 1 Introduction

The foundation of scientific progress relies on rigorous expert feedback, primarily through the peer review system [12]. This system is currently under significant strain [6, 5, 1]. The rapid progress of Large Language Models (LLMs) has led to proposing their use as AI reviewer assistants [9, 14]. However, designing such AI agents is not a straightforward machine learning problem. A naïve formulation, such as training a model to predict accept/reject decisions or to imitate human review text from historical data [2], faces fundamental obstacles. First, the data violates the classical *i.i.d.* assumptions: review criteria shift between venues, scientific standards evolve over time. Second, LLMs are trained on data from the public internet, including many published papers/reviews (*e.g.*, on OpenReview), leading to potential test-set contamination and leakage. Third, human reviews exhibit biases and inconsistencies [8, 3, 15] , thus an agent trained to imitate them may internalize such flaws [7]. Finally, any deployed review assistant will encounter adversarial or strategically crafted inputs, diverging from original training distributions. This motivates reframing the task from prediction to causal evaluation: assessing whether an agent responds to genuine improvements while ignoring irrelevant variation. We implement this via a controlled intervention protocol using triplets of papers: an original flawed version, a substantive human revision (treatment), and a superficial LLM revision (control). Comparing the agent's judgments on treatment versus control yields a causal measure of its sensitivity to real scientific improvement.

*The authors are in alphabetical order of last name, except the first author. Supplemental material is found at: https://github.com/ktgiahieu/eval_llm_surgical_edits. The corresponding author ktgiahieu@gmail.com is funded by a grant of Paris Région Ile-de-France.

## 2 Methodology

### 2.1 Causal Verification Framework

Current approaches to AI-assisted peer review often frame the task as imitating human reviewers [2]. However, human decisions are influenced by both valid causal quality rubrics $Q$ related to paper soundness and unauthorized confounders $U$, such as author prestige, temporal trends, or stylistic biases (App A). Unlike imitation approaches, based on observational data, which absorb both $Q$ and $U$, we propose a counterfactual framework [10] to disentangle the causal effect of $Q$, via targeted sensitivity analysis. By controlling for $U$ via intervention, we define a rigorous evaluator as one sensitive to genuine improvements (treatment) while robust to superficial edits (placebo). Our proposed methodology aims to mitigate the problems mentioned in the introduction, which all stem from evaluating LLMs in an observational rather than interventional setting.

### 2.2 Protocol: The Triplet Intervention

We evaluate triplets $\{P_{\text{bad}}, P_{\text{genuine}}, P_{\text{superficial}}\}$ to isolate evaluator sensitivity to $Q$ (App B). The components are: **(1) Baseline ($P_{\textbf{bad}}$):** A paper containing a specific, fatal scientific flaw; **(2) Treatment ($P_{\textbf{genuine}}$):** The human-revised camera-ready version (gold standard); and **(3) Placebo ($P_{\textbf{superficial}}$):** An LLM-generated "sham" fix (good-looking, but scientifically hollow).

The LLM-verifier rates revision quality $S$ on a 9-point scale (App F.3). $S_{\text{genuine}}$ rates the Treatment *vs.* Baseline revision while $S_{\text{superficial}}$ rates the Placebo *vs.* Baseline revision. We quantify sensitivity via the paired Cohen's $d$:

$$\text{Effect} = \frac{\text{Mean}(S_{\text{genuine}} - S_{\text{superficial}})}{\text{StdDev}(S_{\text{genuine}} - S_{\text{superficial}})} \tag{1}$$

A positive effect implies the model correctly distinguishes substantive scientific work from hallucinations or stylistic edits.

### 2.3 Dataset: Surgical Flaw Injection

Lacking access to original submissions, we reverse-engineer triplets from 1,573 NeurIPS 2024 papers. Following [13], we identify consensus flaws from OpenReview (flaws identified by at least one reviewer, which the authors acknowledge should be corrected) and use an auxiliary LLM to *surgically inject* them into the camera-ready ($P_{\text{genuine}}$) to create $P_{\text{bad}}$ (App C.3). We performed a partial manual inspection of the resulting flawed papers for quality control. We then task an LLM (blind to the original solution) to fix $P_{\text{bad}}$ based solely on the flaw description to generate $P_{\text{superficial}}$. Papers are preprocessed to Markdown, retaining tables and citation abstracts (App C.1). We balanced the dataset across a taxonomy of 13 flaw categories (App C.2), yielding 50 triplets per category.

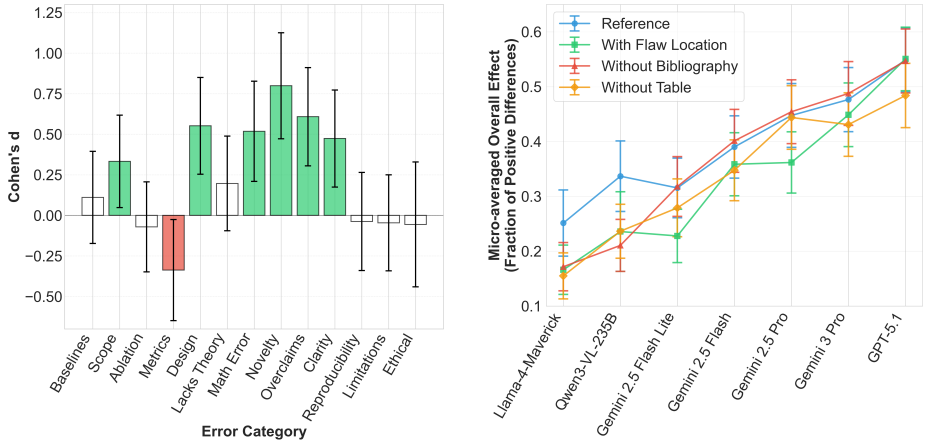



Fig. 1: **Comparison of treatment effects.** a. Effect across flaw rubrics for Gemini 2.5 Pro. b. Average effect for various models and varying experimental conditions (see text).

## 3 Experiments

We evaluated a suite of 7 models: proprietary (Gemini 3 Pro, Gemini 2.5 Pro, GPT 5.1, Gemini 2.5 Flash, Gemini 2.5 Flash Lite) and open-weights (Qwen3-VL-235B-A22B-Instruct, Llama-4-Maverick-17B-128E-Instruct). To mitigate self-preference bias [16], we used a different LLM for error planting/surgical injection (OpenAI o3). All experiments used a fixed zero-shot prompt (App F).

### 3.1 Experimental Setup

Open-weights inference ran on NVIDIA A100-80GB GPUs via vLLM; proprietary models used standard APIs. Hyperparameters are in App E.

### 3.2 Prototypical Evaluation in the Reference Setting

We illustrate the methodology using Gemini 2.5 Pro (Figure 1.a), in the *reference setting*, which will be used as reference in ablation studies. In that setting, the verifier inputs are: full pre- and post-revision papers (including bibliography/abstracts) and the flaw description (App C.2). No flaw location information is provided. $P_{\text{bad}}$ serves as the pre-revision baseline, with $P_{\text{genuine}}$ and $P_{\text{superficial}}$ as post-revision inputs (50 triplets per flaw category). The quality of the revision is judged on a 9-point Likert scale by the LLM verifier. We report the "treatment effect" as calculated by the Cohen's d (Eq. 1). The error bar is the sample error 95% CI, assuming a T-distribution.

Positive effect (green bars) indicates successful distinction of human ($P_{\text{genuine}}$) from LLM ($P_{\text{superficial}}$) revisions, e.g., "Math Error" and "Novelty" ($pvalue < 0.05$). Negligible (white) or negative (red, e.g., "Metrics") effects are observed

for empirical categories, suggesting competence domain limits. Variance from model seed is shown to be negligible (App D.1).

### 3.3 Model Comparison and Verifiability Tiers

All models showed similar performance profiles (App D.2). We group flaw rubrics into three tiers based on consistent Cohen's $d$ magnitude:

**T1 High Reliability ($d > 0.5$): Logical Edits.** (Theory, Math, Novelty, Clarity). The verifiers successfully track internal logic and textual consistency; verification requires no external data.

**T2 Medium Reliability ($0.3 < d \leq 0.5$): Calibration Edits.** (Scope, Overclaims, Design). Verifiers reward language calibration (e.g., claiming less). They may be assisted by the provided citation abstracts.

**T3 Low/Negative Reliability ($d \leq 0.3$): Unverified Claims.** (Experimentation: Baselines, Ablation, Reproducibility, Metrics; Broader Impact: Limitations, Ethical). LLM verifiers are susceptible to "empirical blindness," accepting *claims of existence* (e.g., code, baselines) as *proof of existence* due to an inability to verify external artifacts. The negative effect in Metrics suggests preference for perfect, hallucinated control data over messy, real human data.

Reliability strongly correlates with **verifiability**. To improve Tier 3 performance, LLM verifiers must be trained to analyze empirical evidence (figures/tables) rather than solely relying on textual fluency, potentially using tool-augmented execution environments. We also advocate for LLMs to report non-verifiability when evidence is insufficient.

In App D.3.6, we find that splitting the scalar score into Quality and Verifiability confirms this: T1 and T2 rubrics score high on both, while T3 rubrics show low Verifiability (Reproducibility, Ethical) or low Quality (Baselines, Ablation).

### 3.4 Systematic Ablation Studies

We compared models across three ablations (App D.3.1, D.3.2, D.3.3) to assess ranking robustness (Figure 1.b).

A1 **Localizing Changes:** Providing snippets (diffs) unexpectedly *decreased* effect size. We hypothesize this focuses the model on *textual effort* rather than *scientific outcome*, distracting it from the sham nature of the control.

A2 **Removing Bibliography:** This was largely *effect-neutral*, suggesting citation context is insufficiently utilized to ground verification.

A3 **Removing Tables:** Unsurprisingly, this *decreased* the effect size, confirming that empirical evidence is a necessary, though often misunderstood, signal.

These counter-intuitive results (A1 and A2) suggest that current LLMs suffer from information overload or misdirection in long-context verification tasks. Further work is needed to examine the verifier's textual explanations.

## 4 Discussion, Limitations, and Conclusions

This paper made the following contributions: We re-framed AI reviewing as a causal evaluation problem; we proposed a controlled-intervention protocol with triplets of flawed, human-revised, and LLM-revised papers; we built a benchmark of such triplets with flaw categories; we ran a comparative empirical study across LLM judges; and we showed how intervention tests expose failure modes and help design more reliable AI reviewers.

The intervention-based evaluation reveals two distinct error sources in the treatment-control effect estimates. The first is revision-quality ambiguity: a verifier may not detect that a control revision only partially addresses the stated flaw. The second is revision-verifiability failure: a verifier may accept fabricated evidence as correct when the paper provides insufficient information to validate it. Future LLM verifiers should support three types of diagnoses: (i) wrong or absent change; (ii) change that is in principle verifiable but unsupported by the paper; (iii) change that may be verifiable externally (e.g., via data or code) but is not verifiable by the verifier given access limits. Current experiments mainly separate (i) from ii, iii. Distinguishing (ii) vs. (iii) requires explicit modeling of access to external artifacts and a verifier behavior that can return "not verifiable with provided information."

Additional limitations are methodological. Our reliance on semi-synthetic flawed papers created by injecting flaws in the camera-ready papers ensures perfect ground truth for causal measurement but may lack the nuance of organic submission flaws. The "surgical" nature of our injection might leave artifacts that are easier to detect than natural errors. However, this is a necessary trade-off to scale the evaluation beyond the small number of papers with public pre-rebuttal versions available on arXiv. Additionally, de-planting errors to create control papers assumes that the control revision is superficial relative to the treatment (human revisions). This holds for many flaw types but weakens in categories where human fixes are minimal while LLM controls are expansive. Effect sizes also depend on prompting: prior experiments show near-zero signal for most open models under generic prompts, while a review-form prompt yields measurable sensitivity for Gemini-2.5-Pro. Hence reported effects are properties of (model, prompt, rubric) triples, not of the model alone. The variance in our effect sizes (indicated by wide error bars in Figure 1) suggests that larger sample sizes are needed to achieve significance in subtler flaw categories. In future work will expand the dataset to include ICLR and ICML papers and explore tool-augmented LLMs capable of executing code to verify empirical claims.

In conclusion, the intervention protocol measures causal sensitivity of LLM evaluators to substantive scientific revision, but current estimates are affected by verifiability limits and prompt dependence. Empirically, verifiers often reward fabricated but detailed controls in evidence-heavy rubrics, especially Ethics and Metrics. Future work should (a) report dual scores separating quality from verifiability, (b) expand intervention families toward flaws that are locally checkable, and (c) include verifier options for "insufficient evidence to verify." These

extensions are necessary for characterizing LLM judgment behavior under the evaluation criteria targeted by responsible ML assessment.

## References


[1] M. Breuning, J. Backstrom, J. Brannon, B. I. Gross, and M. Widmeier. Reviewer fatigue? why scholars decline to review their peersâ work. *PS: Political Science & Politics*, 48(4):595–600, 2015.

[2] A. Checco, L. Bracciale, P. Loreti, S. Pinfield, and G. Bianchi. Ai-assisted peer review. *Humanities and social sciences communications*, 8(1):1–11, 2021.

[3] C. Cortes and N. D. Lawrence. Inconsistency in conference peer review: Revisiting the 2014 neurips experiment. *arXiv preprint arXiv:2109.09774*, 2021.

[4] S. Fricke. Semantic scholar. *Journal of the Medical Library Association: JMLA*, 106(1):145, 2018.

[5] R. E. Gropp, S. Glisson, S. Gallo, and L. Thompson. Peer review: A system under stress. *BioScience*, 67(5):407–410, 2017.

[6] M. A. Hanson, P. G. Barreiro, P. Crosetto, and D. Brockington. The strain on scientific publishing. *Quantitative Science Studies*, 5(4):823–843, 2024.

[7] M. Hosseini and S. P. Horbach. Fighting reviewer fatigue or amplifying bias? considerations and recommendations for use of chatgpt and other large language models in scholarly peer review. *Research integrity and peer review*, 8(1):4, 2023.

[8] J. Huber, S. Inoua, R. Kerschbamer, C. König-Kersting, S. Palan, and V. L. Smith. Nobel and novice: Author prominence affects peer review. *Proceedings of the National Academy of Sciences*, 119(41):e2205779119, 2022.

[9] C. Lu, C. Lu, R. T. Lange, J. Foerster, J. Clune, and D. Ha. The ai scientist: Towards fully automated open-ended scientific discovery. *arXiv preprint arXiv:2408.06292*, 2024.

[10] J. Pearl. *Causality: Models, Reasoning and Inference*. Cambridge University Press, USA, 2nd edition, 2009.

[11] J. Priem, H. Piwowar, and R. Orr. Openalex: A fully-open index of scholarly works, authors, venues, institutions, and concepts. *arXiv preprint arXiv:2205.01833*, 2022.

[12] R. Smith. Peer review: a flawed process at the heart of science and journals. *Journal of the royal society of medicine*, 99(4):178–182, 2006.

[13] G. Son, J. Hong, H. Fan, H. Nam, H. Ko, S. Lim, J. Song, J. Choi, c. P. Gon Y. Yu, and S. Biderman. When ai co-scientists fail: Spot-a benchmark for automated verification of scientific research, 2025.

[14] N. Thakkar, M. Yuksekgonul, J. Silberg, A. Garg, N. Peng, F. Sha, R. Yu, C. Vondrick, and J. Zou. Can llm feedback enhance review quality? a randomized study of 20k reviews at iclr 2025. *arXiv preprint arXiv:2504.09737*, 2025.

[15] A. Tomkins, M. Zhang, and W. D. Heavlin. Reviewer bias in single-versus double-blind peer review. *Proceedings of the National Academy of Sciences*, 114(48):12708–12713, 2017.

[16] K. Wataoka, T. Takahashi, and R. Ri. Self-preference bias in LLM-as-a-judge, 2025.

# A Theoretical Framework

## A.1 Formalizing the Review Process

The review process is defined as a decision-making system. A reviewer provides a discriminant function $D(A, L, Q, U)$, where:

- $A$ is the **Article** under review.
- $L$ is the set of past **Literature** publicly available.
- $Q$ is the set of recommended review **Quality rubrics** (e.g., novelty, soundness, clarity) provided in the review instructions.
- $U$ is the set of **Unauthorized** or protected factors (e.g., author prestige, institution, irrelevant stylistic choices).

The accept/reject recommendation is based on applying a threshold $\Theta$ to this function: Accept if $D(A, L, Q, U) > \Theta$, and Reject otherwise.

Factors affecting this decision can be categorized based on their causal relationship to paper quality:

- **Direct Causes:** Factors extracted from the article $A$ that map directly to the quality rubrics $Q$. These are the only factors that *should* affect the decision $D$.
- **Unauthorized Factors (U):** This set includes all other factors that should *not* affect $D$. These can be:
  - **Confounders:** Factors with a common cause, such as an author's writing style. The author is the source of both the scientific quality (a $Q$ factor) and the writing style (a $U$ factor). A reviewer should not use style as a proxy for quality.
  - **Proxies (Shortcuts):** Bias factors sourced from the reviewer, often used to expedite a review. An example is checking if the author cites the reviewer's own papers.
  - **Indirect Causes:** Factors that affect $Q$ but should not affect $D$ directly (e.g., an author's origin might influence their training, which affects paper quality, but $D$ should be independent of origin *given* the paper's quality).

The central problem is that $U$ factors may correlate strongly with $Q$ factors and be highly predictive of human review decisions, even though they are not causally related to paper quality. An AI agent trained to imitate human reviews will learn these spurious correlations.

### A.2 Ideal vs. Human Review

An ideal, transparent AI review process should function only on the quality rubrics, $Q$. The discriminant function should be $D(Q(A, L))$, as shown in Figure 3. In contrast, the human review process is influenced by both $Q$ and $U$ factors (Figure 2). Therefore, human review data is a flawed target for imitation.

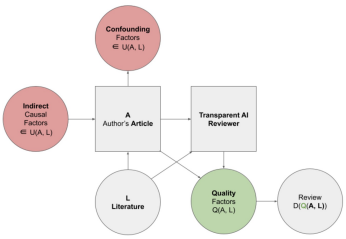



Fig. 2: Human review process. The review $D(A, L, Q, U)$ can be influenced by paper Quality rubrics and Unauthorized factors.

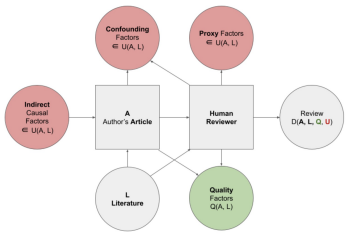



Fig. 3: Ideal transparent AI review process. The AI reviewer extracts and uses only paper Quality rubrics, $Q(A, L)$.

### A.3 Intervention-Based Evaluation

Given that an AI agent $D$ is a black box (e.g., an LLM), it is difficult to guarantee that it only uses $Q$ factors. Its internal reasoning is opaque. Therefore, testing must be done externally by **intervening** on the inputs and observing the outputs (Figure 4).

The core questions for evaluation are:

1. If a $Q$ factor is manipulated (e.g., a fatal flaw is fixed), does the agent's decision $D$ change appropriately?

2. If a $U$ factor is manipulated (e.g., author names are changed from novice to famous), does the agent's decision $D$ remain stable?

A robust AI agent must be sensitive to (1) and insensitive to (2).

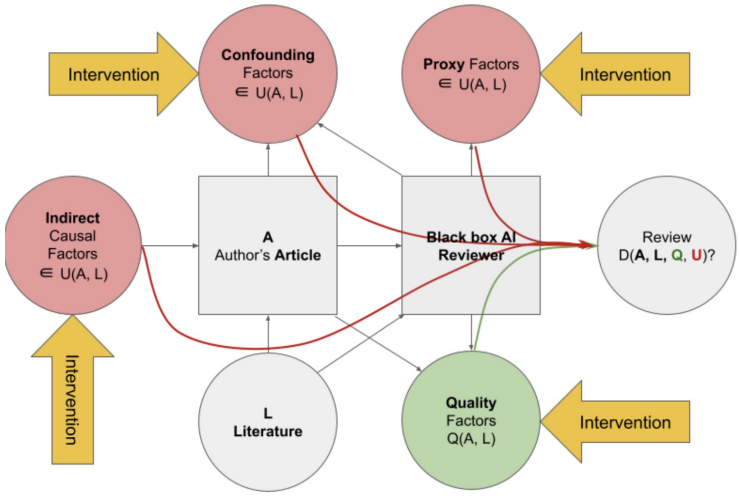



Fig. 4: Testing an AI reviewer must be done by intervening on both Quality ($Q$) and Unauthorized ($U$) factors to check for desired sensitivity and required insensitivity.

## B Experimental Protocol

This specific protocol is designed to test sensitivity to quality-factor interventions. It evaluates an AI agent's ability to **verify the quality of a revision**.

### B.1 The Triplet Design

We use triplets of papers to establish treatment and control groups:

1. **"Bad" (Baseline):** This emulates an initial submission with a single, major flaw. It is created by taking a final camera-ready paper and programmatically "planting" a known, consensus flaw (identified from the original paper's public reviews).
2. **"Genuine" (Treatment Group):** This is the original camera-ready paper, which represents a genuine, best-effort revision where the authors (humans) fixed the flaw.
3. **"Superficial" (Control Group):** This emulates a zero-effort, superficial revision. It is created by taking the "Bad" paper and using an LLM to "fix" the flaw, given only the description of the flaw. This serves as a "sham surgery" control.

### B.2 Evaluation Task

An AI agent (the "LLM verifier") is tasked with scoring the quality of the revision. The agent is provided with:

- The “True Bad” paper (as the ”original submission”).
- The description of the flaw that needed to be fixed.
- EITHER the “True Good” (treatment) OR the “Fake Good” (control) paper (as the “revision”).

The agent outputs a **verification score** on a scale of 1-9, indicating how well the revision addressed the flaw.

### B.3 Analysis and Metric

This design allows for a paired-sample analysis. For each baseline paper, there is a score for the treatment (human fix) and a score for the control (LLM fix). The primary metric is the **Treatment Effect**.

The treatment effect is calculated as the average difference between the paired scores, normalized by the standard deviation of the differences.

$$\text{Treatment Effect} = \frac{\text{Avg}[\text{score(true good)} - \text{score(fake good)}]}{\text{StDev}[\text{score(true good)} - \text{score(fake good)}]}$$

A positive and significant treatment effect indicates that the AI verification agent successfully identifies the ”True Good” (human-revised) paper as a higher-quality revision than the ”Fake Good” (LLM-revised) paper.

## C Dataset and Implementation Details

### C.1 Data Collection and Preprocessing

Our data collection process, managed via our ***public repository***, sourced a corpus of approximately 20,000 camera-ready papers and their corresponding LaTeX source files from arXiv. The dataset for this study consists of 1,573 accepted papers from NeurIPS 2024 available in Markdown format.

To facilitate automated processing and flaw injection, we converted each paper from its original LaTeX source into Markdown format using a custom-modified version of Pandoc. This conversion preserved essential structural elements such as formulas, figures, and tables as distinct, identifiable components. Furthermore, to provide the LLM reviewers with necessary bibliographic context, we utilized the Semantic Scholar API [4] and OpenAlex API [11] to retrieve the abstracts of all cited articles. These abstracts were then concatenated with the main text of the paper. We selected Markdown as the intermediate format for three key reasons: it simplifies the programmatic injection of flaws into the text; it allows for easy identification and processing of the bibliography; and it aligns with the input requirements of modern multimodal LLMs, which often necessitate handling figures separately from the main text body.

To enhance the context available to the model, the textual data was enriched by appending the abstracts of all referenced papers in the bibliography.

Additionally, all figures were extracted and saved to ensure full multimodal accessibility during the review generation process.

| Stage | Implementation Details |
|---|---|
| Source Data | 1,573 NeurIPS 2024 accepted papers (LaTeX source + PDF). |
| Expansion | Recursive injection of all local '.tex' files via 'input' resolution to create a monolithic source. |
| Normalization | Regex-based simplification of complex 'tabular' environments and 'algorithmic' blocks; removal of visual-only macros (e.g., 'vspace', 'centering'). |
| Conversion | Custom Pandoc pipeline with iterative compilation: on failure, the engine identifies the error line, disables the specific command, and recompiles. |
| Reference Enrichment | Waterfall resolution for **cited abstracts**: OpenAlex API → Semantic Scholar API → Elsevier/Springer API → Google Search & HTML Scraping → PDF text extraction. |
| Visuals | Figures in Vector PDFs and other formats are converted to PNGs via 'pdf2image'/Poppler for multimodal compatibility. |

Table 1: Preprocessing pipeline summary.

### C.2 Flaw Categorization

To evaluate the sensitivity of LLMs to specific scientific deficits, papers were categorized according to the following thirteen flaw types:

1. **Baselines** *(Insufficient Baselines/Comparisons)*: The evaluation is missing comparisons to relevant, state-of-the-art, or obvious alternative methods.

2. **Scope** *(Weak or Limited Scope of Experiments)*: The experiments are too narrow to support the paper's general claims (e.g., "toy" problems, insufficient data, no real-world testing).

3. **Ablation** *(Lack of Necessary Ablation or Analysis)*: The paper fails to analyze why its method works, missing ablation studies, cost/scalability analysis, or parameter sensitivity checks.

4. **Metrics** *(Flawed Evaluation Metrics or Setup)*: The metrics used are inappropriate or misleading, or the experimental setup is unreliable.

5. **Design** *(Fundamental Technical Design Limitation)*: The proposed method has an inherent design flaw that severely restricts its applicability or performance (e.g., requires unrealistic inputs, cannot scale by design).

6. **Lacks Theory** *(Missing or Incomplete Theoretical Foundation)*: The paper requires but lacks a formal theoretical justification for its method (e.g., no convergence guarantees, no formal proof).

7. **Math Error** *Technical or Mathematical Error:* The paper contains a demonstrable error in its mathematical derivations, proofs, or algorithm description.

8. **Novelty** *Insufficient Novelty / Unacknowledged Prior Work:* The core contribution is highly similar to or an uncredited rediscovery of existing work.

9. **Overclaims** *(Overstated Claims or Mismatch Between Claim and Evidence)*: The paper's claims in the abstract, introduction, or conclusion are stronger than what the experimental results actually support.

10. **Clarity** *(Lack of Clarity / Ambiguity)*: The paper is written in a way that is ambiguous or difficult to understand, preventing an expert from properly interpreting the work.

11. **Reproducibility** *(Missing Implementation or Methodological Details)*: The paper omits crucial details needed for reproduction (e.g., key hyperparameters, data processing steps, source code).

12. **Limitations** *(Unacknowledged Technical Limitations)*: The paper fails to discuss or downplays obvious or crucial limitations of its method, evaluation, or theoretical assumptions.

13. **Ethical** *(Unaddressed Ethical or Societal Impact)*: The paper fails to address potential negative societal impacts, risks of misuse, fairness, or other ethical considerations raised by the research.

The taxonomy of scientific flaws used in this study was derived from the **NeurIPS 2024 Review Guidelines**. We manually refined these guidelines to ensure orthogonality between categories, resulting in distinct flaw types grouped into five high-level categories.

To automate the categorization of flaws from raw review texts, we created a "Golden Set" of 100 manually annotated flaw descriptions. We tuned the prompt provided in Appendix F.4 until it achieved 100% classification accuracy against this human-verified dataset.

From this corpus, 1,573 NeurIPS papers were organized based on specific error categories. The available pool for each category ranged from approximately 120 to 300 papers. For the final evaluation set, 50 pairs of papers were randomly selected for each Error Category.

### C.3 Flaw Re-introduction (Injection)

The core of our benchmark is the controlled re-introduction of major, rejection-worthy flaws into high-quality papers. We began by manually annotating consensus flaws from a "golden set" of rebuttals and reviews for 50 randomly sampled papers from our corpus. These were flaws identified by human reviewers that the authors acknowledged and agreed to fix in the final camera-ready version.

We specifically filtered for "actionable" weaknesses—concrete errors or omissions that could be rectified—while excluding minor issues like typos or stylistic preferences.

Using this annotated set, we designed a few-shot prompt in Appendix F.1.1 for OpenAI's o3 model to identify similar consensus flaws across the broader dataset. For each identified flaw, we recorded the anonymized reviewer ID and the specific description of the error.

For each paper $i$ in our source corpus (denoted as the original camera-ready version $o^i$), we identified a set of $K_i$ consensus flaws, $\{f_1^i, f_2^i, ..., f_{K_i}^i\}$. For each flaw $f_k^i$, using the prompt from Appendix F.1.2, we created a corresponding degraded version of the paper, $x_k^i$, using the following automated procedure:

1. **Localization:** We identified the 1–3 specific paragraphs in the original paper $o^i$ that were directly related to the flaw $f_k^i$.

2. **Injection:** We prompted OpenAI o3 to rewrite these identified paragraphs to re-introduce the specific flaw $f_k^i$ described in the review history. This effectively "degrades" the paper back to its state before the flaw was fixed (or to a similar flawed state).

This process yields a set of degraded papers, where each $x_k^i$ differs from the high-quality original $o^i$ only by the presence of a single, known, and significant flaw. The validity of this re-introduction process was manually verified on a random sample of 50 instances, all of which were confirmed to successfully contain the intended flaw.

### C.4 Revision attempt to address flaw (De-planting)

To establish a rigorous control group (the "fake good" papers), we simulate a scenario where an author attempts to resolve a major scientific flaw solely through LLM generation, without conducting the necessary underlying research or experiments.

For each degraded paper $x_k^i$ containing a specific planted flaw $f_k^i$, we use LLM to generate a revision. There exist 2 variations:

1. **With** flaw location information (prompt F.3.2): The automated pipeline first locates the flawed section using the structural metadata (headings and text markers) recorded during the injection phase. The LLM is then provided with the passage to rewrite, the location of the flaw, and the specific description of the flaw.

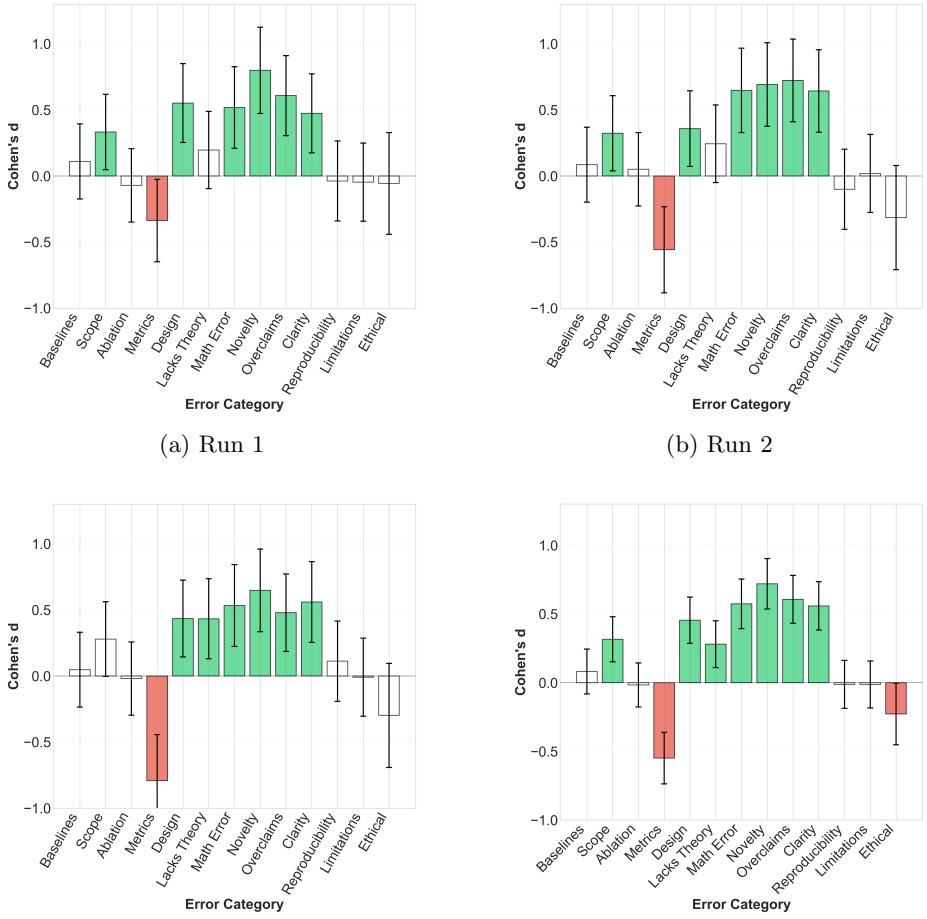
2. **Without** flaw location information (prompt F.3.1): The LLM is only provided with the passage to rewrite and a specific description of the flaw.

Crucially, the prompting strategy is designed to emulate a "sham surgery": the model is explicitly instructed to fabricate plausible-sounding evidence to cover the flaw. The system instructions require the model to "provide concrete experimental results and quantitative evidence... as if they have been conducted"

and to include “specific numerical metrics” and “experimental details” (such as ablation studies or hyperparameter analyses) to substantiate the new claims. This approach ensures that the resulting control papers contain hallucinated but textually convincing fixes that mimic the style and structure of the original manuscript. The generated revisions are programmatically spliced back into the document to produce the de-planted paper.

# D Results and Experimental Ablations

## D.1 Reference Experiment: Between-run Stability Analysis




(a) Run 1

(b) Run 2

(c) Run 3

(d) Average (3 runs) with SE bars

Fig. 5: Stability analysis of Gemini 2.5 Pro in the Reference Setting.

To validate the reliability of our evaluation pipeline, we first established the stability of the *reference setting* with Gemini 2.5 Pro. In this setting, the inputs to the LLM verifier are the full pre-revision (planted error) and post-revision papers. The post-revision papers are either the camera-ready version (treatment) or the LLM-fixed version (control). The input includes the bibliography and abstracts of cited papers, but explicitly excludes information regarding the specific location of the flaw.

The quality of the revision is judged on a 9-point Likert scale. We report the effect size calculated by Cohen's $d$ (Eq. 1). In the bar charts presented throughout this section, we utilize the following color convention:

- **Green:** Indicates a statistically significant positive effect ($p < 0.05$), meaning the model successfully distinguishes the substantive human revision from the superficial LLM revision.
- **White:** Indicates no statistically significant difference (inconclusive).
- **Red:** Indicates a significant negative effect, where the model prefers the superficial LLM revision over the human expert revision.

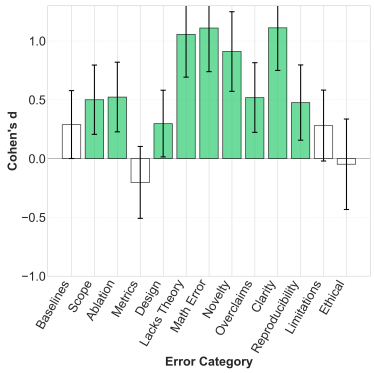
We repeated the reference experiment on Gemini 2.5 Pro three times with different random seeds. As shown in Figure 13, we observed negligible variance in the effect sizes across the 13 flaw categories. The error bars in Figure 5d represent the standard error over the three experiments (calculated as $\overline{\sigma}/\sqrt{3}$). Given this high stability, we proceeded with a single experimental run for the subsequent model comparisons and ablation studies to maximize computational efficiency.

### D.2 Experiment 2: Model Comparisons in Reference Setting

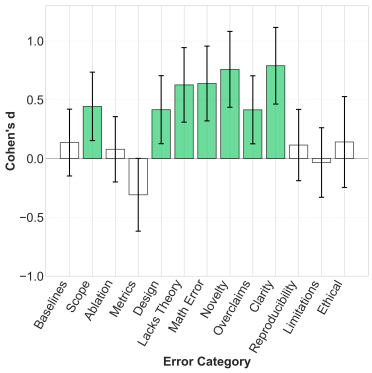
We evaluated a suite of models in the reference setting to benchmark current capabilities. Figure 6 displays the detailed bar graphs for all tested models.

**Analysis of Strong Models:** When aggregating results across the strongest models (GPT 5.1, Gemini 3 Pro, and Gemini 2.5 Pro) in Figure 7, we observe a distinct hierarchy in rubric reliability:

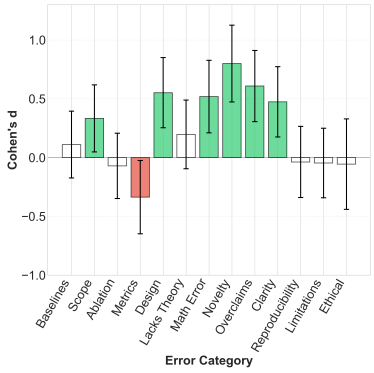
1. **Consistent Positive Effect ($d > 0.5$):** Flaw categories such as *Lack of Theory*, *Math Error*, *Novelty*, and *Clarity*. These rely on internal logical consistency, which strong models can verify effectively.
2. **Moderate Effect ($0.3 < d < 0.5$):** Categories like *Scope*, *Overclaims* and *Design*. Here, verifiers reward calibrated language (e.g., authors claiming less), but quality scores are marginally significant.
3. **Insignificant or Negative Effect:** Categories requiring external verification, such as *Baselines*, *Ablation*, *Reproducibility*, and *Ethical*. Notably, *Metrics* often displays a negative effect, suggesting that LLMs prefer the "perfect" but hallucinated control data over messy real-world results.



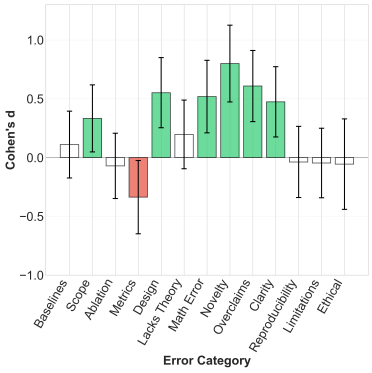

(a) GPT 5.1



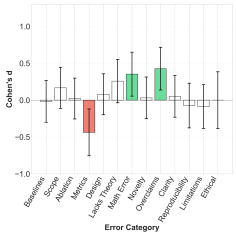

(b) Gemini 3 Pro



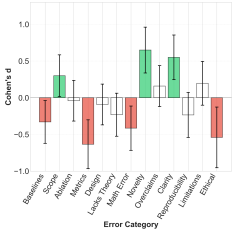

(c) Gemini 2.5 Pro



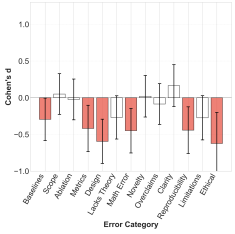

(d) Gemini 2.5 Flash




(e) Gemini 2.5 Flash Lite




(f) Qwen3-VL-235B




(g) Llama-4-Maverick

Fig. 6: Performance comparison across all evaluated models in the Reference Setting.

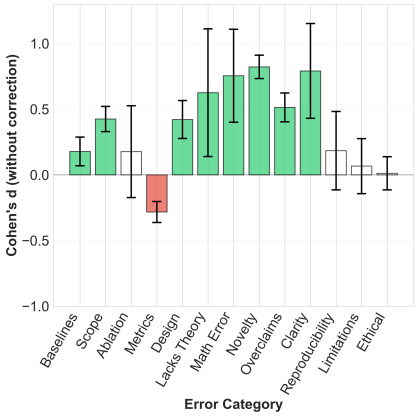



Fig. 7: Aggregated performance across the three strongest models: Gemini 3 Pro, Gemini 2.5 Pro, and GPT 5.1. Error bars represent variance across models.

### D.3 Experiment 3: Ablation Studies (Gemini 2.5 Pro)

To further investigate the drivers of verifier performance, we conducted extensive ablation studies using Gemini 2.5 Pro.

#### *D.3.1 A1: Include Flaw Location in Prompt*

In this ablation, the verifier is explicitly told where the flaw is located. Counter-intuitively, providing the location generally *decreases* the effect size compared to the reference. We hypothesize that localizing the change focuses the model on *textual effort* rather than *scientific outcome*, making the superficial LLM edits (which are textually dense) appear more convincing.

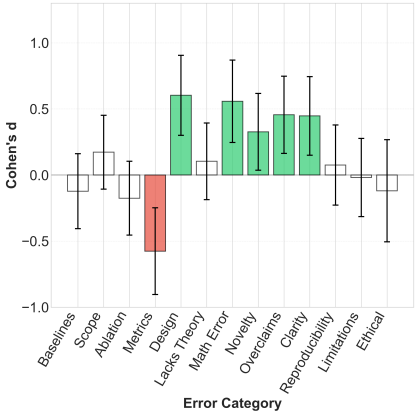



Fig. 8: Effect of providing explicit flaw location (A1). Performance degrades relative to the Reference setting.

### *D.3.2 A2: Remove Bibliography*

Removing the bibliography resulted in performance highly correlated with the Reference setting. While this simplifies the context window, it indicates that current models do not effectively utilize citation graphs or abstract metadata to verify claims, representing a missed opportunity for grounding.

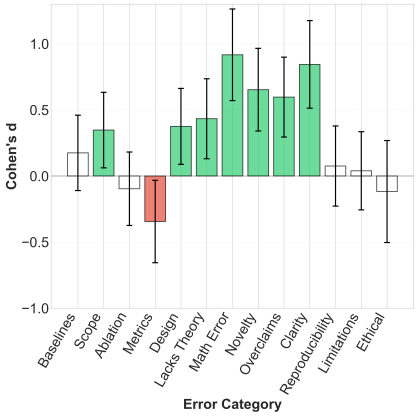



Fig. 9: Performance without bibliography (A2). Results are largely neutral/similar to Reference.

### *D.3.3 A3: Remove Tables*

Removing tables consistently reduced the effect size. This confirms that while models struggle to fully verify empirical data (as seen in the Metrics rubric), the presence of tabular data provides a necessary signal for rubrics like *Baselines* and *Ablation*.

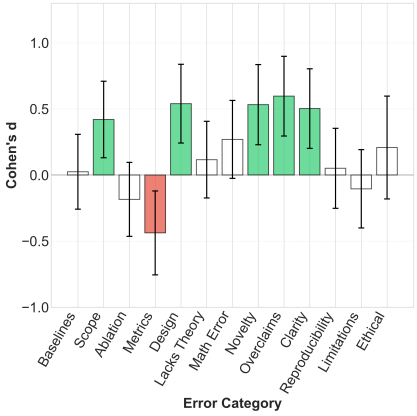



Fig. 10: Performance without tables (A3). A decrease in sensitivity is observed.

### *D.3.4 B1: "Real" Submission (arXiv)*

Here, we replaced the "planted error" paper with the "supposedly original arXiv submission" (pre-rebuttal), by obtaining the corresponding Arxiv papers before the submission deadline. We hope this reflects a more realistic "in-the-wild" setting. We find that it introduces significant variance (large error bars) because the flaws are not surgical. The high number of "white" bars indicates the difficulty of detecting organic flaws compared to planted ones.

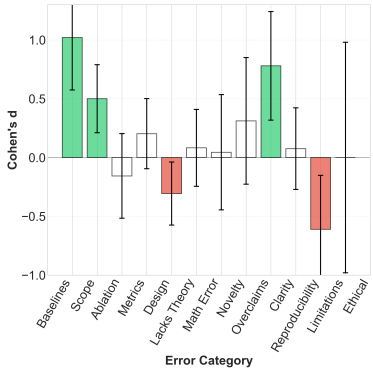



Fig. 11: Performance on "Real" submissions (B1). High variance and lower sensitivity compared to surgical flaws.

### *D.3.5 B2: Snippets Only*

We provided only the paragraphs containing the flaw/fix. The resulting plot is dominated by white bars (insignificant results). This suggests that snippets alone lack the global context required for a reliable verdict; the verifier needs the full paper to assess coherence and consistency.

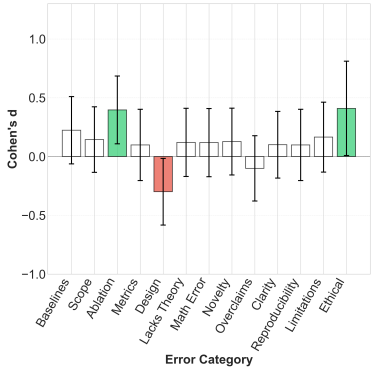



Fig. 12: Performance using only snippets/paragraphs (B2). Lack of context leads to inconclusive results.

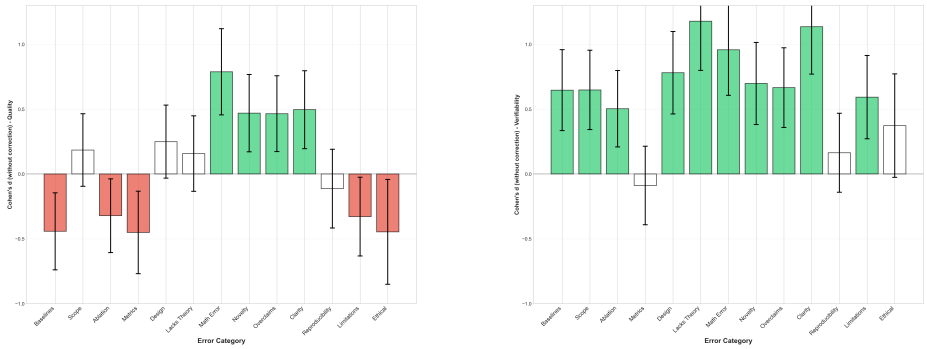
### *D.3.6 B3: Split Scores (Quality vs. Verifiability)*

In Section 3.3, we found that Reliability strongly correlates with **verifiability**. To improve Tier 3 performance, LLM verifiers must be trained to analyze empirical evidence (figures/tables) rather than solely relying on textual fluency, potentially using tool-augmented execution environments.

As a preliminary test, we prompted the model to provide two distinct scores:

1. **Quality of Revision:** Did the change logically address the concern?
2. **Verifiability:** Is the evidence supported by sufficient detail/data?

This ablation clarifies the failure modes in Tier 3 rubrics. For *Baselines* and *Ablation*, the Quality score is negative (the model dislikes the change). For *Reproducibility* and *Ethical*, the Verifiability score drops to near zero, explaining the overall lack of signal.




(a) Quality score (b) Verifiability score

Fig. 13: Performance when splitting the metric into Quality and Verifiability. This breakdown helps explain negative results in empirical categories.

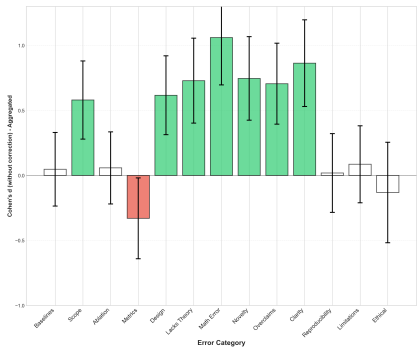



Fig. 14: Aggregated performance after splitting the metric into Quality and Verifiability (B3). This resembles the performance in the *reference setting*.

# E Experimental Settings and Compute Resources

## E.1 Model Specifications

We evaluated a diverse set of state-of-the-art proprietary and open-weights models. Table 2 summarizes the models used in our experiments.

Table 2: Summary of models used in verification experiments.

| Model Name | Type | Access Method |
|---|---|---|
| **GPT 5.1** | Proprietary (OpenAI) | API |
| **Gemini 3 Pro** | Proprietary (Google) | API |
| **Gemini 2.5 Pro** | Proprietary (Google) | API |
| **Gemini 2.5 Flash** | Proprietary (Google) | API |
| **Gemini 2.5 Flash-Lite** | Proprietary (Google) | API |
| **Qwen3-VL-235B-A22B-Instruct** | Open Weights | Local Inference (Jean Zay) |
| **Llama-4-Maverick-17B-128E-Instruct** | Open Weights | Local Inference (Jean Zay) |

## E.2 Hyperparameters

To ensure a fair evaluation, for each model, we adopt the provider's recommended generation parameters when available. In cases where specific recommendations are not provided, we utilize the following default configuration.

- **Temperature:** 0.6
- **Top-p:** 0.95
- **Repetition Penalty:** 1.0 (no penalty)
- **Max Output Tokens:** 8,192 (to allow extensive reasoning)

## E.3 Compute Infrastructure and Grants

Our experiments leveraged cloud APIs and national supercomputing resources.

1. *Proprietary Models:* Access to OpenAI's o3 model for dataset generation and manipulation was supported by the Microsoft Accelerate Foundation Models Research (AFMR) grant program. Gemini models were accessed via the standard Google Cloud Vertex AI API.
2. *Open-Weights Models:* Inference for large-scale open-weights models (Qwen3-VL and Llama-4-Maverick) was performed on the Jean Zay supercomputer, hosted by IDRIS (Institut du développement et des ressources en informatique scientifique).
    - Hardware Specs: We utilized the accelerated A100 partition, consisting of nodes equipped with 8× NVIDIA A100 SXM4 80GB GPUs and dual AMD EPYC 7543 (Milan) 64-core processors.

- Parallelism: Models were deployed using tensor parallelism (vLLM backends) distributed across all 8 GPUs to accommodate their VRAM requirements.

*Acknowledgment:* This work was performed using HPC resources from GENCI-IDRIS (Grant 2025-AD011016658). This work was supported by the Microsoft Accelerate Foundation Models Research (AFMR) grant program and a grant of Paris Région Ile-de-France.

# F LLM Prompts

## F.1 Error Planting Process

### *F.1.1 Detect Consensus Flaw*

**Consensus Flaw Detection Prompt**

You will be provided with the review and rebuttal process for a research paper, as well as its content in markdown format. Your task is to analyze the dialogue and identify only the most crucial flaws that the authors acknowledge and must address for this paper to be publishable.
**Rules:**

1. Focus strictly on significant weaknesses in methodology, statistical rigor, or experimental scope that impact the paper's core claims. Ignore minor points like typos or grammatical errors. If there is no specific section for weaknesses, obtain the crucial flaws based on the main idea being discussed between the reviewers and the authors.
2. A flaw should only be included in the final list if it meets this specific "consensus" criteria:
    - A reviewer raises it as a major weakness, question, or limitation that needs to be addressed.
    - The authors' reply confirms its importance by either: (a) Promising to add the required analysis, data, or clarification to the final, camera-ready version of the paper, or (b) Explicitly agreeing that it is a major limitation of the current study's scope.
3. Do not include issues that the authors state will be addressed in separate "follow-up work" unless it is also clearly acknowledged as a critical limitation of the current paper.
4. Your output MUST be a JSON object that conforms to the provided schema.

**Input:**

```
JSON schema: {json_schema}
Review Process:
<BEGIN REVIEW PROCESS> {review_process} <END REVIEW PROCESS>
Paper:
<BEGIN PAPER> {paper_text} <END PAPER>
```

*F.1.2 Error Planting (Injection)*

**Error Injection Prompt**

You are an AI model creating a benchmark of flawed scholarly papers for a research experiment. Your objective is to modify a published research paper to re-introduce a specific flaw in a way that makes the paper's claims methodologically unsound.
You must adopt the persona of an author who is either unaware of the flaw or actively downplaying it to make their results seem more significant. Stay true to the authors' original material and personas as much as possible, use scientific terms, avoid writing in an obviously different manner in order to create an illusion that the paper is coherent, well-written and the introduced flaws are as hard to detect as possible.
**Your Task:** For each section that needs modification, you must:

1. Generate the new, complete text for only the sections you identified.
2. The rewritten text must maintain the original paper's style but reflect the overconfident persona.

**Input:**

```
The flaw to re-introduce is:
<BEGIN FLAW> {flaw_description} <END FLAW>

The original paper:
<BEGIN PAPER> {original_paper} <END PAPER>
```

### F.2 Attempt to Fix Error (De-planting)

**Error Remediation Prompt (De-planting)**

You are revising a research paper to address identified issues in specific sections.
**Overall Issue Description:** {flaw_description}
**Context Before the Section:** {context_before}
**Sections to Revise:** {sections_prompt}

**Context After the Section:** {context_after}
**Revision Guidelines:**

1. Provide concrete experimental results and quantitative evidence. Include specific numerical metrics (accuracy, F1 scores, loss values, p-values, confidence intervals) rather than statements about future work or planned evaluations.
2. When comparisons or evaluations are mentioned, include detailed quantitative results with specific numbers comparing against baseline methods.
3. Add experimental details such as ablation study results, hyperparameter analyses, and performance metrics to substantiate claims.
4. Maintain the paper's writing style, structure, and formatting. Preserve all existing tables, figures, and citations. Each section must start with its exact heading line.
5. Ensure revisions are coherent across sections and consistent with the overall paper context.

Return the revised sections in the same order as provided.

### F.3 Verify if Revision Addresses Reviewer's Concerns

#### *F.3.1 Verification Prompt (Without Location)*

**Verification Prompt (Blind to Location)**

You are evaluating how well a revised research paper addresses a specific flaw that was previously identified by reviewers. Your task is to provide a careful, critical, and professional assessment based on the actual content of the papers, not on the authors' claims alone.
**Evaluation Approach:** Authors may sometimes make revisions that sound plausible but do not fully resolve the underlying issue. They may describe additional experiments, analyses, or baselines without providing sufficient supporting evidence, or update wording without making substantive methodological or empirical changes. Therefore, you must verify what is actually present in the revised paper by comparing it directly with the original version.
**Check systematically for:**

1. Fabricated or Unrealistic Results
2. Incomplete or Vague Experimental Details

3. Cherry-Picking and Selective Reporting
4. Over-Exaggeration and Unsubstantiated Claims
5. Low-Effort and Superficial Changes
6. Lack of Genuine Understanding
7. Insufficient Depth and Rigor
8. Poor Integration and Coherence
9. Discrepancies between Claims and Evidence

**Scoring (1-9):** Assign a score from 1 to 9 based on how well the revisions address the flaw:

- **1: Invalid or Ignored.** The flaw is ignored, or results are statistically implausible. No credible evidence is provided.
- **2: Methodologically Void.** Claims lack supporting data (e.g., no $n$-values). Resolution is deferred to future work.
- **3: Superficial or Anomalous.** Revisions are cosmetic. Results lack expected variance or error analysis, indicating reporting anomalies.
- **4: Ambiguous Verification.** Evidence is present but vague; the generation of results is obscured. Content lacks depth.
- **5: Partially Reproducible.** Experimental context is provided, but critical reproducibility parameters (e.g., splits, hyperparameters) are omitted.
- **6: Generally Valid.** Methodology is described in broad terms. Expectations are met, though specific details for reproduction are absent.
- **7: Verifiable and Substantive.** Methods are transparent and results appear organic. Changes substantively resolve the flaw.
- **8: Rigorous and Transparent.** Detailed methodology is provided, including error bars and limitation analysis. Evidence strongly supports the resolution.
- **9: Fully Reproducible and Comprehensive.** Complete transparency is achieved (exact metrics, settings, data). The response fully resolves the issue.

**Assessment Output:** Provide:

- A numerical revision score from 1 to 9.

- A clear explanation (in a few sentences) identifying problems, commenting on depth/rigor, and summarizing if the fix is substantive or superficial.

**Input:**

```
Identified Flaw: {flaw_description}
Original Paper (with flaw): {original_paper}
Revised Paper: {revised_paper}
```

*F.3.2 Verification Prompt (With Location)*

**Verification Prompt (Location Aware)**

You are evaluating how well a revised research paper addresses a specific flaw that was previously identified by reviewers. Your task is to provide a careful, critical, and professional assessment based on the actual content of the papers, not on the authors' claims alone.
**Evaluation Approach:** Authors may sometimes make revisions that sound plausible but do not fully resolve the underlying issue. You must verify what is actually present in the revised paper by comparing it directly with the original version.
**Check systematically for:**

1. Fabricated or Unrealistic Results
2. Incomplete or Vague Experimental Details
3. Cherry-Picking and Selective Reporting
4. Over-Exaggeration and Unsubstantiated Claims
5. Low-Effort and Superficial Changes
6. Lack of Genuine Understanding
7. Insufficient Depth and Rigor
8. Poor Integration and Coherence
9. Discrepancies between Claims and Evidence

**Scoring (1-9):** Assign a score from 1 to 9 based on how well the revisions address the flaw:

- **1: Invalid or Ignored.** The flaw is ignored, or results are statistically implausible. No credible evidence is provided.

- **2: Methodologically Void.** Claims lack supporting data (e.g., no $n$-values). Resolution is deferred to future work.
- **3: Superficial or Anomalous.** Revisions are cosmetic. Results lack expected variance or error analysis, indicating reporting anomalies.
- **4: Ambiguous Verification.** Evidence is present but vague; the generation of results is obscured. Content lacks depth.
- **5: Partially Reproducible.** Experimental context is provided, but critical reproducibility parameters (e.g., splits, hyperparameters) are omitted.
- **6: Generally Valid.** Methodology is described in broad terms. Expectations are met, though specific details for reproduction are absent.
- **7: Verifiable and Substantive.** Methods are transparent and results appear organic. Changes substantively resolve the flaw.
- **8: Rigorous and Transparent.** Detailed methodology is provided, including error bars and limitation analysis. Evidence strongly supports the resolution.
- **9: Fully Reproducible and Comprehensive.** Complete transparency is achieved (exact metrics, settings, data). The response fully resolves the issue.

**Assessment Output:** Provide:

- A numerical revision score from 1 to 9.
- A clear explanation (in a few sentences) identifying problems, commenting on depth/rigor, and summarizing if the fix is substantive or superficial.

**Input:**

```
Identified Flaw: {flaw_description}

Specific Sections That Were Changed:
Original (flawed) content: {original_flawed_content}
Revised content: {revised_content}

Original Paper (with flaw): {original_paper}
Revised Paper: {revised_paper}
```

### F.4 Flaw Categorization Prompt

**Flaw Categorization Prompt**

You are an expert research analyst. Your task is to read a CSV file where each unique instance is defined by a (openreview_id, flaw_id) pair, and each flaw_description is a short paragraph summarizing flaws identified in a scientific paper's peer review process. You must categorize the described issues into one or more of the predefined categories below. Do not attempt to categorize by flaw_id only.
**Input Format:** The input will be a CSV file with at least: `openreview_id`, `flaw_id`, and `flaw_description`.
**Output Format:** Your output should be the original CSV file with one new column added: `category_ids`. This column should contain a comma-separated list of all applicable category IDs (e.g., "1b, 5a").
**Flaw Categories:**

- **Category 1: Empirical Evaluation Flaws** (1a: Insufficient Baselines, 1b: Weak Scope, 1c: No Ablation, 1d: Flawed Metrics)
- **Category 2: Methodological Flaws** (2a: Fundamental Limitation, 2b: Missing Theory, 2c: Technical Error)
- **Category 3: Positioning** (3a: Novelty, 3b: Overclaims)
- **Category 4: Presentation** (4a: Clarity, 4b: Reproducibility)
- **Category 5: Limitations** (5a: Unacknowledged, 5b: Ethics)

**Disambiguation Rules:**

- *Method vs. Experiment (2a vs 1b):* If the method is inherently limited by design, use 2a. If the method could have been tested more broadly but wasn't, use 1b.
- *Clarity vs. Error (4a vs 2c):* If confusing, 4a. If demonstrably incorrect, 2c.
- *Omission (5a):* If a paper has a flaw (e.g., 1b) and fails to mention it, include both 1b and 5a.