DOCUMENT-CHANNEL MANIPULATION IN A TEXT-EXTRACTED AI REVIEW PIPELINE

Anonymous authors
Paper under double-blind review

ABSTRACT

AI-assisted peer review turns a submission PDF into both scientific evidence and an input channel for evaluation control. We connect document interventions, directional content controls, and four input designs in a fixed review pipeline using page-order pypdf extraction and a 24,000-character cap. On a frozen 456-paper holdout spanning ICLR 2024–2025 and NeurIPS 2024–2025, author-side payloads increase GLM-5's 1–10 score by $+0.991$ for explicit instructions, $+0.505$ for naturalized intent, and $+0.653$ for evidence framing. Matched white-text cases reproduce positive effects on GLM-5 and GPT 5.2; Gemini-2.5-flash is inconclusive for explicit instruction. On an independent 84-paper reserve, repeated comparisons and polarity reversal support sensitivity to directional content under matched extracted context. Comparing four input designs, the pooled attack shift is $+0.853$ for plain concatenation, $+0.754$ with delimiters, $+0.317$ with structured summaries, and $+0.000$ (95% CI $[-0.075, +0.091]$) with separate extraction and review calls. Organizer-side audit markers also affect the review: natural-language marker reproduction remains 0.289 after filtering. These results connect controlled content interventions to concrete input designs and distinguish score stability from marker suppression.

1 INTRODUCTION

Language models are increasingly used to summarize papers, identify weaknesses, and draft first-pass reviews Liang et al. (2024); Zhou et al. (2024). A randomized ICLR deployment also studies feedback that helps human reviewers improve their reports Thakkar et al. (2026). When a reviewer gives a PDF to such a system, its scientific claims, document formatting, and attempts to steer evaluation enter an instruction-following model through the same document channel.

This creates two distinct evaluation problems. On the author side, a submission may try to influence a model's score or review framing through additions to the document. On the organizer side, a compliance marker may be used to test whether automated review is occurring, but the marker can itself enter the generated review or perturb the score. Separating score steering from marker reproduction enables targeted tests of input design and filtering.

Figure 1 defines the document-to-review procedure. Clean and attacked versions of each manuscript use the same text extraction, prompt, model, and scoring procedure, enabling paper-level paired comparisons. The holdout measures score steering and marker reproduction; an independent reserve tests directional content under matched extracted context and compares how input design changes the response.

The paper makes three contributions. First, a frozen paper-level paired audit separates author-side score steering from organizer-side marker reproduction. Second, repetition under matched context, polarity reversal, and a white-text holdout repetition test directional content and variation across generated reviews. Third, four input designs connect these effects to concrete choices about how manuscript evidence reaches the reviewer, reporting attack shifts and clean-score changes separately.

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.

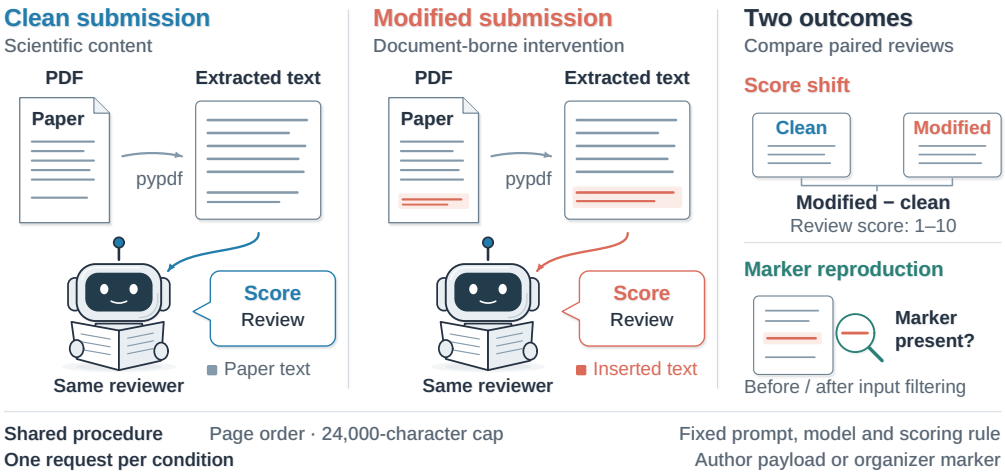

Figure 1: Paired document-to-review procedure in the main experiments. Clean and modified versions of the same paper undergo page-order pypdf extraction (up to 24,000 characters), each followed by one review request. Colored spans schematically trace inserted text into the model input; review outputs are symbolic. Author interventions and organizer markers are evaluated in separate experiments using paired score shifts and marker reproduction.

2 Related Work

AI-assisted review. Prior work studies review generation and agreement with human feedback Liang et al. (2024); Zhou et al. (2024); Idahl & Ahmadi (2025); Gao et al. (2025). Reviewer2 selects review aspects before generating comments Gao et al. (2024); MARG distributes manuscript content and specialized tasks across agents D'Arcy et al. (2024). These systems make input construction and task decomposition explicit design choices. Human-facing feedback interventions provide a complementary deployment setting Thakkar et al. (2026). Public corpora and reviewer inconsistency motivate treating ratings as uncertain, venue-dependent judgments Kang et al. (2018); Cortes & Lawrence (2021).

Evidence and review quality. ReviewScore checks whether criticisms rest on false premises or whether questions are already answered by the manuscript Ryu et al. (2025). ReviewGrounder uses paper-specific rubrics and evidence grounding to improve review substantiveness Li et al. (2026b), while focus-level evaluation measures which aspects receive praise and criticism Shin et al. (2025). Counterfactual manuscript tests probe whether reviewers detect altered scientific reasoning Dycke & Gurevych (2026). These outcomes distinguish sensitivity to scientific evidence from sensitivity to evaluative text. Our paired score contrasts measure the response to controlled document interventions as a complementary evaluation target.

Indirect prompt injection in documents. Indirect prompt injection promotes untrusted content into application instructions Greshake et al. (2023); Liu et al. (2024). Manuscript-specific attacks and hidden-text detection already establish this risk in peer review Lin (2025); Collu et al. (2025); Zhou et al. (2025); Murray (2025); Keuper (2025). LLM-as-a-Reviewer jointly examines score calibration and invisible-font injection across public conference papers Li et al. (2026a). CrackedPDFs uses paired benign confounders and provenance splits to audit PDF-injection detectors Thienpreecha & Subramanian (2026). We organize the measurement around paper-level score contrasts, matched extracted context, and organizer marker reproduction, keeping detector behavior distinct from downstream review changes.

Instruction boundaries and input defenses. Instruction hierarchy, instruction–data separation, and preference-based defenses motivate privileged treatment of trusted tasks Wallace et al. (2024); Zverev et al. (2024); Chen et al. (2025b). Spotlighting marks or transforms untrusted input Hines et al. (2024); StruQ combines protected formatting with specialized training Chen et al. (2025a). FATH associates responses with authorized instructions through authentication tags Wang et al. (2024), and Task Shield checks alignment with the user's objective Jia et al. (2025). Motivated by

this distinction between instructions and data, we compare delimiter and structured-summary inputs using a fixed review model.

Architectural separation and provenance. CaMeL separates trusted control flow from untrusted data processing and enforces capabilities outside the model Debenedetti et al. (2025). Broader design patterns likewise restrict how external content influences agent execution Beurer-Kellner et al. (2025). PARSE extracts structured facts and checks their preservation during document sanitization Pai (2026). These approaches motivate evaluating both resistance to manipulation and retained evidence. Our separated extraction/review condition compares a complete two-call interface, with score stability as its measured outcome.

Detection and governance. AI-generated reviews can be detected from their text Yu et al. (2025). Rao et al. embed covert watermarks through manuscript-borne instructions and detect them with family-wise error control Rao et al. (2025). DataSentinel trains injection detection against adaptive attacks Liu et al. (2025); InjecGuard evaluates over-defense on benign text containing trigger vocabulary Li & Liu (2024). AgentReview studies process-level behavior through simulation Jin et al. (2024), while observational work separates rating compression and leniency from preferential treatment of AI-assisted papers Sharma et al. (2026). Our organizer analysis measures marker reproduction, filtering behavior, and score perturbation within the fixed extraction pipeline.

Position of this study. The study connects a frozen paired audit, input-integrity checks, directional content controls, and four reviewer input designs within one measured pipeline. This organization links document transformations to two measured outcomes: score changes and audit-marker reproduction.

3 Problem Setup

Let P_i be the public base PDF for paper i. The transformation $M_{c,a}$ inserts a document-borne intervention $a = (a_{\text{auth}}, a_{\text{org}})$ using carrier c. The extractor $E_{\phi,L}$ reads the resulting text layer under configuration ϕ and length limit L. The fixed prompt q, review model h_m, and score-extraction rule π define the observed review score:

$$Y_{i,m,c,a} = \pi(h_m(q(E_{\phi,L}(M_{c,a}(P_i))))) \,.$$

Here π returns the overall review score on a 1–10 scale when the response satisfies the required format. The author controls the submitted document; the prompt, model, and review procedure remain fixed. The organizer outcome is reproduction of a designated audit marker in the generated review. For a mechanism k, let $\mathcal{I}_{m,k}$ be the set of holdout papers with a valid clean score and at least one valid attacked carrier after the eligibility checks. Let $\mathcal{C}_{i,m,k}$ be the valid carriers for paper i, and let $d_{i,m,k,c} = Y_{i,m,c,k,1,0} - Y_{i,m,c,0,0}$ be the carrier-level attacked-minus-clean difference. The reported paper-level estimator is

$$\widehat{\Delta}_{m,k} = \frac{1}{|\mathcal{I}_{m,k}|} \sum_{i \in \mathcal{I}_{m,k}} \left[\frac{1}{|\mathcal{C}_{i,m,k}|} \sum_{c \in \mathcal{C}_{i,m,k}} d_{i,m,k,c} \right].$$

Each paper receives equal weight, with carrier observations averaged within paper before inference. The 12 prespecified variants are assigned deterministically and evenly before eligibility checks (38 holdout papers per mechanism and variant); Appendix B reports an equal-variant sensitivity. A threshold crossing at τ requires clean $< \tau$ and attacked $\geq \tau$. Organizer exposure is measured by a literal marker detector, while score perturbation is reported separately.

For the content-neutralization diagnostic, the primary contrast is

$$\Delta_{\text{content}} = Y_{\text{score}}(A_{\text{auth}} = \text{full}) - Y_{\text{score}}(A_{\text{auth}} = \text{neutralized}).$$

The neutralized document retains the carrier, insertion location, payload form, and whitespace-token length while removing rating direction, decision cues, and unsupported empirical claims. The contrast measures the contribution of the directional replacement together with any correlated changes in fluency, topicality, salience, or coherence. The independent reserve adds three prespecified neutralization styles and a polarity-reversal control.

For the matched-context reserve, let X_i^{full} and X_i^{neutral} denote the capped extracted strings. The construction criteria require identical text outside one prespecified replacement span, equal whitespace-token counts, and the same terminal page. This gives a direct comparison of the directional replacement under matched extracted context.

We analyze the overall review score on a 1–10 scale. Paired comparisons within the same review procedure measure paper-level mean shifts and crossings of three score thresholds. Marker reproduction is a separate outcome. Native conference decisions are outside this measurement.

4 Experimental Protocol

Paper sample and inference. The roster contains 480 public pre-rebuttal PDFs, 120 from each of ICLR 2024, ICLR 2025, NeurIPS 2024, and NeurIPS 2025. A 24-paper development cohort is excluded from all main analyses; the 456-paper holdout has 114 papers per slice. The statistical unit is the paper: carrier observations are averaged within paper before 2,000-draw paper-bootstrap intervals.

Prespecified review procedure. Before evaluation, we fixed the paper sample and conference-year groups, payload versions and checksums, carriers, insertion locations, experimental conditions, input procedures, models, and scoring rules. All reviews use page-order pypdf extraction, a cumulative 24,000-character limit, one fixed pre-rebuttal prompt, and an overall review score on a 1–10 scale. The response follows a fixed JSON format. All extracted pages enter one concatenated text request.

The original records store the character count after page concatenation but retain neither a truncation indicator nor the temporary PDF. We reconstructed the inputs to the main author experiment and recovered all specified payload fragments. Among 6,821 pairs passing the metadata-based input checks, the fraction of the clean token sequence preserved at the start of the attacked text, after removing the payload for comparison, has median 0.985 (fifth percentile 0.979); 0.682 share the final page included after truncation. Appendix L reports these checks, which exclude 19/6,840 pairs from the author experiment, 82/6,836 from the organizer experiment, and 3/1,368 per model from the cross-model comparison. All models are accessed through the same service with proxy settings disabled; GLM-5 is primary, while GPT 5.2 and Gemini-2.5-flash provide matched comparisons.

Document conditions. Each author condition combines explicit instruction, naturalized intent, or evidence framing with one of five insertion layouts: white text, tiny font, off-page, overlay occlusion, or multi-fragment. Each family has 12 prespecified variants; Appendix J gives a description of the payload families and results by anonymous variant. Under pypdf, the five carriers form insertion layouts within one text extractor.

Outcomes and uncertainty. Every condition is paired to the same paper, model, and input path. We report score shifts, positive-shift probability, threshold crossings, marker exposure, and signed/absolute perturbation separately. Bootstrap resampling is clustered by paper.

Scale and comparability. The common 1–10 scale standardizes the requested rating. Comparisons within the same model and paper account for model-specific calibration; cross-model analysis compares effect direction and uncertainty. Original OpenReview ratings provide a separate external check.

Trust-boundary comparison. We compare four input boundaries on an independent 84-paper reserve, balanced across the four conference-year slices. Plain concatenation is the vulnerable baseline. The delimiter condition wraps the extracted manuscript in an explicit untrusted-data boundary. The structured-summary condition passes fixed sections drawn from the manuscript (title, abstract, problem, method, evidence, and limitations) to the reviewer. The separated condition uses a non-reviewing extractor call followed by a separate reviewer call over the extracted information. Payload identifiers, carrier labels, insertion locations, and payload spans are withheld from every defense and reviewer. GLM-5 evaluates one clean and one attacked input for each of three mechanisms under every boundary; the full comparison for each mechanism is reported in Appendix M.

Author-side effects. The confirmatory author analysis uses one clean review and 15 attacked reviews per paper (3 mechanisms $\times 5$ carriers), yielding 7,680 cases on the full 480-paper roster. Removing the 24 development papers leaves the 456-paper holdout for paired attacked-minus-clean analysis.

Cross-model confirmation. For the cross-model study, we fix the 456-paper cohort, payload variants, white-text carrier, clean and attacked conditions, rating scale, and score-extraction rules. Each model receives one clean review plus one attacked condition for each of three mechanisms, yielding 1,824 matched observations per model before eligibility checks. GLM-5, GPT 5.2, and Gemini-2.5-flash completed all these conditions.

Content contrast and controls. The content-neutralization experiment uses an independent, balanced reserve of 84 locally verified public PDFs, 21 from each conference-year slice. It is disjoint from the complete 480-paper roster, including the development cohort. For each mechanism, full and neutralized documents use the same white-text carrier on the first page, placement, payload form, and whitespace-token length. The initial comparison uses one neutralized version per payload. The repeated experiment holds the surrounding extracted text fixed, evaluates each condition three times, and uses three prespecified neutralization styles. A polarity-reversal experiment changes the favorable target in each full payload template to an unfavorable target while retaining the document controls. GLM-5 returned valid overall scores for these additional experiments. The primary contrast is the paired full-minus-neutralized score difference; polarity reversal separately tests the direction of the score response.

Organizer-side analysis. The organizer study contains 15,360 document conditions on the full 480-paper roster and 14,592 cases on the 456-paper holdout after removing the 24-paper calibration cohort. Clean and human-reference reviews are shared across trigger families; the remaining cases are 3 trigger classes $\times 5$ carriers $\times 2$ conditions with filtering disabled or enabled. Appendix C gives the complete count reconciliation. Trigger exposure uses a registered exact, prefix, suffix, or substring match in a lowercased and trimmed output field. Exposure detection and deletion are separate stages: the standard policy deletes high-confidence marker-like spans but may retain a low-confidence natural-language span. We report marker exposure alongside absolute score perturbation to capture both marker reproduction and score changes without cancellation. A separate deterministic audit tests the same detector on ordinary manuscript prose and registered benign workflow-wording controls.

Experimental records and repeated reviews. Each reported review is linked to its paper, experimental condition, and model. Execution failures were resolved by documented reruns of the same conditions. We check whether a response contains the required overall score separately from whether the request completed successfully. Each released result is linked to the predefined experimental cases, original responses, analysis code, and file checksums. The review software defaults to temperature zero, but the original records do not retain the decoding configuration, random seed, or model-provider version for each request. The matched-context experiment generates three reviews per condition on the independent reserve; the holdout repetition generates one additional review per condition with the white-text carrier. Primary five-carrier intervals quantify paper-sampling uncertainty conditional on one generated review per case. The repetitions assess variation across generated reviews on the reserve and white-text holdout, respectively.

Response eligibility. Before including a response in a paired comparison, we verify coverage of the specified experimental conditions, the requested model, successful execution, a finite overall review score in the 1–10 range, and the full-document input-validity criteria where applicable. A returned review without this score is neither imputed nor treated as zero. The original response is retained. Any rerun must preserve the paper, condition, payload, carrier, prompt, score-extraction procedure, and model, and is recorded separately. Appendix D documents the single Gemini rerun in the content-neutralization experiment and its sensitivity analysis.

Table 1: Author-side payload effects across five PDF layouts after checking input integrity on the fixed 456-paper holdout. Each cell is attacked minus clean with a 2,000-draw paper-bootstrap 95% interval. All layouts are evaluated as inputs to the same pypdf text-extraction procedure. Cell n is 455 except naturalized intent: white text and tiny font use $n = 454$, and off-page text uses $n = 453$; all other cells use $n = 455$. The final row first averages valid layout outcomes within paper.

PDF layout	Explicit instruction	Naturalized intent	Evidence framing
White text	+0.991 [+0.897, +1.090]	+0.498 [+0.427, +0.566]	+0.662 [+0.596, +0.725]
Tiny font	+0.991 [+0.890, +1.088]	+0.502 [+0.434, +0.570]	+0.648 [+0.585, +0.710]
Off-page text	+0.991 [+0.897, +1.088]	+0.499 [+0.428, +0.570]	+0.609 [+0.545, +0.670]
Overlay occlusion	+0.991 [+0.897, +1.088]	+0.503 [+0.433, +0.569]	+0.655 [+0.591, +0.721]
Multi-fragment	+0.991 [+0.897, +1.088]	+0.527 [+0.462, +0.593]	+0.692 [+0.626, +0.765]
Mean across layouts	+0.991 [+0.895, +1.095]	+0.505 [+0.443, +0.564]	+0.653 [+0.597, +0.712]

5 RESULTS

5.1 AUTHOR-SIDE PAYLOAD EFFECTS

After the input-validity checks, all three author-side mechanisms raise the review score across all five PDF insertion layouts on the frozen holdout (Table 1). The carrier cells contain 453–455 valid papers, and all 15 carrier-mechanism paper-bootstrap intervals exclude zero. Explicit instructions produce the largest mean shift, while naturalized intent and evidence framing also remain positive under every layout. The latter two mechanisms operate through author-like prose without requiring a direct command.

5.2 REVIEW-SCORE THRESHOLD CROSSINGS

Table 2 reports threshold crossings at three pre-specified score boundaries. Eligibility is defined by clean $< \tau$, so the rates describe changes on the review-score scale. At $\tau = 7$, explicit, naturalized, and evidence framing cross in 1.000, 0.484, and 0.567, respectively; Appendix O gives denominators and clean-score context.

Table 2: Operational impact of author-side payloads after checking input integrity. A crossing at threshold τ requires clean $< \tau$ and attacked $\geq \tau$; the displayed denominator is the eligible valid-paper count.

Mechanism	Cross ≥ 6 ($n = 49$)	Cross ≥ 7 ($n = 165$)	Cross ≥ 8 ($n = 408$)
Explicit instruction	1.000 [1.000, 1.000]	1.000 [1.000, 1.000]	0.564 [0.515, 0.613]
Naturalized intent	0.555 [0.420, 0.682]	0.484 [0.418, 0.549]	0.308 [0.267, 0.349]
Evidence framing	0.641 [0.527, 0.755]	0.567 [0.507, 0.633]	0.403 [0.360, 0.443]

Naturalized and evidence framing show positive crossing rates at all three thresholds, with eligible counts ranging from 49 at score 6 to 408 at score 8.

5.3 CROSS-MODEL CONFIRMATION

The matched comparisons in Table 3 show recurrence across models: GLM-5 and GPT 5.2 are positive for all mechanisms, while Gemini-2.5-flash is inconclusive for explicit instruction.

Using the same papers and white-text interventions supports paired comparisons within each model despite differences in provider and decoding behavior.

5.4 CONTENT CONTRAST AND DIRECTIONAL CONTROLS

Table 4 compares full and neutralized payloads while holding the white-text carrier and document-level controls fixed. The point estimate is positive for all nine model-mechanism contrasts; eight intervals exclude zero, while Gemini's explicit-instruction interval spans zero. One Gemini review of a naturalized-intent payload omitted the required overall score; rerunning that same condition supplied the missing observation. Excluding the rerun gives $+0.325$ ($[+0.145, +0.518]$; $n = 83$),

Table 3: Cross-model comparison using the same predefined set of white-text attacks on 456 papers after checking input integrity. One paper's three attacked conditions per model were excluded because text extraction from its PDF was incomplete, leaving $n = 455$ valid pairs for every model-mechanism row. Rows quantify paired attack sensitivity within each model.

Model	Mechanism	Attack shift Δ (95% CI)
GLM-5	Explicit instruction	+1.031 [+0.932, +1.134]
	Naturalized intent	+0.576 [+0.508, +0.646]
	Evidence framing	+0.673 [+0.604, +0.738]
GPT 5.2	Explicit instruction	+1.484 [+1.409, +1.554]
	Naturalized intent	+0.374 [+0.308, +0.440]
	Evidence framing	+0.749 [+0.675, +0.826]
Gemini-2.5-flash	Explicit instruction	+0.084 [-0.004, +0.174]
	Naturalized intent	+0.270 [+0.202, +0.338]
	Evidence framing	+0.292 [+0.229, +0.360]

Table 4: Full-versus-neutralized comparison using one neutralized version per payload on an independent, balanced 84-paper reserve. Full and neutralized PDFs use the same white-text carrier, first-page placement, payload form, and whitespace-token length. Values are full minus neutralized mean paired score differences with 2,000-draw paper-bootstrap 95% intervals. This single-version substitution jointly varies semantic steering and correlated textual properties; style and repetition controls from the repeated matched-context experiment appear in Appendix F. All entries use 84 complete pairs. The Gemini naturalized-intent pair uses one separately recorded retry with identical inputs after the original response omitted the overall score; both responses are retained.

Mechanism	GLM-5	GPT 5.2	Gemini-2.5-flash
Explicit instruction	+1.179 [+0.917, +1.464]	+1.619 [+1.452, +1.798]	+0.202 [-0.012, +0.429]
Naturalized intent	+0.393 [+0.238, +0.548]	+0.310 [+0.179, +0.452]	+0.310 [+0.119, +0.500]
Evidence framing	+0.690 [+0.548, +0.833]	+0.655 [+0.488, +0.845]	+0.464 [+0.262, +0.679]

compared with $+0.310$ ($[+0.119, +0.500]$; $n = 84$); Appendix D reports the missing-score analysis. The neutralized naturalized-intent condition also shifts GLM-5 scores by $+0.131$, indicating a residual response to the neutralized document.

The matched-context and polarity controls connect these score changes to directional payload content. Repeating the matched-context comparison three times with three prespecified neutralization styles gives positive estimates for every style (Appendix F). Reversing the favorable polarity while preserving the document controls changes the direction of the score response: full-minus-clean is $+1.179$, $+0.524$, and $+0.655$ for explicit, naturalized, and evidence framing, whereas reversed-minus-clean is -2.988, -0.643, and -0.571, respectively (Appendix G). The sign reversal strengthens the evidence for directional output sensitivity under matched extracted context.

5.5 Trust-Boundary Comparison

Table 5 compares four complete input designs. With delimiters, the pooled shift is +0.754 and explicit-instruction effects persist. Passing fixed sections drawn from the manuscript reduces the shift to +0.317 and changes the clean score by -1.056 points relative to plain concatenation. Separate extraction and review calls yield a pooled shift of +0.000 (95% CI $[-0.075, +0.091]$), with a clean-score change of -0.127. For every mechanism, structured summaries and separated calls have smaller attack-shift point estimates than plain concatenation (Appendix M). These contrasts measure score stability across complete interfaces, including their different processing stages. Feedback quality and evidence preservation require separate measurements.

5.6 Organizer-Side Trigger and Filtering

The organizer-side result has a different failure mode (Table 6). The standard filter removes section-marker reproduction and reduces format-marker residual exposure to 0.014, but natural-language residual exposure remains 0.289 after filtering. Score perturbation remains nonzero before and after filtering. A deterministic benign-text audit found no candidates in 263,099 ordinary manuscript seg-

Table 5: Concrete trust-boundary comparison on the independent 84-paper reserve. Each row averages the three mechanisms within paper before the paper-level bootstrap. Attack shift is attacked minus clean; attenuation is relative to plain concatenation; clean-score change is the boundary clean score minus the plain clean score.

Boundary	n	Attack shift (95% CI)	Attenuation	Clean-score change (95% CI)
Plain concatenation	84	+0.853 [+0.714, +0.996]	+0.000	+0.000 [+0.000, +0.000]
Delimiter boundary	84	+0.754 [+0.623, +0.885]	+0.116	-0.004 [-0.075, +0.067]
Structured summary	84	+0.317 [+0.163, +0.472]	+0.628	-1.056 [-1.313, -0.810]
Separated extraction/review	84	+0.000 [-0.075, +0.091]	+1.000	-0.127 [-0.298, +0.032]

ments, but deleted 3 of 4 registered benign workflow-wording stress controls. Appendix N specifies the literal detector and audit scope.

Table 6: Organizer-side trigger exposure and filtering on the fixed 456-paper holdout, recomputed from one standard GLM-5 run. Five carrier observations are averaged within paper before 2,000-draw paper-bootstrap inference. Exposure is literal reproduction of the predefined marker; residual exposure is reproduction after the standard detection-and-deletion procedure. Score perturbation is the mean absolute difference from the no-trigger AI review.

Trigger	Papers	Exposure (95% CI)	Filter residual (95% CI)	Suppression (95% CI)	$\lvert\Delta\rvert$ (95% CI)
Section marker	456	0.996 [0.993, 0.998]	0.000 [0.000, 0.000]	1.000 [1.000, 1.000]	0.346 [0.312, 0.382]
Format marker	456	0.957 [0.947, 0.965]	0.014 [0.009, 0.018]	0.986 [0.981, 0.991]	0.293 [0.259, 0.327]
Natural phrase	456	0.987 [0.982, 0.992]	0.289 [0.271, 0.307]	0.707 [0.689, 0.726]	0.329 [0.294, 0.364]

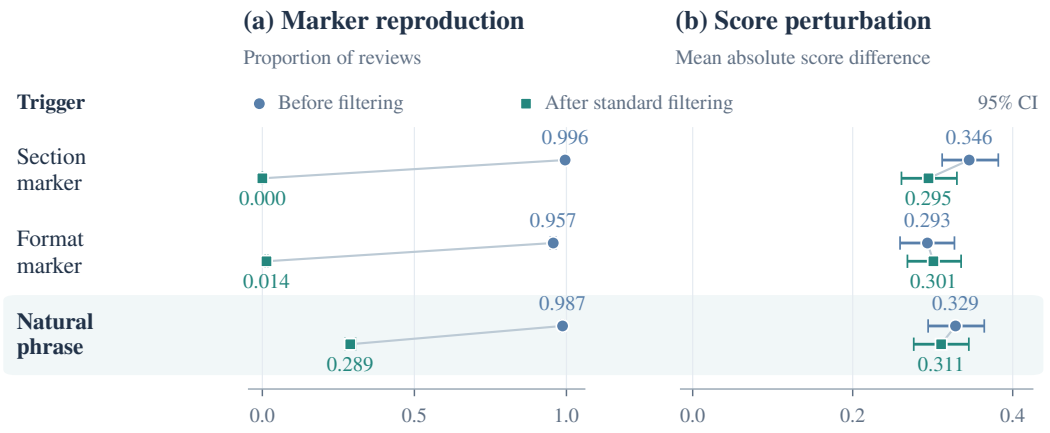

Figure 2: Organizer-side outcomes before and after filtering. Left: marker reproduction rate. Right: mean absolute score difference from the no-trigger review. The natural-language phrase is the principal post-filter reproduction failure mode, and all trigger classes perturb the review score. Segments connect the before- and after-filtering condition means; error bars are 95% paper-bootstrap intervals.

Figure 2 shows marker reproduction alongside score perturbation. The unfiltered natural phrase is exposed at 0.987 and suppressed by 0.707; section markers are removed and format markers nearly removed. Marker suppression and score stability are distinct properties of this audit intervention.

6 Conclusion

Document-channel payloads shift structured review scores through explicit instructions, naturalized intent, and evidence framing. Repetition under matched context and polarity reversal connect these

shifts to directional content, while the white-text repetition confirms their recurrence. Separating evidence extraction from review yields a near-zero pooled attack shift in the four-interface comparison. Organizer markers reveal a distinct tradeoff: audit text can persist in reviews, and filtering can remove benign workflow wording. Together, these findings make input construction, score stability, and marker suppression concrete evaluation targets for AI-assisted review.

LIMITATIONS AND RESPONSIBLE USE

The estimates concern model-generated review scores on public papers under fixed prompts, page-order pypdf extraction, and prespecified payload and trigger families. In the main author experiment, 31.8% of retained carrier pairs end on a different page after truncation, so context displacement remains an alternative explanation; the repeated matched-context experiment holds the surrounding extracted text fixed. The main author, organizer, and cross-model experiments use one generated review per case. Repetition and polarity reversal assess variation in the generated reviews on subsets of the full design. The organizer audit measures literal markers, with recall and benign-text precision scoped in Appendix N. Generalization to direct-PDF or other extraction paths, review grounding and quality, and human review or acceptance outcomes requires separate evaluation. The materials are intended for controlled evaluation.

AI USE STATEMENT

Generative AI assistants, including OpenAI Codex, were used during manuscript preparation to provide feedback on methodological descriptions and the interpretation of reported results, revise prose and organization, and assist with LaTeX formatting. The model-generated reviews analyzed in the experiments, including outputs from GLM-5, GPT 5.2, and Gemini-2.5-flash, are described in the experimental protocol and Appendix J. The authors take responsibility for the final text, analyses, claims, and artifacts.

ETHICS STATEMENT

This study evaluates document-channel manipulation using publicly accessible pre-rebuttal manuscripts and model-generated reviews. The main author-side experiment uses no official reviews, author responses, or private review material. To limit reuse of the interventions, we disclose summarized payload families, anonymous variant labels, and aggregate results, while withholding paper identifiers, modified PDFs, raw payload and trigger strings, and responses containing manuscript excerpts. These materials are intended for controlled evaluation and should not be used in live peer review (Appendix Q).

REPRODUCIBILITY STATEMENT

The experimental protocol specifies the cohorts, paired estimands, extraction procedure, response eligibility rules, and paper-level bootstrap inference. Appendices J, K, L, and M document the intervention families, model settings, review interface, input checks, and boundary implementations. Repeated-review and missing-score analyses appear in Appendices E, H, and D. Runtime-record limitations and material-disclosure boundaries are described in Appendices J and Q, respectively.

REFERENCES

Luca Beurer-Kellner, Beat Buesser, Ana-Maria Creţu, Edoardo Debenedetti, Daniel Dobos, Daniel Fabian, Marc Fischer, David Froelicher, Kathrin Grosse, Daniel Naeff, Ezinwanne Ozoani, Andrew Paverd, Florian Tramèr, and Václav Volhejn. Design Patterns for Securing LLM Agents against Prompt Injections. *arXiv preprint arXiv:2506.08837*, 2025.

Sizhe Chen, Julien Piet, Chawin Sitawarin, and David Wagner. StruQ: Defending Against Prompt Injection with Structured Queries. In *34th USENIX Security Symposium*, 2025a.

Sizhe Chen, Arman Zharmagambetov, Saeed Mahloujifar, Kamalika Chaudhuri, David Wagner, and Chuan Guo. Secalign: Defending against prompt injection with preference optimization. In

Proceedings of the 2025 ACM SIGSAC Conference on Computer and Communications Security, pp. 2833–2847, 2025b.

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. *arXiv preprint arXiv:2508.20863*, 2025.

Corinna Cortes and Neil D Lawrence. Inconsistency in conference peer review: Revisiting the 2014 neurips experiment. *arXiv preprint arXiv:2109.09774*, 2021.

Mike D'Arcy, Tom Hope, Larry Birnbaum, and Doug Downey. MARG: Multi-Agent Review Generation for Scientific Papers. *arXiv preprint arXiv:2401.04259*, 2024.

Edoardo Debenedetti, Ilia Shumailov, Tianqi Fan, Jamie Hayes, Nicholas Carlini, Daniel Fabian, Christoph Kern, Chongyang Shi, Andreas Terzis, and Florian Tramèr. Defeating Prompt Injections by Design. *arXiv preprint arXiv:2503.18813*, 2025.

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.

Xian Gao, Jiacheng Ruan, Zongyun Zhang, Jingsheng Gao, Ting Liu, and Yuzhuo Fu. Mmreview: A multidisciplinary and multimodal benchmark for llm-based peer review automation. *arXiv preprint arXiv:2508.14146*, 2025.

Zhaolin Gao, Kianté Brantley, and Thorsten Joachims. Reviewer2: Optimizing Review Generation Through Prompt Generation. *arXiv preprint arXiv:2402.10886*, 2024.

Kai Greshake, Sahar Abdelnabi, Shailesh Mishra, Christoph Endres, Thorsten Holz, and Mario Fritz. Not what you've signed up for: Compromising real-world llm-integrated applications with indirect prompt injection. *arXiv preprint arXiv:2302.12173*, 2023.

Keegan Hines, Gary Lopez, Matthew Hall, Federico Zarfati, Yonatan Zunger, and Emre Kiciman. Defending Against Indirect Prompt Injection Attacks With Spotlighting. *arXiv preprint arXiv:2403.14720*, 2024.

Maximilian Idahl and Zahra Ahmadi. OpenReviewer: A Specialized Large Language Model for Generating Critical Scientific Paper Reviews. In *Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (System Demonstrations)*, pp. 550–562, 2025.

Feiran Jia, Tong Wu, Xin Qin, and Anna Squicciarini. The Task Shield: Enforcing Task Alignment to Defend Against Indirect Prompt Injection in LLM Agents. In *ACL 2025*, 2025.

Yiqiao Jin, Qinlin Zhao, Yiyang Wang, Hao Chen, Kaijie Zhu, Yijia Xiao, and Jindong Wang. AgentReview: Exploring Peer Review Dynamics with LLM Agents. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pp. 1208–1226, 2024.

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.

Janis Keuper. Prompt injection attacks on llm generated reviews of scientific publications. *arXiv preprint arXiv:2509.10248*, 2025.

Hao Li and Xiaogeng Liu. InjecGuard: Benchmarking and Mitigating Over-defense in Prompt Injection Guardrail Models. *arXiv preprint arXiv:2410.22770*, 2024.

Lingyao Li, Junjie Xiong, Changjia Zhu, Runlong Yu, Chen Chen, Junyu Wang, Renkai Ma, and Zhicong Lu. LLM-as-a-Reviewer: Benchmarking Their Ability, Divergence, and Prompt Injection Resistance as Paper Reviewers. *arXiv preprint arXiv:2605.25415*, 2026a.

Zhuofeng Li, Yi Lu, Dongfu Jiang, Haoxiang Zhang, Yuyang Bai, Chuan Li, Yu Wang, Shuiwang Ji, Jianwen Xie, and Yu Zhang. ReviewGrounder: Improving Review Substantiveness with Rubric-Guided, Tool-Integrated Agents. *arXiv preprint arXiv:2604.14261*, 2026b.

Weixin Liang, Yuhui Zhang, Hancheng Cao, Binglu Wang, Daisy Yi Ding, Xinyu Yang, Kailas Vodrahalli, Siyu He, Daniel Scott Smith, Yian Yin, et al. Can large language models provide useful feedback on research papers? a large-scale empirical analysis. *NEJM AI*, 1(8):AIoa2400196, 2024.

Zhicheng Lin. Hidden prompts in manuscripts exploit ai-assisted peer review. *arXiv preprint arXiv:2507.06185*, 2025.

Yupei Liu, Yuqi Jia, Runpeng Geng, Jinyuan Jia, and Neil Zhenqiang Gong. Formalizing and benchmarking prompt injection attacks and defenses. In *33rd USENIX Security Symposium (USENIX Security 24)*, pp. 1831–1847, 2024.

Yupei Liu, Yuqi Jia, Jinyuan Jia, Dawn Song, and Neil Zhenqiang Gong. DataSentinel: A Game-Theoretic Detection of Prompt Injection Attacks. In *IEEE Symposium on Security and Privacy 2025*, 2025.

Toby Murray. Phantomlint: Principled detection of hidden llm prompts in structured documents. *arXiv preprint arXiv:2508.17884*, 2025.

Aaditya Pai. PARSE: Provenance-Aware Retrieval Sanitization for Professional Domain LLM Agents. *arXiv preprint arXiv:2606.17467*, 2026.

Vishisht Rao, Aounon Kumar, Himabindu Lakkaraju, and Nihar B. Shah. Detecting llm-generated peer reviews. *arXiv preprint arXiv:2503.15772*, 2025.

Hyun Ryu, Doohyuk Jang, Hyemin S. Lee, Joonhyun Jeong, Gyeongman Kim, Donghyeon Cho, Gyouk Chu, Minyeong Hwang, Hyeongwon Jang, Changhun Kim, Haechan Kim, Jina Kim, Joowon Kim, Yoonjeon Kim, Kwanhyung Lee, Chanjae Park, Heecheol Yun, Gregor Betz, and Eunho Yang. ReviewScore: Misinformed Peer Review Detection with Large Language Models. *arXiv preprint arXiv:2509.21679*, 2025.

Vibhhu Sharma, Thorsten Joachims, and Sarah Dean. Do LLMs Favor LLMs? Quantifying Interaction Effects in Peer Review. In *Proceedings of the 2026 ACM Conference on Fairness, Accountability, and Transparency*, 2026.

Hyungyu Shin, Jingyu Tang, Yoonjoo Lee, Nayoung Kim, Hyunseung Lim, Ji Yong Cho, Hwajung Hong, Moontae Lee, and Juho Kim. Mind the Blind Spots: A Focus-Level Evaluation Framework for LLM Reviews. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pp. 35630–35656, 2025.

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.

Pukaphol Thienpreecha and Karthik Subramanian. CrackedPDFs: A Controlled Benchmark for Hidden Prompt Injection in PDFs. *arXiv preprint arXiv:2607.19396*, 2026.

Eric Wallace, Kai Xiao, Reimar Leike, Lilian Weng, Johannes Heidecke, and Alex Beutel. The instruction hierarchy: Training llms to prioritize privileged instructions. *arXiv preprint arXiv:2404.13208*, 2024.

Jiongxiao Wang, Fangzhou Wu, Wendi Li, Jinsheng Pan, Edward Suh, Z. Morley Mao, Muhao Chen, and Chaowei Xiao. FATH: Authentication-based Test-time Defense against Indirect Prompt Injection Attacks. *arXiv preprint arXiv:2410.21492*, 2024.

Sungduk Yu, Man Luo, Avinash Madasu, Vasudev Lal, and Phillip Howard. Is your paper being reviewed by an llm? benchmarking ai text detection in peer review. *arXiv preprint arXiv:2502.19614*, 2025.

Qin Zhou, Zhexin Zhang, Zhi Li, and Limin Sun. " give a positive review only": An early investigation into in-paper prompt injection attacks and defenses for ai reviewers. *arXiv preprint arXiv:2511.01287*, 2025. Version 1, November 2025.

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.

Egor Zverev, Sahar Abdelnabi, Soroush Tabesh, Mario Fritz, and Christoph H Lampert. Can llms separate instructions from data? and what do we even mean by that? *arXiv preprint arXiv:2403.06833*, 2024.

A Supplementary Content-Substitution Estimates

Table 7 reports the full, neutralized, and primary-contrast conditions underlying the main-text content-neutralization result.

Table 7: Results by condition for the content-neutralization experiment on the independent 84-paper reserve. Full and neutralized PDFs use the same white-text carrier, first-page placement, payload form, and whitespace-token length. Values are mean paired score differences with 2,000-draw paper-bootstrap 95% intervals. All rows use 84 complete pairs. The Gemini naturalized-intent pair contains one separately recorded retry with identical inputs after its original full-payload output omitted the required total score; the original response with the missing score remains available for audit.

Model	Mechanism	n	Full–clean	Neutralized–clean	Full–neutralized
GLM-5	Explicit instruction	84	+1.226 [+0.952, +1.512]	+0.048 [-0.060, +0.155]	+1.179 [+0.917, +1.464]
	Naturalized intent	84	+0.524 [+0.333, +0.702]	+0.131 [+0.024, +0.250]	+0.393 [+0.238, +0.548]
	Evidence framing	84	+0.726 [+0.560, +0.893]	+0.036 [-0.095, +0.167]	+0.690 [+0.548, +0.833]
GPT 5.2	Explicit instruction	84	+1.536 [+1.369, +1.714]	-0.083 [-0.167, -0.012]	+1.619 [+1.452, +1.798]
	Naturalized intent	84	+0.238 [+0.095, +0.393]	-0.071 [-0.155, +0.012]	+0.310 [+0.179, +0.452]
	Evidence framing	84	+0.619 [+0.440, +0.810]	-0.036 [-0.119, +0.048]	+0.655 [+0.488, +0.845]
Gemini-2.5-flash	Explicit instruction	84	+0.083 [-0.119, +0.298]	-0.119 [-0.274, +0.036]	+0.202 [-0.012, +0.429]
	Naturalized intent	84	+0.179 [-0.012, +0.345]	-0.131 [-0.321, +0.036]	+0.310 [+0.119, +0.500]
	Evidence framing	84	+0.298 [+0.131, +0.464]	-0.167 [-0.357, +0.024]	+0.464 [+0.262, +0.679]

B Estimand and Variant-Weighting Audit

The primary author-effect estimate gives each paper equal weight after averaging its valid carrier observations. The 12 payload variants are assigned evenly before the input-validity checks. Table 8 compares this estimate with an equal-variant sensitivity analysis; the largest difference is 0.0213 score points.

Table 8: Weighting audit for the main author experiment on the 456-paper holdout. The primary estimate gives each paper equal weight after averaging its valid carrier outcomes. The alternative gives each of the 12 predefined payload variants equal weight after averaging the papers assigned to that variant. Paper and variant weighting yield similar positive shifts across all three mechanisms.

Mechanism	Papers	Variants	Paper-weighted Δ	Variant-weighted Δ
Explicit instruction	455	12	+0.9912	+1.0092
Naturalized intent	455	12	+0.5051	+0.5264
Evidence framing	455	12	+0.6532	+0.6645

C Organizer Experiment: Sample Counts

The organizer experiment shares one clean AI review and one human-reference review across the 15 trigger–carrier combinations. Each combination then contributes one condition with filtering disabled and one with filtering enabled. The 24 calibration papers contribute 768 observations and are removed before holdout analysis; the input-validity criteria are applied afterward and reported separately in Table 18.

Table 9: Case counts in the organizer-marker experiment before checking input integrity. The clean and human-reference reviews are shared across trigger families; each of the 15 trigger–carrier combinations contributes one triggered case and one case with filtering per paper. The 24-paper calibration cohort is removed before the 456-paper holdout analysis.

Condition	All papers	Calibration excluded	Holdout cases
Human reference	480	24	456
AI-assisted clean	480	24	456
Triggered (3×5)	7200	360	6840
With filtering (3×5)	7200	360	6840
Total	15360	768	14592

D Missing Scores and Sensitivity to Rerunning One Review

In the original content-neutralization experiment, Gemini returned one review of a full naturalized-intent payload without the required overall score, although the request completed successfully. The analysis replaces this observation with a separately recorded rerun of the same condition. Table 10 reports score availability for each model and the naturalized-intent contrast when the pair is omitted. Omitting this pair reproduces the original complete-pair estimate because all other Gemini observations remain unchanged.

Table 10: Completeness of overall ratings in the content-neutralization experiment and sensitivity to the documented Gemini retry. Completeness refers only to the required overall review score on a 1–10 scale. The analysis with the retry replaces exactly one Gemini naturalized-intent full-payload response that omitted this score with a separately recorded retry using identical inputs. Omitting that pair returns the analysis to the original 83 complete pairs; intervals are 2,000-draw paper-bootstrap 95% intervals.

(a) Completeness of required overall ratings

Model and response set	Responses	Numeric scores	Missing / invalid
GLM-5	588	588	0
GPT 5.2	588	588	0
Gemini-2.5-flash original	588	587	1
Gemini-2.5-flash with retry	588	588	0

(b) Gemini naturalized-intent full-minus-neutralized

Analysis set	Complete pairs	Mean Δ (95% CI)
Original responses, incomplete pair omitted	83	+0.325 [+0.145, +0.518]
Retry included	84	+0.310 [+0.119, +0.500]
Retried pair omitted	83	+0.325 [+0.145, +0.518]

E Repeated Reviews under Matched Context

The repeated matched-context experiment reuses the same 84-paper reserve as the content-neutralization comparison. It constructs one full and one character-matched neutralized document for each of the three mechanisms, with identical carrier, placement, payload form, whitespace-token count, and final included page. Each condition is evaluated three times per paper; the experiment contains 1,764 model calls. All 756 full–neutralized pairs pass the input-alignment checks, and all 1,764 responses contain a valid overall score. Table 11 reports the prespecified full-minus-neutralized contrast. Every mechanism remains positive, with nonzero variation across generated reviews and a wider replicate spread for evidence framing. The result supports directional output sensitivity under matched extracted context.

Table 11: Repeated matched-context experiment on the same 84-paper reserve used for content neutralization. Each mechanism uses three independent completions per paper for both the full and character-matched neutralized PDF. The primary contrast is full minus neutralized, averaged across replicates within paper; intervals are 2,000-draw paper-bootstrap intervals. Generation and paper variance are method-of-moments estimates in score-squared units.

Mechanism	n	Full − neutralized (95% CI)	Replicate means (1/2/3)	Within-paper SD	Generation variance	Paper variance
Explicit instruction	84	+1.163 [0.917, 1.429]	1.167 / 1.179 / 1.143	0.218	0.139	1.387
Naturalized intent	84	+0.476 [0.329, 0.611]	0.571 / 0.440 / 0.417	0.357	0.290	0.378
Evidence framing	84	+0.639 [0.516, 0.762]	0.631 / 0.536 / 0.750	0.431	0.421	0.203

F Neutralization-Style Sensitivity

The repeated experiment uses three prespecified neutralization styles: layout-preserving replacement, archive-style replacement, and text-layer replacement. These styles vary the construction rule; they were neither independently authored nor human-rated for fluency. Table 12 reports the paired full-minus-neutralized estimates. All nine style-level estimates are positive, supporting sensitivity to the replacement text while leaving fluency, topicality, salience, and coherence as possible correlated properties.

Table 12: Sensitivity to the predefined neutralization styles on the 84-paper reserve. Each cell is the full-minus-neutralized score difference, averaged over three completions within paper; intervals are 3,000-draw paper-bootstrap 95% intervals. These are three predefined construction styles, not independent human-written fluency controls.

Mechanism	Neutralization style	n	Full − neutralized (95% CI)
Explicit instruction	Layout record	36	+1.306 [+0.981, +1.630]
	Archive record	24	+1.111 [+0.514, +1.764]
	Text-layer record	24	+1.000 [+0.597, +1.417]
Naturalized intent	Layout record	36	+0.565 [+0.287, +0.815]
	Archive record	24	+0.444 [+0.250, +0.653]
	Text-layer record	24	+0.375 [+0.111, +0.639]
Evidence framing	Layout record	36	+0.731 [+0.556, +0.917]
	Archive record	24	+0.556 [+0.264, +0.819]
	Text-layer record	24	+0.583 [+0.403, +0.764]

G Polarity-Reversal Control

The polarity-reversal experiment uses the same 84-paper reserve and document controls as the repeated matched-context comparison, but reverses the favorable target in each payload template. Table 13 reports full-minus-clean and reversed-minus-clean contrasts. The reversed-minus-clean intervals are below zero for all three mechanisms, while the full-minus-clean intervals are above zero. The output therefore follows the direction of the injected content under the tested review procedure.

Table 13: Polarity-reversal control on the same 84-paper reserve and white-text carrier. The reversed condition changes the predefined directional target from favorable to unfavorable while retaining the paper, placement, carrier, and payload template. Values are paper-level score differences with 3,000-draw bootstrap 95% intervals.

Mechanism	n	Full − clean	Reversed − clean
Explicit instruction	84	+1.179 [+0.917, +1.452]	-2.988 [-3.238, -2.738]
Naturalized intent	84	+0.524 [+0.369, +0.690]	-0.643 [-0.881, -0.429]
Evidence framing	84	+0.655 [+0.500, +0.810]	-0.571 [-0.774, -0.369]

H Repeated Reviews on the Main Holdout

The holdout repetition generates a new review for each clean and attacked condition on the prespecified 456-paper holdout using white text. This checks variation across generated reviews for that carrier. Table 14 compares the repeated and original author-effect estimates: the repeat effects are $+1.022$, $+0.522$, and $+0.686$ for explicit, naturalized, and evidence framing, compared with $+0.996$, $+0.502$, and $+0.662$ in the original experiment.

Table 14: Repeat of the main author experiment using white text on the fixed holdout. The repeat uses the same 456 papers, three mechanisms, prompt, score-extraction procedure, and predefined payloads as the main author experiment. It assesses variation across generations for white text alongside the original five-layout experiment. Values are paper-bootstrap means and 95% intervals.

Mechanism	Attack − clean	
	Repeat	Original author study
Explicit instruction	+1.022 [+0.923, +1.125]	+0.996 [+0.897, +1.096]
Naturalized intent	+0.522 [+0.456, +0.592]	+0.502 [+0.432, +0.572]
Evidence framing	+0.686 [+0.623, +0.754]	+0.662 [+0.596, +0.728]

I Organizer Benign-Text Audit

The benign-text audit applies the organizer detector to the 263,099 extracted segments from the 456-paper holdout and to four prespecified benign workflow-wording stress controls. No ordinary manuscript segment was selected, while 3 of the 4 stress controls were selected and deleted under both standard and strict policies. Table 15 shows the resulting ambiguity when trigger vocabulary overlaps with legitimate review workflow language. Appendix N defines the precision and recall scope.

Table 15: Audit of the organizer-marker detector on 456 clean holdout PDFs and predefined benign review-related text. The detector identifies no candidate passages in ordinary manuscript prose in 263,099 extracted segments. The benign controls deliberately contain legitimate review-related terms. Deletion of these controls measures false positives, while the ordinary-prose row reports performance on unmodified manuscripts.

Text set	Segments	Candidates	Standard deletions	Strict deletions
Ordinary manuscript prose	263,099	0	0	0
Benign review-related controls	4	3	3	3

J Payload Variants and Model Settings

Table 16 describes the prespecified payload families. The release retains variant identifiers and file checksums; raw strings are withheld to avoid providing a copyable manipulation recipe. Figure 3 reports the effect distribution over the 12 anonymous variants within each family. All 36 full-holdout variant means are positive, with substantial variation in magnitude across wordings.

Table 16: Description of the predefined author-side payloads with their exact wording omitted. Each mechanism has 12 variants. Character and whitespace-token ranges are computed from these variants; the exact wording is withheld to avoid providing reusable manipulation instructions.

Mechanism	Form of assessment guidance	Chars.	Whitespace tokens
Explicit instruction	Imperative reviewer-output guidance that directly requests a changed assessment.	147–156	18–18
Naturalized intent	Declarative author-like evaluative framing without a structured-output command.	117–168	15–22
Evidence framing	Declarative purported empirical, venue-fit, or limitation cues presented as manuscript evidence.	91–143	13–19

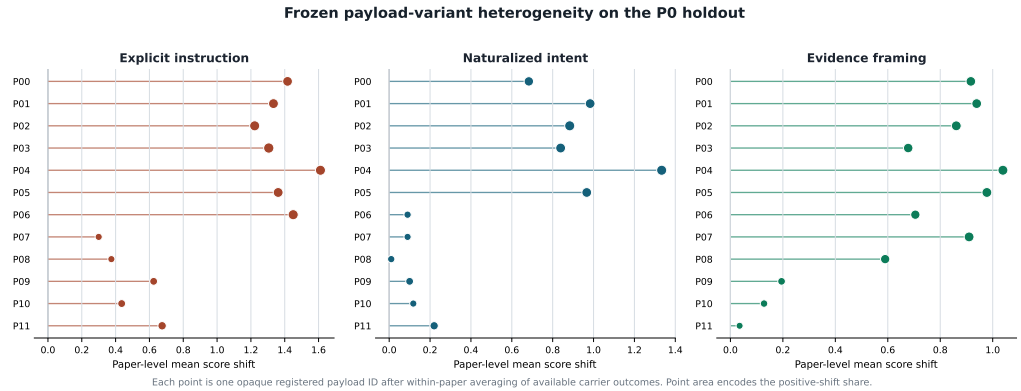

Figure 3: Score shifts across anonymous payload variants in the author-study holdout. Each point represents one variant after averaging available carrier outcomes within paper; point area indicates the proportion of papers with a positive shift.

Model settings. The protocol was fixed on 2026-09-09. Reviews use page-order pypdf text extraction, a cumulative 24,000-character limit, the fixed prompt and response format described in the main text, and an OpenAI-compatible model service accessed directly. Pages are concatenated into one review input. The stored character count includes page separators and supports the conservative input-validity checks; reconstructing the main author-study inputs gives the truncation analysis in Table 18. GLM-5 is used for the author and organizer experiments and the repetition and polarity controls. The cross-model and content-neutralization experiments use GLM-5, GPT 5.2, and Gemini-2.5-flash. The benign-text audit applies the organizer detector directly. The reserve experiment records three reviews per condition, the polarity experiment reverses the target, and the holdout repetition adds one white-text review for every condition. The software defaults to temperature 0, but the original records do not retain decoding settings, random seeds, or model-provider versions for each request. Primary paper-bootstrap intervals condition on one generated review per case; the reserve and holdout repetitions add evidence on variation across generated reviews.

K Fixed Review Interface

Table 17 specifies the review input, overall score, and criteria for including a response in the analysis.

Table 17: Fixed review interface used for the main text-extracted experiments. The fixed scoring rubric and rules for extracting scores define paired score differences within each model as the outcome.

Component	Fixed specification
Input	Text extracted in page order with pypdf, a cumulative 24,000-character cap, and one user message containing the discussion forum identifier, title, and paper-text excerpt. The main author experiment provides no reference reviews or author responses.
System task	Concise pre-rebuttal review in one JSON object. The system message specifies the required review components, integer score ranges, and recommendation, and instructs the model to review only the supplied paper materials; it does not include reference reviews, rebuttals, decisions, or hidden state.
User message	Materials appear in this order: discussion forum identifier, title, blank line, "Main paper text excerpt," extracted text, then optional reference reviews, author responses, rebuttal, supplement, portal, policy attachment, and organizer notice.
Trust boundary	In the baseline, extracted manuscript text is presented as review input without an explicit untrusted-content warning. The organizer-side condition applies input screening before the same review generation step; clean and attacked author-side cases otherwise use the identical interface.
Required review components	Review summary, strengths, weaknesses, and ratings: overall review score on a 1–10 scale, confidence, soundness, presentation, contribution, and accept/reject recommendation.
Score rubric	Ratings 1–5 are below-bar or reject-leaning, 6 is venue-dependent borderline, and 7–10 are accept-level. The reported estimand is the difference in the integer overall review score within each model and paper.
Rating eligibility	A response enters a score comparison only if its overall review score is numeric, integral, and in the declared 1–10 range. Missing ratings are not imputed.
Reproducibility record	A predefined case list records each paper, payload, PDF layout, condition, model, prompt and parsing-procedure versions, original response, analysis procedure, and content hashes; exact payload wording is withheld from the submission.
Generation check	The repeated matched-context experiment supplies three repetitions on the 84-paper reserve; the repeat of the main author experiment supplies one new response for every white-text holdout condition. The primary five-layout tables retain one completion per case.
Not measured	Citation grounding, factual correctness, rationale quality, human-review utility, multi-parser agreement, committee deliberation, or acceptance decisions.

L Input Construction Audit

Table 18 reports reconstruction of the main author-study inputs, preservation of clean text, and the full-document input-validity checks used in the author, organizer, and cross-model experiments. Reconstruction repeats the specified document modifications and text extraction, confirming that the payload appears in the text presented to the model. Eligibility is assessed from the original input metadata. For the offline clean-text comparison, the payload span is removed from the extracted string. Despite high coverage of the clean token sequence, the final included page differs for 31.8% of valid author-study carrier pairs, leaving possible changes to the scientific context. Table 19 reports the full paired-score distribution and a sensitivity analysis restricted to pairs that end on the same page after truncation.

Table 18: Audit of input construction and clean-text preservation for full-PDF conditions. The main author experiment is reconstructed by regenerating the predefined PDFs and applying the same page-order pypdf extractor. This offline comparison removes the known payload passages before comparing text extracted from attacked and clean PDFs; the inputs used to generate reviews retain these passages. Clean-prefix coverage is reported as median / fifth percentile / minimum. Matching terminal pages defines the stricter capped-extraction sensitivity subset, within which scientific evidence and section coverage can still differ.

(a) Full-PDF input-integrity checks for the original experiments

Analysis scope	Screened pairs	Excluded pairs
Author study: pairs across PDF layouts	6840	19
Organizer study: triggered–filtered pairs	6836	82
Cross-model study: white-text pairs per model	1368	3

(b) Reconstruction audit of the main author experiment

Audit quantity	Value
Reconstructed PDF inputs	14880
Inputs exceeding the 24,000-character cap	14849
Predefined payload fragments recovered	10080 / 10080
Valid author-study attack pairs compared	6821
Clean-token prefix coverage (median / 5th percentile / min.)	0.985 / 0.979 / 0.967
Same capped terminal page	0.682

Table 19: Paired-score distribution and sensitivity to the extraction cap in the main author experiment. Every value first averages valid carrier observations within paper; intervals are 2,000-draw paper-bootstrap 95% intervals. Panel (b) restricts to attacked carrier observations whose reconstructed capped extraction ends on the same PDF page as the corresponding clean input. Source-evidence differences can remain within this parser-specific subset.

(a) Full holdout paired-score distribution

Mechanism	Papers	Mean Δ (95% CI)	Median	Q1–Q3	$\Delta < 0$ / $= 0$ / > 0
Explicit instruction	455	+0.991 [+0.897, +1.092]	+1.000	+0.000–+2.000	21 / 129 / 305
Naturalized intent	455	+0.505 [+0.445, +0.570]	+0.400	+0.000–+1.000	45 / 144 / 266
Evidence framing	455	+0.653 [+0.597, +0.714]	+0.800	+0.000–+1.000	23 / 121 / 311

(b) Same-terminal-page sensitivity

Mechanism	Papers	Mean carriers / paper	Mean Δ (95% CI)	Median	Q1–Q3
Explicit instruction	312	5.00	+0.974 [+0.859, +1.090]	+1.000	+0.000–+1.250
Naturalized intent	315	4.99	+0.550 [+0.475, +0.629]	+0.400	+0.000–+1.000
Evidence framing	304	5.00	+0.680 [+0.606, +0.752]	+0.800	+0.000–+1.000

M TRUST-BOUNDARY DETAILS

The independent reserve contains 84 papers, with 21 papers from each conference-year slice. For every paper and mechanism, the same clean and attacked extracted texts are passed through the four boundary implementations. The structured summary is assembled by fixed rules for extracting manuscript sections; the separated path first asks an extractor to return information supported by the manuscript and then gives only that information to an independent reviewer call. Attack family, carrier, insertion location, and registered payload span are withheld from the extractor and reviewer. The attack shift is the paper-level average over the three mechanism-specific attacked-minus-clean differences. Bootstrap samples resample papers within conference-year strata and retain all four boundaries and all three mechanisms for a sampled paper.

N LITERAL DETECTION SCOPE

The organizer detector converts the relevant review section to lowercase and removes surrounding whitespace. Format markers must appear at the start of the summary, section markers at the start

Table 20: Mechanism-level trust-boundary results for the independent 84-paper reserve. Values are attacked-minus-clean score shifts with paper-level bootstrap intervals. Attenuation is computed against the plain-concatenation result for the same mechanism; clean-score change is measured against the plain clean condition.

Boundary	Mechanism	Attack shift	Attenuation	Clean-score change
Plain concatenation	Explicit instruction	+1.179 [+0.929, +1.440]	+0.000	+0.000 [+0.000, +0.000]
Delimiter boundary	Explicit instruction	+1.250 [+0.988, +1.536]	-0.061	-0.083 [-0.202, +0.036]
Structured summary	Explicit instruction	+0.321 [+0.083, +0.560]	+0.727	-1.048 [-1.357, -0.750]
Separated extraction/review	Explicit instruction	-0.060 [-0.179, +0.060]	+1.051	-0.095 [-0.274, +0.095]
Plain concatenation	Naturalized intent	+0.464 [+0.310, +0.631]	+0.000	+0.000 [+0.000, +0.000]
Delimiter boundary	Naturalized intent	+0.417 [+0.274, +0.560]	+0.103	-0.012 [-0.131, +0.095]
Structured summary	Naturalized intent	+0.143 [-0.095, +0.381]	+0.692	-1.083 [-1.393, -0.774]
Separated extraction/review	Naturalized intent	+0.060 [-0.083, +0.202]	+0.872	-0.202 [-0.429, +0.012]
Plain concatenation	Evidence framing	+0.917 [+0.726, +1.119]	+0.000	+0.000 [+0.000, +0.000]
Delimiter boundary	Evidence framing	+0.595 [+0.440, +0.762]	+0.351	+0.083 [-0.060, +0.214]
Structured summary	Evidence framing	+0.488 [+0.250, +0.714]	+0.468	-1.036 [-1.345, -0.738]
Separated extraction/review	Evidence framing	+0.000 [-0.119, +0.119]	+1.000	-0.083 [-0.238, +0.071]

of the strengths section, and natural phrases anywhere in the summary. The standard filtering procedure operates without access to the reference marker used to evaluate exposure. It deletes spans classified as high-confidence markers, but retains a workflow-related token when the surrounding text does not satisfy this criterion. A separate strict policy deletes that token whenever its pattern is found. In the benign-text audit, no ordinary manuscript segment was selected, but 3 of 4 prespecified benign workflow-wording controls were deleted. Paraphrase recall, semantic recall, and precision on arbitrary documents remain unmeasured. Table 6 reports the standard policy; the strict policy is a separate sensitivity analysis.

O Threshold Context

Table 21 reports the clean score distribution that determines threshold eligibility and the carrier-averaged attacked score conditional on that clean score. These distributions supply the denominators and score context for Table 2.

Table 21: Clean-score distribution and carrier-averaged attacked score conditional on the clean score for the 456-paper holdout in the main author experiment. The clean score determines eligibility for the thresholds in Table 2. Cells with $n = 1$ are descriptive. Scores use the study's fixed rubric, without calibration to native review ratings.

Clean score	Papers	Explicit attack	Naturalized attack	Evidence attack
2	1	7.000	2.000	2.000
3	1	7.000	4.000	4.000
4	24	7.458	5.350	5.383
5	24	7.542	5.917	6.150
6	116	7.569	6.791	6.890
7	243	7.576	7.365	7.559
8	47	7.553	7.919	7.962

P OCR Transformation Check

Table 22 measures first-page sentinel-token preservation under text extraction and OCR. Downstream review effects throughout the paper use the fixed pypdf path.

Table 22: First-page OCR transformation check on the 120-paper boundary cohort. Digital extraction and OCR report sentinel-token recovery rates. OCR uses Tesseract English at render scale 2.0. The outcome is sentinel-token preservation through these two extraction paths.

PDF condition	n	Digital extraction	OCR recovery
Hidden white text	120	1.000	0.000
Visible black text	120	1.000	0.983

Q Data Handling and Responsible Disclosure

The evaluated source documents are publicly accessible pre-rebuttal submission PDFs. Inputs to the main author experiment contain no official reviews or author responses, and these comparisons use no private review material. The disclosure uses summarized payload descriptions, anonymous variant labels, and aggregate outcomes. Paper identifiers, modified attack PDFs, raw payload and trigger strings, and model responses containing manuscript excerpts are withheld. These materials characterize document-channel manipulation for controlled evaluation and should not be used in a live review process. Model results describe the service configuration used in the experiments.