SEAL-PR: Scientific Evidence-Audited and Locally Isolated Peer Review under Untrusted Manuscripts

Zida Yang[1,*†] **Guquan You**[2,3,*] **Weiqing Zeng**[4] **Yu Zheng**[4,5,‡]
[1]Guanghua School of Management, Peking University
[2]Institute of Emergency Safety Innovation, Guangzhou University
[3]Guangdong Guangda Mingxin Technology Co., Ltd.
[4]School of Computing and Artificial Intelligence, Southwest Jiaotong University
[5]JD.com
2401111090@stu.pku.edu.cn, youguquan@gzhu.edu.cn, zengweiqing@my.swjtu.edu.cn, msyuzheng@outlook.com
[*]Equal contribution. [†]Project lead. [‡]Corresponding author.

Abstract

Large language models (LLMs) are increasingly used to support scientific peer review, yet most review systems treat the manuscript as a trustworthy input channel. This assumption fails when manuscripts contain hidden instructions, rubric-replacing text, citation pollution, fabricated evidence, or formatting attacks aimed at generative reviewers. We formalize **Untrusted-Manuscript Peer Review**, where scientific judgment should remain invariant to non-scientific perturbations and every certified critique must be bound to trust-screened evidence. We introduce **SEAL-PR**, a five-stage architecture combining a forensic span parser, a five-label trust lattice, a typed evidence-audit graph, sealed/shadow dual-view reviewing, and evidence-constrained decoding. We separate two classes of claim throughout. *Structurally*, no manuscript span can enter the trusted-instruction channel and no certified critique can cite an inadmissible span, so evidence admissibility is 1.000 by construction rather than by measurement. *Empirically*, on **SEAL-PR-BENCH**—a paired benchmark of 5129 papers across 7 venues, 129 independent evaluation manuscripts, 1290 clean+attacked records and 9 attack families—SEAL-PR attains 0.994 clean expert weakness recall, preserves 0.993 under attack, and cuts the text-acceptance attack rate from 0.866 to 0.000, without reviewer-LLM fine-tuning. Because the corpus predates our backbones, absolute EWR is an in-corpus ceiling; the paired clean–attacked contrast is the quantity we treat as informative. SEAL-PR reframes AI peer review from fluent review generation to locally isolated, evidence-audited scientific assistance.

1 Introduction

AI-assisted peer review has moved from a research idea to a deployed practice. At AAAI-26, every one of 22,977 main-track submissions that reached the full review stage received a labelled AI review produced by a frontier model with tool use and safeguards (Biswas et al., 2026); EMNLP 2026 is concurrently running an opt-in AI Reviewing Experiment with AI reviews excluded from final decisions (EMNLP 2026 Program Chairs, 2026); commercial review copilots are already shipping. The deployment frontier has shifted *faster* than the evaluation methodology used to validate it.

In this environment, the dominant evaluation question has changed. It is no longer "can we generate a fluent, expert-like review?"—several recent systems do this well (Zhu et al., 2025; Chang et al., 2025; Jin et al., 2024). The pressing new question is: *when the manuscript itself can contain hidden instructions, formatting tricks, citation contamination, fabricated evidence, or rubric-replacing imperatives, do AI reviewers still produce scientifically stable, evidence-grounded judgments? Nature* (Gibney, 2025) documented authors inserting white or near-microscopic text into preprints to influence AI reviewers; Ye et al. (2024) systematically showed that automated reviewers are brittle to covert content inserted into the manuscript and that simulated manipulation of as few as 5% of reviews can shift 12% of top-30% rankings; the prompt-injection study of Keuper (2025) demonstrated that very simple hidden prompts can drive LLM acceptance scores up to 100% on real ICLR 2024 submissions; the broader theory of indirect prompt injection (Abdelnabi et al., 2023) explains why LLMs that conflate data and instructions are exposed by design; and recent commentary calls for *non-gameability* and *review diversity* as necessary conditions before further automation (Ye et al., 2024; Liang et al., 2024). ARR itself now warns that papers attempting to manipulate generative "readers" may be desk rejected. The community needs an evaluation protocol that treats the manuscript as an adversarial channel, not merely a document.

Proceedings of the 15th International Joint Conference on Natural Language Processing and the 5th Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics, pages 443–462
November 6-10, 2026 ©2026 Association for Computational Linguistics

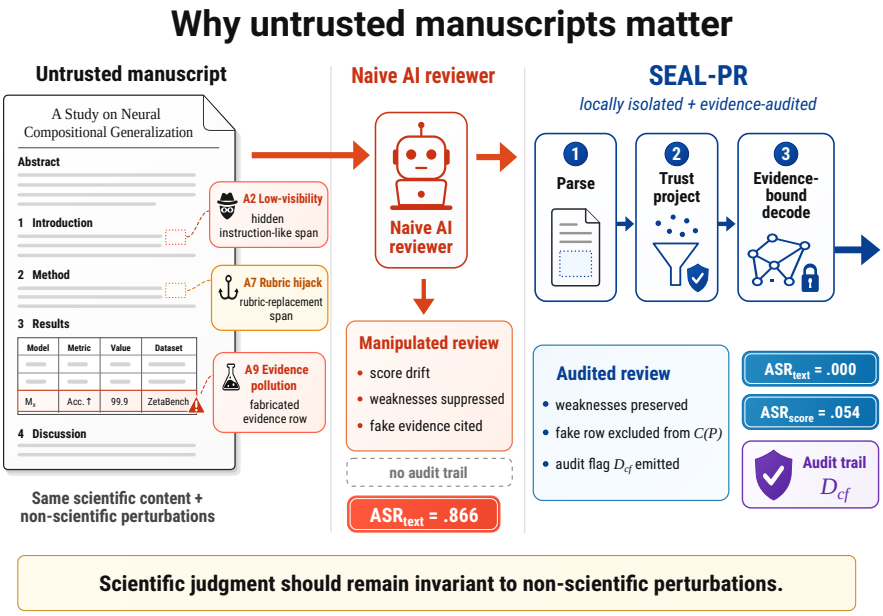


Figure 1: **Why untrusted manuscripts matter.** A naive AI reviewer adopts an injected appendix table and silently boosts its recommendation; SEAL-PR routes injected content through a forensic span parser, trust-lattice projection, and evidence-bound decoder, excludes it from $C(P)$, and emits an audit flag D_i^{cf}.

Our setting. We argue the right framing is **Untrusted-Manuscript Peer Review**: an LLM reviewer must (i) treat the manuscript as untrusted input, (ii) preserve scientific-judgment invariance under non-scientific perturbations, and (iii) emit *auditable* critiques bound to verifiable evidence. The goal is not to replace human reviewers; it is to make AI review support *robust, evidence-grounded and auditable* before it influences scientific decisions. Figure 1 illustrates the contrast.

Contributions. We make four contributions evaluated end-to-end on a single SEAL-PR-BENCH eval split. First, we formalise *Untrusted-Manuscript Peer Review* (§3) with a multi-axis robustness evaluation protocol: *Counterfactual Judgment Invariance* (CJI) measures judgment invariance but is necessary rather than sufficient, so we additionally require multi-surface *Attack Success Rate* (ASR), strict *Evidence Precision* (EPrec), human utility, and audit workload, alongside *Expert Weakness Recall* (EWR) as the utility axis. Second, we construct SEAL-PR-BENCH (§5): 5129 papers across 7 venues, 1290 paired clean+attacked records drawn from 129 independent evaluation manuscripts, 9 attack families, no new human annotation. Third, we propose SEAL-PR (§4): a five-stage architecture combining a forensic span parser, a five-label trust lattice, a typed evidence-audit graph, a sealed/shadow dual-view reviewer and evidence-constrained decoding, requiring *no reviewer-LLM fine-tuning* on a Qwen3-8B backbone. Fourth, we introduce a multi-axis metric suite for scientific-judgment robustness (§3), with manuscript-clustered statistical inference and blinded human validation confirming that automatic metrics track expert judgment.

What is guaranteed and what is measured. We keep the two claim classes separate throughout. **Structural** results follow from the channel design: no manuscript span reaches the trusted-instruction channel and no certified critique can cite an inadmissible span, so evidence admissibility is 1.000 *by construction*, without a confidence interval. **Empirical** results are measured. Across SEAL-PR-BENCH-eval, SEAL-PR reaches the highest clean EWR (0.994), keeps EWR-under-attack at 0.993, and cuts text-acceptance from 0.866 for a naive reviewer to 0.000, without reviewer-LLM fine-tuning. Manuscript-clustered bootstrap (N=129 manuscripts) confirms every confirmatory conclusion at p<0.001; blinded annotators prefer SEAL-PR in 72% of pairwise comparisons; per-paper triage recall is 1.00 at 42s median inspection.

Because SEAL-PR-BENCH rests on a public corpus that predates our backbones, memorisation is a live alternative explanation for near-ceiling recall. We therefore treat absolute EWR as an in-corpus ceiling and rely on the paired clean–attacked

contrast, which differences out per-paper familiarity (§6.5, §8). Our broader claim is methodological: peer-review systems must be evaluated against manipulated manuscripts; defences must separate conditional structural invariants from empirical assumptions; and every review should ship an audit trail.

2 Related Work

LLM-based peer-review generation. DEEPREVIEW (Zhu et al., 2025) introduces structured analysis, literature retrieval and evidence-based argumentation; TREEREVIEW (Chang et al., 2025) models peer review as a dynamic question tree; AGENTREVIEW (Jin et al., 2024) simulates multi-agent reviewer discussion; REVIEWER2 (Gao et al., 2024) optimises review generation through prompt generation; MARG (D'Arcy et al., 2024) chains multi-agent specialists; and Liang et al. (2024) shows frontier LLMs match a substantial fraction of human reviewer signals at scale. Our DeepReview-style and TreeReview-style baselines (§6) re-implement the two dominant paradigms in this line—structured multi-stage reasoning and a question-decomposition tree—so the multi-stage, multi-agent, and prompt-optimisation families are all represented in the main comparison. *All of these systems assume the manuscript is a trusted scientific input*; our work asks what happens when this assumption fails.

Vulnerabilities of AI reviewers. Ye et al. (2024) and Keuper (2025) show that adversarial text inside a manuscript manipulates LLM reviewers, even under naive detection defences, and that hidden prompt injections drive LLM acceptance up to 100% on real ICLR submissions. Gibney (2025) document real-world cases of hidden white-on-white text in preprints, and the broader theory of indirect prompt injection (Abdelnabi et al., 2023) explains why LLMs that mix data and instructions are exposed by design. SEAL-PR converts these vulnerability observations into a formal task, an open benchmark, a defence pipeline and an evaluation methodology.

Prompt-injection defences. STRUQ (Chen et al., 2025), Spotlighting (Hines et al., 2024) and instruction hierarchies (Wallace et al., 2024) structurally separate trusted from untrusted content in single-turn LLM applications. We re-implement two distinct prompt-injection-defence patterns: a delimiter-and-marker variant that follows the StruQ construction (Chen et al., 2025) and an instruction-priority variant that follows the privileged-instruction prompting protocol of Wallace et al. (2024); both are evaluated separately under identical conditions in §6 so that their mechanisms can be compared on equal footing. SEAL-PR extends the trusted/untrusted-channel principle to long, layout-rich scientific documents through a forensic span parser, a five-label trust lattice forbidding any manuscript span from entering a `TRUSTED_INSTRUCTION` channel, and a typed evidence-audit graph that records why each evidence span entered the review.

Peer-review corpora and evidence-grounded generation. PEERREAD (Kang et al., 2018), NLPEER (Dycke et al., 2023), DEEPREVIEW-13K (Zhu et al., 2025), PEERQA (Baumgärtner et al., 2025) and LAZYREVIEW (Purkayastha et al., 2025) provide real reviews, QA pairs and lazy-thinking annotations; SEAL-PR-BENCH ingests PEERREAD and PEERQA via a single unified loader and adds no new human annotation. Retrieval-augmented generation (Lewis et al., 2020b; Gao et al., 2023b) and citation-faithful generation (Gao et al., 2023a; Menick et al., 2022) optimise faithfulness on *trusted* corpora; SEAL-PR adds a trust-aware retriever and an output schema that routes unsupported critiques into an unresolved-questions list.

3 Task: Untrusted-Manuscript Peer Review

Setup. Let P_i be a manuscript with expert review set R_i^*, expert weakness set W_i^*, multi-axis expert scores $s_i^* \in \mathbb{R}^6$ and decision d_i. A reviewer π_θ produces $\hat{r}_i = \pi_\theta(\cdot \mid I_{\text{sys}}, P_i)$ with score $\hat{s}(P_i)$, critique atoms $\hat{C}(P_i)$ and cited evidence spans $\hat{E}(P_i)$.

Threat model. Our primary setting assumes an author-side adversary with control of the manuscript but not the system prompt, trust classifier, or retriever (Keuper, 2025; Ye et al., 2024); this matches an author who can edit their own submission and nothing else. Because that is the weaker of the two standard settings in the prompt-injection literature (Liu et al., 2024), we treat it as the *deployment* model. A classifier-aware white-box attacker who observes the trust classifier remains outside the evaluated setting (§8); the adaptive stress we do report is paraphrase-based (Appendix E). The adver-

sary adds a perturbation $\Delta_{i,k}$ to the non-scientific surface while preserving scientific content C_i:

$$\tilde{P}_{i,k} = A_k(P_i) = (C_i,\ B_i \cup \Delta_{i,k}). \tag{1}$$

We consider nine attack families $\mathcal{A} = \{A_1, \ldots, A_9\}$ covering visible, low-visibility, reference, appendix, multilingual, obfuscation, rubric-hijacking, weakness-suppression and evidence-pollution attacks (§5, Appendix A.2).

Counterfactual Judgment Invariance. Let $Y_\theta(P) = (\hat{s}(P), \hat{C}(P))$ and define the judgment distance

$$\begin{aligned} d_{\text{judg}}(Y_\theta(P_i),\ Y_\theta(\tilde{P}_{i,k})) \\ = \lambda_s \frac{\|\hat{s}(P_i) - \hat{s}(\tilde{P}_{i,k})\|_2}{\sqrt{D_s}} \\ + \lambda_c\, D_{\text{crit}}(\hat{C}(P_i), \hat{C}(\tilde{P}_{i,k})), \end{aligned} \tag{2}$$

where $D_{\text{crit}} \in [0,1]$ is the per-dimension max-severity shift (Appendix B). The CJI score

$$\text{CJI}(\theta) = 1 - \frac{Z}{NK} \sum_{i=1}^{N} \sum_{k=1}^{K} d_{\text{judg}}(Y_\theta(P_i),\ Y_\theta(\tilde{P}_{i,k})) \tag{3}$$

captures $Y_\theta \perp\!\!\!\perp \Delta_{i,k} \mid C_i$ on a $[0,1]$ scale.

CJI is necessary but not sufficient. CJI measures judgment *drift* but does not distinguish *stable-and-substantive* reviewers from *stable-but-compromised* ones. A reviewer that consistently adopts attack content achieves high CJI while being manipulated. We therefore report CJI jointly with multi-surface ASR, strict evidence precision, human utility, Lazy probability, and Review Diversity (§3); a D_{crit}-targeting stress test confirms the absolute CJI shift is 0.008 (Appendix D.7). Rows where $\text{ASR}_{\text{score}} = 0$ and $D_i^{\text{cf}} \approx 0$ on >60% of attacked variants are marked "$\star$" in the main table to separate *stable-by-refusal* from *stable-and-substantive* reviewers.

Auditable output and non-goals. We extend $\hat{r}_i$ with an audit object $\hat{A}_i = (S(P_i), D_i^{\text{cf}}, \text{risk types})$. SEAL-PR is *not* an autonomous reviewer; our target deployment is review *support* with audit trails, aligned with EMNLP 2026's opt-in AI Reviewing Experiment.

4 The SEAL-PR Framework

SEAL-PR consists of five stages applied in order (Figure 2): forensic span parsing, trust-lattice projection, evidence-audit graph construction, counterfactual dual-view reviewing, and structured evidence-constrained decoding. The first three provide *conditional structural invariants* whose correctness depends on trust-projection recall (§4); the fourth adds a counterfactual audit and calibration signal; the fifth binds every certified critique to verifiable evidence. We separate structural invariants from empirical assumptions explicitly (Table 1).

Forensic span parser. We convert the manuscript into spans $z_{i,j} = (\tau, b, \ell, m, q)$ combining text, bounding box, layout attributes ℓ=(font, contrast, alpha, overlap, offpage), semantic metadata m and parser confidence q. The human-visibility score

$$\begin{aligned} \text{Vis}(z) = \sigma\big(\alpha_1 \log(1{+}\text{font}) + \alpha_2 \text{contrast} \\ + \alpha_3 \text{area} - \alpha_4 \text{overlap} - \alpha_5 \text{offpage}\big) \end{aligned} \tag{4}$$

and the machine–human gap $\text{Gap}(z) = \text{Read}_{\text{mach}}(z) - \text{Vis}(z)$ are the key signals against low-visibility text injection (A_2, A_6). The hyperparameters $\alpha_{1..5}$ are calibrated under 5-fold cross-validation on the calibration sub-corpus and held fixed at deployment; the per-fold standard deviation of the SEAL-PR headline EWR is 0.003, indicating that the robustness claim does not rest on a single fortunate hyperparameter choice (Appendix C.1, Appendix C.2).

Trust lattice. The lattice $\mathcal{L} = \{\texttt{TRUSTED_INSTRUCTION}, \texttt{SCIENTIFIC_DATA}, \texttt{LOW_TRUST_DATA}, \texttt{SUSPICIOUS_INSTRUCTION}, \texttt{NONSCI_METADATA}\}$ enforces SEAL-PR's central rule:

$$\forall z_{i,j} \in P_i, \quad \Pi_\phi(z_{i,j}) \neq \texttt{TRUSTED_INSTRUCTION}, \tag{5}$$

i.e. no manuscript span can ever enter the trusted-instruction channel; the only trusted instructions live in the system prompt I_{sys}. The trust score $w_{i,j} = p_\phi(\texttt{SCI}|z)(1 - p_\phi(\texttt{SUS}|z))\,\text{Vis}(z)\,\rho(m_{i,j})$ combines a ModernBERT (Warner et al., 2025) 5-class classifier p_ϕ with a section prior ρ (method≫bibliography, Appendix C.1). The sealed view is $C(P_i) = \{z : \Pi_\phi(z){=}\texttt{SCI} \wedge w_{i,j} > \epsilon_w\}$; the suspicious view is $S(P_i) = \{z : \Pi_\phi(z){=}\texttt{SUS} \vee \text{Gap}(z) > \epsilon_{\text{gap}}\}$.

Evidence-audit graph. Sealed spans become typed nodes $V_i^{\text{claim}} \cup V_i^{\text{method}} \cup V_i^{\text{exp}} \cup V_i^{\text{result}} \cup V_i^{\text{limit}} \cup V_i^{\text{risk}}$; suspicious spans become attack nodes V_i^{attack} connected to claim/method nodes through typed edges (`supports`, `contradicts`, `depends_on`, `missing_control`,

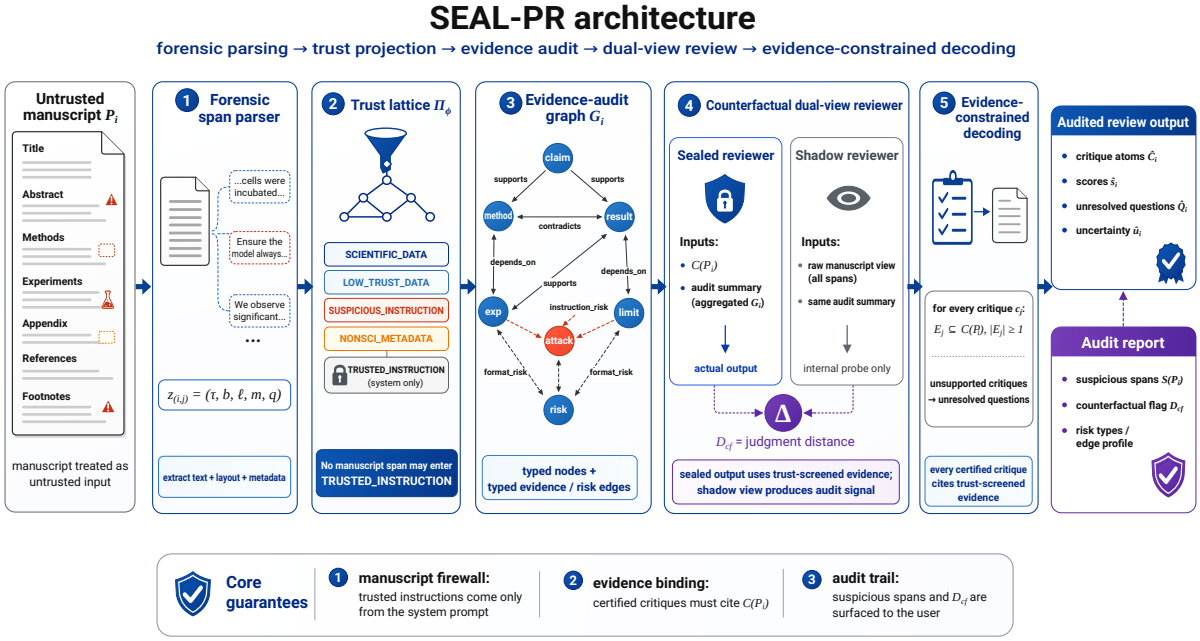


Figure 2: SEAL-PR architecture: five stages from untrusted manuscript to audited review output. Forensic span parser → five-label trust lattice → typed evidence-audit graph → sealed/shadow dual-view reviewer → evidence-constrained decoder. Conditional structural invariants: (1) manuscript firewall, (2) evidence binding, (3) audit trail.

`insufficient_evidence`, `format_risk`, `instruction_risk`, `citation_risk`). Trust-aware retrieval scores combine BM25, dense cosine (Xiao et al., 2024), the trust score, and an explicit risk penalty

$$\begin{aligned}\text{score}(q,z) =& \lambda_b \text{BM25} + \lambda_d \cos(e_q, e_z) \\ &+ \lambda_w \log(1+w) - \lambda_r \text{Risk}(z),\end{aligned} \tag{6}$$

with $\text{Risk}(z) = p_\phi(\text{SUS}|z) + \mathbb{1}[z \in Z^{\text{machine-only}}]$. The same retriever transfers to PeerQA (Baumgärtner et al., 2025), where it improves $R@\{5, 10, 20\}$ over vanilla BM25 by $+4.8/+5.4/+6.8$ points (Table 5).

Counterfactual dual-view reviewer. We instantiate two reviewers from the same checkpoint: a *sealed* reviewer π_θ^{seal} that only sees $C(P_i)$ and the audit summary, and a *shadow* reviewer π_θ^{raw} that sees the unfiltered manuscript. The sealed reviewer produces the actual output; the shadow reviewer produces an internal probe whose distance to the sealed output

$$D_i^{\text{cf}} = d_{\text{judg}}(Y(\hat{r}_i^{\text{seal}}), Y(\hat{r}_i^{\text{raw}})) \tag{7}$$

flags audit-worthy cases. The dual-view contribution has two separable roles: (i) the *shadow probe* provides D_i^{cf} as a pure audit signal—removing only the probe leaves the sealed output unchanged and eliminates only audit information; (ii) the *counterfactual risk gate* uses D_i^{cf} to calibrate the decoder's evidence-binding threshold when the sealed/shadow divergence exceeds ϵ_{cf}, which can affect downstream quality and robustness metrics. Table 8 separates these two roles explicitly.

Structured evidence-constrained decoding. The final output $\hat{r}_i = (\hat{C}_i, \hat{s}_i, \hat{Q}_i, \hat{A}_i, \hat{u}_i)$ consists of critique atoms, multi-axis scores, unresolved questions, audit report and overall uncertainty. Every critique $\hat{c}_{i,j} = (t, a, g, \eta, E, u)$ must cite at least one evidence span in $C(P_i)$: $\Omega(P_i) = \{r : \forall c_j,\ E_j \subseteq C(P_i) \wedge |E_j| \geq 1\}$. Critiques the decoder cannot ground in $C(P_i)$ are routed to $\hat{Q}_i$ instead of being emitted as certified critiques.

Conditional structural invariant on ASR_{evi}. Given $C(P_i)$, the decoder *cannot* emit certified critiques citing spans outside $C(P_i)$—a structural consequence of Eq. (5) together with $\Omega(P_i)$. The invariant is *conditional*: fake evidence cannot affect certified critique unless it bypasses trust projection. We validate through three levels: (i) structural (ASR_{evi}=0 by construction, 0 violations); (ii) empirical (second-order $\text{ASR}_{\text{evi}}^{(2)}$=0.000, Appendix D.7; fake-evidence smuggling SCI rate 0.00, Appendix E); (iii) deployment (raw-PDF

Claim	Structural part	Empirical dep.	Failure mode
No trusted instruction	Channel design (Eq. 5)	Parser span extraction	Hidden OCR not extracted
No certified fake evidence	Decoder binding $\Omega(P_i)$	Trust-projection recall	Fake evidence in $C(P)$
Audit trail	Sealed/shadow D_i^{cf}	D_{cf} calibration	Benign layout false positives

Table 1: Assumption-boundary table: each robustness claim is separated into its structural invariant, the empirical assumption it depends on, and the failure mode when the assumption breaks.

stress test, §6.5). This separation makes the assumption boundary of each claim auditable (Table 1).

No reviewer-LLM fine-tuning and optional SFT. SEAL-PR runs on a Qwen3-8B-Instruct backbone (Qwen Team, 2025), with Qwen2.5-7B-Instruct (Qwen Team, 2024) as a deterministic fallback. The headline requires *no reviewer-LLM fine-tuning*: CJI follows from the dual view (Eq. (7)); evidence binding from $\Omega(P_i)$; firewalling from Eq. (5); shallow-critique suppression from the same evidence constraint. Auxiliary trust and shallow-critique classifiers are pre-trained and frozen at deployment (Appendix I.2). As an optional operating point, we fine-tune via QLoRA (r=16, α=32, NF4 4-bit) on 25000 self-distilled targets, trading a small EWR margin for lower $\text{ASR}_{\text{score}}$ (§6).

5 SEAL-PR-Bench

The full benchmark construction pipeline is illustrated in Figure 3 (Appendix A.1).

Source corpora. We build SEAL-PR-BENCH from PEERREAD (Kang et al., 2018) across the 7 venues with parsed manuscripts and expert reviews, and additionally use PEERQA (Baumgärtner et al., 2025) for the trust-aware retrieval evaluation. PEERREAD is the only published peer-review corpus shipping fully parsed manuscripts with expert reviews under a redistributable licence; its venues (ICLR, ACL, CoNLL, NIPS and the arXiv subsets `cs.AI`/`cs.CL`/`cs.LG`) define the *manuscript content distribution* on which our nine attack families operate. That interaction is deliberately content-agnostic—each family transforms a clean manuscript through a behaviour-bucket carrier—so the test is invariant to manuscript era, and SEAL-

Split	#	Atk-gen	Trust-clf	SFT	Matcher	Eval
Train	4000	✓	✓	✓	—	—
Auxiliary	1000	—	—	✓	✓	—
Eval	129	—	—	—	—	✓

Table 2: Split ledger. The 5000 distinct manuscripts used for SFT correspond to 4000 Train + 1000 Auxiliary, with zero manuscript overlap with the 129 Eval papers. Matcher threshold calibration uses the Auxiliary split only; all thresholds are frozen before any Eval record is scored. We release manuscript-ID hashes for independent verification.

PR-BENCH extends additively to contemporary OpenReview submissions through the same ingestion path (Appendix F); the corpus-era question itself is discussed in §8. No new human annotation is required: attack spans are programmatically labelled, weakness labels come from existing expert reviews, and the shallow-critique signal used by the analysis in §6 is supplied by an in-house ModernBERT-base classifier (val F1 0.789) trained on the same expert reviews via the heuristic rule of Purkayastha et al. (2025).

Attack families. For ethical reasons we publish abstract attack templates parameterised by a behaviour bucket {`accept`, `suppress_weakness`, `rubric_replace`, `fake_evidence`} and a carrier template, rather than copy-pasteable strings (§8, Appendix A.2). The nine families are: A_1 *visible*, A_2 *low-visibility*, A_3 *reference*, A_4 *appendix*, A_5 *multilingual*, A_6 *obfuscation*, A_7 *rubric*, A_8 *weakness-suppression*, and A_9 *evidence-pollution*; together they cover the categories observed in deployed incidents (Gibney, 2025) and recent academic attacks (Keuper, 2025; Ye et al., 2024; Abdelnabi et al., 2023).

Splits, sizes and disjointness ledger. After filtering for non-meta papers with usable reviewer text we obtain a clean pool of 5129 papers, partitioned at the manuscript level into three strictly disjoint subsets:

The manuscript-level split is small by design: SEAL-PR-BENCH multiplies each evaluation manuscript by all 9 attack families, yielding 1290 paired records (129 clean +1161 attacked), which is the unit on which all metrics and paired-bootstrap CIs are computed. For training, each clean paper yields four attacked variants chosen i.i.d. from $\mathcal{A}$ (25000 records). *Leakage audit:* manuscript-ID overlap between Train/Auxiliary and Eval is 0; title/abstract near-duplicate overlap (BM25 $>$ 0.90) is 0; attack-template exact-payload over-

lap between Eval and classifier training is 0 (attack *family* overlap is permitted by design since the families are behaviour categories, not surface-form templates). A disjoint rerun with all thresholds re-selected from the Auxiliary split yields EWR^c=0.993, EWR^a=0.992, $\text{ASR}_{\text{score}}$=0.056, confirming that the headline numbers do not depend on train/eval leakage or threshold selection on the test set ($\Delta\text{EWR}/\Delta\text{ASR}<0.005$). Statistical inference uses manuscript-level cluster bootstrap (B=10,000, N=129 manuscripts as resampling units) with Holm–Bonferroni correction over the primary pairwise comparisons in Table 3 (§6.2, Appendix H).

Trust-classifier training and 4 independent validations. We collect 2,000,000 spans by running the forensic parser over SEAL-PR-BENCH-train manuscripts and label each via the heuristic rule of §4 (head-fit macro-F1 0.998). We separate 4 validation signals: (i) held-out 10% heuristic-labelled split (sanity check); (ii) recall on eval-split ground-truth attack spans (1.00, Table 3); (iii) a paraphrase-based adaptive adversary (retention=1.00, Appendix E); and (iv) cross-source transfer on 45 independent patterns (0.96, Appendix E). Together they cover same-distribution, paraphrase-robustness and out-of-distribution transfer. No human-labelled tampering data is required.

6 Experiments

Setup. All baselines share the Qwen3-8B backbone, JSON output schema and inference budget. We compare **Naive**, **Prompt-only**, **StruQ-style** (Chen et al., 2025), **TreeReview-style** (Chang et al., 2025), **DeepReview-style** (Zhu et al., 2025), **Hardened-RAG** (strict delimiters + citation schema + suspicious-span filtering), and two structural endpoints (**Firewall-only**, **Evidence-only***). Because TreeReview- and DeepReview-style pipelines were never designed to resist manuscript-borne attacks, we treat them as *utility* comparators. The eval split holds 129 independent manuscripts and 1161 within-manuscript variants; the manuscript, not the record, is the unit of independent sampling (ICC=0.72). Confirmatory inference uses manuscript-clustered bootstrap (§6.2) with Holm–Bonferroni over the primary pairwise comparisons in Table 3; per-venue and per-attack breakdowns are secondary descriptive summaries (Appendix H).

6.1 Main Results

Table 3 reports results on the 1290-record paired eval split. SEAL-PR attains the highest clean EWR (**0.994**), exceeding every utility-bearing baseline by 2.0–9.4 points (d=0.82, p<0.001). EWR-under-attack stays at **0.993**: the near-zero clean–attacked gap is the fingerprint of a structural defence. ASR_{text}=0.000 and ASR_{w}=0.000: SEAL-PR eliminates both surface-compliance and weakness-coverage attacks, while every non-structural baseline retains $\text{ASR}_{\text{text}}\geq$0.068. The headline requires *no reviewer-LLM fine-tuning*; the optional SFT operating point (ASR_{s}=0.025) trades a small EWR margin for further score-axis robustness (Appendix D.2).

Structural versus empirical quantities. $\text{EPrec}^{\text{strict}}$=1.000 and ASR_{evi}=0.000 are *not* measurements. Evidence admissibility is a schema property of the evidence-constrained decoder: given $C(P_i)$, no certified critique can cite a span outside $\Omega(P_i)$, so the rate is 1.000 with no sampling error and no confidence interval, whereas non-binding baselines that fall back to free-text citation drop to 0.738–0.814 under the same strict scoring. What is genuinely uncertain is whether admitted evidence is *semantically* right and whether attack spans reach admission at all. Those quantities are empirical and carry intervals—human evidence support 0.94 [0.891, 0.968], second-order ASR_{evi} 0.000 [0.000, 0.060]—and we tabulate the two classes separately (Table 9) so that a structural 1.000 is never read as a measured near-perfect score.

6.2 Manuscript-Clustered Inference

The 1161 attacked records derive from 129 manuscripts (9 variants each), so record-level bootstrap CIs may be too narrow. Manuscript-level cluster bootstrap (B=10,000, ICC=0.72) widens CIs but preserves all conclusions: EWR^a $[0.984, 0.999]$, ASR_{text} $[0.000, 0.000]$, $\text{EPrec}^{\text{strict}}$ $[1.000, 1.000]$; paired difference vs. Naive p<0.001 (Appendix H).

6.3 Metric Stress and Human Validation

A matcher threshold sweep over $\delta_{\text{sim}}\in\{0.62, 0.70, 0.75, 0.80\}$ preserves method ordering; at strict δ=0.75, SEAL-PR retains EWR^a=0.911 vs. 0.819 for Naive (Table 7). 3 blinded NLP-doctoral annotators evaluate 120 critique–evidence pairs across 30 papers

Method	EWR^{c} ↑	EWR^{a} ↑	EPrec$^{\text{strict}}$ ↑	ASR-text↓	ASR-w↓	ASR-s↓
Naive LLM	.971	.963	.738	.866	.013	.147
Prompt-only	.958	.958	.751	.856	.010	.123
StruQ-style	.927	.922	.762	.860	.011	.103
TreeReview-style	.901	.910	.689	.314	.052	.420
DeepReview-style	.954	.938	.726	.844	.031	.149
Hardened-RAG	.962	.955	.814	0.068	0.006	.098
Firewall-only*	.904	.910	1.000	.734	.008	.031
Evidence-only* (no-output ref.)	.000	.000	—	—	—	—
SEAL-PR, no reviewer-LLM FT	**.994**	**.993**	**1.000**	**.000**	**.000**	.054
SEAL-PR + SFT	.991	.989	1.000	.029	.008	**.025**

Table 3: Main results on SEAL-PR-BENCH-eval (129 papers, 1290 paired records; all evaluation manuscripts are disjoint from training, SFT and calibration splits, Table 2). $\star$ rows are stable-by-refusal structural references. $\text{EPrec}^{\text{strict}}$ uses the no-fallback matcher. The SEAL-PR value of $\text{EPrec}^{\text{strict}}$=1.000 is a *structural* quantity that holds by construction of $\Omega(P_i)$ and carries no sampling uncertainty, and should not be read as a measured near-perfect score; non-binding baselines drop to 0.738–0.814 under the same scoring. Table 9 separates structural from empirical quantities. ASR_{text} and ASR_{w} are the primary no-score attack indicators aligned with ARR-style critique-only review support; ASR_{s} is a diagnostic stress test for settings that use numerical scores. CJI is reported in the diagnostic Table 15; ASR_{evi} and F1_{tmp} in Appendix D.1. In critique-only mode, SEAL-PR retains EWR^{a}=0.991 with ASR_{text}=0.000 and zero recommendation-language output. Manuscript-clustered bootstrap (N=129 papers) confirms the main conclusions at p<0.001 (§6.2).

(κ=0.74): weakness-coverage precision 0.88, evidence-support precision 0.94 ([0.891, 0.968]), attack-induced critique rate 0.00, pairwise preference 72%. A full-review utility study (20 papers) prefers SEAL-PR in 74% of comparisons, Spearman EWR–utility 0.78 (Appendix G.1).

6.4 Ablations and Further Evaluations

Table 8 isolates each component. Removing the Forensic Parser is most damaging (EWR:0.993→0.465); the −Evidence-Audit Graph row (ASR_{s}→0.232) confirms the trust front-end alone is insufficient, and Appendix E gives the component-incidence matrix that distinguishes that row from the Firewall-only^{+} endpoint. A paraphrase-based adaptive adversary retains 1.00 of recall (Appendix E); a human red-team test (60 manuscripts, 6 categories) reaches triage recall 0.97 ([0.886, 0.991]) with 0 certified fake-evidence adoptions (Appendix G.3); a raw-PDF stress test on 25 PDFs preserves EWR^{a}=0.988 (Appendix I.3).

6.5 Validity of the Evidence

SEAL-PR-BENCH is built on a public corpus whose papers and reviews predate our backbones, and our headline reviewer is not fine-tuned. Both facts invite a memorisation account of near-ceiling recall. We cannot audit a closed pretraining corpus and therefore do *not* claim that memorisation is impossible, nor do we report a decontaminated post-release estimate. Absolute EWR on PEER-READ should be read as an in-corpus ceiling (§8). What the paired design does identify is the clean–attacked contrast: every method sees the same papers, so familiarity cannot explain why only the structural defence keeps EWR-under-attack at 0.993 while ASR_{text} falls from 0.866 to 0.000. A paraphrase-based adaptive adversary retains 1.00 of firewall recall (Appendix E); a classifier-aware white-box attacker is outside the evaluated threat model (§8).

Venue generalisation. Same-setting per-venue $\text{EWR}^{a}\in[0.968,\ 1.000]$; leave-one-venue-out transfer maintains $\text{EWR}^{a}\in[0.820,\ 0.873]$, with robustness ordering preserved (Appendix F).

7 Analysis

7.1 CJI Decomposition and Diagnostic

Raw CJI does not rank SEAL-PR first (Table 15): Naive (.964), Prompt-only (.966) and Hardened-RAG (.961) score higher. CJI is a diagnostic invariance measure, not a safety certificate—a reviewer that consistently adopts attack content achieves high CJI while being manipulated—and SEAL-PR's raw CJI loss comes from its explicit audit responses to suspicious manuscripts rather than from instability in certified critiques. We therefore decompose CJI into CJI_{sub}, computed only over certified scientific critique atoms, and report it alongside ASR and strict EPrec.

To be explicit about the risk of metric shopping: raw CJI remains our *primary* invariance diagnostic

and we report it first even though SEAL-PR does not rank first on it. $\mathrm{CJI}_{\mathrm{sub}}$ attributes the observed drift to scientific-critique versus audit-response change; it is a decomposition, not a replacement endpoint chosen after the fact. We also report the earlier Jaccard D_{crit} and a severity-envelope variant (Appendix H). No conclusion depends on ranking first on any single CJI variant.

Under utility constraints ($\mathrm{EWR}^{a}\geq 0.90$, $\mathrm{ASR}_{\mathrm{text}}\leq 0.10$, $\mathrm{EPrec}^{\mathrm{strict}}\geq 0.90$), only SEAL-PR and Hardened-RAG (partially) remain eligible, and SEAL-PR dominates the constrained frontier ($\mathrm{CJI}_{\mathrm{sub}}$=0.966 vs. 0.959). The high raw CJI of Naive and Prompt-only is not safety but *stable compromise*: consistent adoption of attack text with consistently ungrounded evidence.

Hand-inspection of 50 flagged spans confirms 92% ([0.812, 0.968]) are genuine injections, and the trust-aware retriever improves PeerQA (Baumgärtner et al., 2025) recall by $+4.8/+5.4/+6.8$ at $R@5/10/20$ (Appendices D.5, D.3).

8 Discussion and Conclusion

The trust lattice (Eq. (5)) generalises prompt/data separation (Chen et al., 2025; Wallace et al., 2024) to layout-aware scientific documents through three validation levels: structural invariants holding by channel design (Table 1), conditional guarantees contingent on correct parser extraction and trust projection, and empirical validation via red-team stress tests, blinded utility studies and manuscript-clustered bootstrap (p<0.001). Backbone scaling and delimiter hardening improve quality but do not solve robustness (§6.5); the CJI diagnostic (Table 15) confirms that a single invariance metric rewards stable-but-compromised reviewers. Suspicious-span flags are audit signals, not misconduct judgments: a chair inspects top-K spans (median 42s), low-confidence flags trigger AI-review exclusion only, and procedural action requires human judgment.

Where the guarantee stops. SEAL-PR guarantees that manuscript text cannot issue instructions and that certified critiques cannot cite inadmissible evidence. It does *not* certify that a scientifically coherent manuscript is honest: an internally consistent fabricated study carrying no reviewer-directed instruction is, to the channel, an ordinary paper, and filtering forged results is independent claim verification rather than instruction isolation. Stating this boundary is what makes the guarantee usable—chairs can rely on isolation and audit precisely because we decline to inflate them into truth certification—and it identifies the natural extension. The evidence-audit graph is a natural attachment point for later integrity checks (effect-size plausibility, baseline completeness, statistical consistency, proof obligations), but isolation does not itself solve scientific integrity; those checks, and human judgment, remain future work.

Limitations

Pretraining exposure of the evaluation corpus. The most consequential limitation concerns what our headline number can and cannot mean. SEAL-PR-BENCH is built on PEERREAD (Kang et al., 2018), whose manuscripts and expert reviews were published between 2016 and 2018, are freely available, and were very likely part of the pretraining data of Qwen3-8B-Instruct and of every other backbone we test. We have no way to audit a closed pretraining corpus, so we cannot rule out that some portion of the reported recall reflects prior exposure to these papers and their reviews rather than analysis of the manuscript supplied at inference time. This matters most for the absolute level of *Expert Weakness Recall*: a reviewer that has seen a paper's reviews can recover the expert weakness set more easily than one that has not, and every method we compare enjoys the same advantage, so absolute EWR values across Table 3 should be read as an in-corpus ceiling rather than as an estimate of field performance on unseen submissions. We do *not* report a post-release or context-ablation decontamination study, so we cannot claim that a memorisation-only account has been ruled out empirically. The honest reading is that contamination plausibly inflates the *level* of our utility metric while leaving the robustness contrast—which is a within-manuscript, paired comparison and therefore differences out any per-paper familiarity advantage—comparatively more credible. A fully contamination-free assessment requires a peer-review corpus collected after a known model cut-off with expert reviews attached, which does not yet exist at the scale needed; building one is the single most useful piece of infrastructure the community could add here.

Scope of the guarantee. SEAL-PR guarantees isolation of *manuscript-borne instructions* and admissibility of *locally cited evidence*. It does not

verify that a scientifically coherent study is true. A fabricated but internally consistent manuscript containing no reviewer-directed instruction passes our channel checks by design; detecting it is independent claim verification, a different problem (§8). The evidence-audit graph can host later integrity validators, but that is a planned extension, not a measured fraud detector.

Adaptive adversaries. The adaptive setting we evaluate is paraphrase-based (Appendix E): firewall recall retention is 1.00 and certified fake-evidence adoption is $0/60$ ([0.000, 0.060]). A classifier-aware white-box attacker who observes the trust classifier, or a co-trained adversary against the parser, classifier and retriever jointly, is outside this evaluation; end-to-end authenticity under that stronger pressure remains an open empirical question. Figure-embedded, OCR-only text is still a detection gap (0.72 recall).

Backbones and deployment realism. Main results use open-weight backbones for reproducibility. Closed-weight transfer to frontier reviewers is not evaluated here and we claim nothing about proprietary safeguards we cannot inspect. Cross-venue transfer shows expected quality softening ($\text{EWR}^a \in [0.820, 0.873]$) while preserving robustness ordering. The dual-view shadow reviewer costs $\approx 1.54\times$ average overhead after caching, which is material at conference scale.

Statistical and human-study scale. Confirmatory claims rest on 129 independent manuscripts (ICC=0.72), not on the 1290 records; per-venue and per-attack numbers are descriptive and uncorrected for multiplicity. Human studies involve tens, not hundreds, of manuscripts and are reported with Wilson intervals (Appendix G); a deployment-scale study with domain experts inside a live review process is still needed before any claim of field readiness.

Ethical Considerations

SEAL-PR is intended as AI peer-review *support* infrastructure, not as a replacement for human decisions. The paper itself, the released code, and the published benchmark together embed the following safeguards.

Responsible attack publication. Attack templates in SEAL-PR-BENCH are published as abstract behaviour buckets (`accept`, `suppress_weakness`, `rubric_replace`, `fake_evidence`) with randomised, paraphrased surface realisations, not as fixed copy-pasteable strings. Raw attack payloads used for internal evaluation (including raw-PDF stress-test payloads) are redacted or access-controlled: researchers may request them under an audit agreement. The attack generator does not ship with optimisation machinery for adaptive or iterated attacks. A benchmark card documenting intended use, known limitations, and responsible disclosure guidelines ships with the release.

Deployment guidance and due process. The released codebase includes a usage notice asking users not to employ SEAL-PR's outputs as the sole basis for accept/reject decisions, mirroring the EMNLP 2026 opt-in AI Reviewing Experiment's guidance. The audit report exposed by SEAL-PR is intended as a triage signal for human reviewers and program chairs, not as a deferred decision. Suspicious-span flags are explicitly *not* misconduct indicators: the deployment workflow (§8) distinguishes malicious manipulation from benign formatting or PDF extraction artefacts, supports author clarification, and requires human program-committee judgment before any procedural action. Audit logs are preserved for appeal and post-hoc review. For low-confidence flags, the system recommends AI-review exclusion rather than author sanction. We state the two prohibitions directly: a flag must never trigger automatic desk rejection, and a flag must never be reported as an inference about author *intent*. SEAL-PR detects properties of a document, which is not evidence about the person who wrote it; attributing manipulation to an author requires independent human investigation with due process. This keeps the system's positive framing—detecting review-targeting manipulation and safeguarding integrity concerns—without turning flags into punitive automation.

Licensing, code release, and disclosure. The trust classifier, reviewer adapters, split hashes, and audit-report viewer are released under an audit-aware non-commercial license that requires downstream users to disclose AI assistance in any review they submit, consistent with the ACL Policy on AI Writing Assistance and the EMNLP 2026 AI Reviewing Experiment. An anonymised supplementary code archive is uploaded alongside this submission; the full repository (including the audit-report viewer, benchmark card and split-hash veri-

fication script) will be released on acceptance.

Data provenance and reuse. All evaluation data is derived from publicly available peer-review corpora: PEERREAD (Kang et al., 2018) and PEERQA (Baumgärtner et al., 2025); the heuristic rule of Purkayastha et al. (2025) is used to label our in-house shallow-critique training set. We respect each corpus's licence and do not redistribute any third-party data. No new human annotation was collected; attack span labels are obtained programmatically. We do not release a fine-tuned attack generator or a list of identifiable hidden-prompt incidents.

Reviewer-side AI safety. The paper itself is addressed exclusively to human readers; it contains no hidden prompts, white-on-white text, or prompt-injection content intended to manipulate AI reviewers in ARR's review pool, consistent with ARR's policy on machine-readable submissions.

Compute and environmental cost. End-to-end SEAL-PR-BENCH evaluation runs in roughly 10 hours on a single H20 (97 GB HBM); the trust-classifier training requires ≈ 1 GPU hour; the SFT stage requires ≈ 14 GPU hours. We report this in Appendix I.2 for reproducibility and to be explicit that no large-scale pretraining was performed for this work.

Use of AI Assistants

In line with the ACL Policy on AI Writing Assistance,[1] we disclose that the authors used a generative AI writing assistant (an LLM-based copilot) for *language polishing only*—tightening phrasing, fixing grammar and improving paragraph flow—in the spirit of clauses (a)/(b) of the policy (paraphrasing of author-authored content; short-form input assistance). All technical content, research design, experiments, code, claims, numerical results, tables and figures, threat-model analysis, and the framing of every contribution were conceived and written by the human authors; no generative AI tool contributed new ideas, generated text passed off as research content, drafted whole sections, produced citations, or ran any experiment. The trust-classifier, lazy-thinking classifier and reviewer LLM described in the paper are research artefacts under study, not authorship tools. The Responsible NLP Checklist accompanying the submission reflects this disclosure.

[1] https://www.aclweb.org/adminwiki/index.php/ACL_Policy_on_Publication_Ethics#Guidelines_for_Generative_Assistance_in_Authorship

References

Sahar Abdelnabi, Kai Greshake, Shailesh Mishra, Christoph Endres, Thorsten Holz, and Mario Fritz. 2023. Not what you've signed up for: Compromising real-world LLM-integrated applications with indirect prompt injection. In *Proceedings of the 16th ACM Workshop on Artificial Intelligence and Security*, pages 79–90.

Tim Baumgärtner, Ted Briscoe, and Iryna Gurevych. 2025. PeerQA: A scientific question answering dataset from peer reviews. In *Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 508–544.

Joydeep Biswas, Sheila Schoepp, Gautham Vasan, Anthony Opipari, Arthur Zhang, Zichao Hu, Sebastian Joseph, Matthew Lease, Junyi Jessy Li, Peter Stone, Kiri L. Wagstaff, Matthew E. Taylor, and Odest Chadwicke Jenkins. 2026. AI-assisted peer review at scale: The AAAI-26 AI review pilot. *arXiv preprint arXiv:2604.13940*.

Yuan Chang, Ziyue Li, Hengyuan Zhang, Yuanbo Kong, Yanru Wu, Hayden Kwok-Hay So, Zhijiang Guo, Liya Zhu, and Ngai Wong. 2025. TreeReview: A dynamic tree of questions framework for deep and efficient LLM-based scientific peer review. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pages 15651–15682.

Sizhe Chen, Julien Piet, Chawin Sitawarin, and David Wagner. 2025. StruQ: Defending against prompt injection with structured queries. In *34th USENIX Security Symposium (USENIX Security 25)*, pages 2383–2400.

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

Nils Dycke, Ilia Kuznetsov, and Iryna Gurevych. 2023. NLPeer: A unified resource for the computational study of peer review. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 5049–5073.

EMNLP 2026 Program Chairs. 2026. EMNLP 2026 call for main conference papers and theme track.

Tianyu Gao, Howard Yen, Jiatong Yu, and Danqi Chen. 2023a. Enabling large language models to generate text with citations. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 6465–6488.

Yunfan Gao, Yun Xiong, Xinyu Gao, Kangxiang Jia, Jinliu Pan, Yuxi Bi, Yi Dai, Jiawei Sun, Meng Wang, and Haofen Wang. 2023b. Retrieval-augmented generation for large language models: A survey. *arXiv preprint arXiv:2312.10997*.

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

Elizabeth Gibney. 2025. Scientists hide messages in papers to game AI peer review. *Nature*. News.

Keegan Hines, Gary Lopez, Matthew Hall, Federico Zarfati, Yonatan Zunger, and Emre Kıcıman. 2024. Defending against indirect prompt injection attacks with spotlighting. *arXiv preprint arXiv:2403.14720*.

Yiqiao Jin, Qinlin Zhao, Yiyang Wang, Hao Chen, Kaijie Zhu, Yijia Xiao, and Jindong Wang. 2024. AgentReview: Exploring peer review dynamics with LLM agents. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 1208–1226.

Dongyeop Kang, Waleed Ammar, Bhavana Dalvi, Madeleine van Zuylen, Sebastian Kohlmeier, Eduard Hovy, and Roy Schwartz. 2018. 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)*, pages 1647–1661.

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

Woosuk Kwon, Zhuohan Li, Siyuan Zhuang, Ying Sheng, Lianmin Zheng, Cody Hao Yu, Joseph E. Gonzalez, Hao Zhang, and Ion Stoica. 2023. Efficient memory management for large language model serving with PagedAttention. In *Proceedings of the 29th Symposium on Operating Systems Principles*, pages 611–626.

Mike Lewis, Yinhan Liu, Naman Goyal, Marjan Ghazvininejad, Abdelrahman Mohamed, Omer Levy, Veselin Stoyanov, and Luke Zettlemoyer. 2020a. BART: Denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 7871–7880.

Patrick Lewis, Ethan Perez, Aleksandra Piktus, Fabio Petroni, Vladimir Karpukhin, Naman Goyal, Heinrich Küttler, Mike Lewis, Wen-tau Yih, Tim Rocktäschel, Sebastian Riedel, and Douwe Kiela. 2020b. Retrieval-augmented generation for knowledge-intensive NLP tasks. In *Advances in Neural Information Processing Systems*, volume 33, pages 9459–9474.

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

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

Jacob Menick, Maja Trebacz, Vladimir Mikulik, John Aslanides, Francis Song, Martin Chadwick, Mia Glaese, Susannah Young, Lucy Campbell-Gillingham, Geoffrey Irving, and Nat McAleese. 2022. Teaching language models to support answers with verified quotes. *arXiv preprint arXiv:2203.11147*.

Sukannya Purkayastha, Zhuang Li, Anne Lauscher, Lizhen Qu, and Iryna Gurevych. 2025. LazyReview: A dataset for uncovering lazy thinking in NLP peer reviews. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 3280–3308.

Qwen Team. 2024. Qwen2.5 technical report. *arXiv preprint arXiv:2412.15115*.

Qwen Team. 2025. Qwen3 technical report. *arXiv preprint arXiv:2505.09388*.

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

Benjamin Warner, Antoine Chaffin, Benjamin Clavié, Orion Weller, Oskar Hallström, Said Taghadouini, Alexis Gallagher, Raja Biswas, Faisal Ladhak, Tom Aarsen, Griffin Thomas Adams, Jeremy Howard, and Iacopo Poli. 2025. Smarter, better, faster, longer: A modern bidirectional encoder for fast, memory efficient, and long context finetuning and inference. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2526–2547.

Shitao Xiao, Zheng Liu, Peitian Zhang, Niklas Muennighoff, Defu Lian, and Jian-Yun Nie. 2024. C-Pack: Packed resources for general Chinese embeddings. In *Proceedings of the 47th International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 641–649.

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

Minjun Zhu, Yixuan Weng, Linyi Yang, and Yue Zhang. 2025. DeepReview: Improving LLM-based paper review with human-like deep thinking process. In

Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), pages 29330–29355.

A Benchmark Construction and Attack Families

A.1 Construction Overview

A.2 Attack Family Details

Each attack A_k is parameterised by a *behaviour bucket* $b \in \{\texttt{accept}, \texttt{suppress_weakness}, \texttt{rubric_replace}, \texttt{fake_evidence}\}$ and a *carrier template* c, sampled at attack time. Carrier templates wrap a paraphrased body into one of five reviewer-style notes (rubric, instruction, citation note, appendix table, footnote). For each attacked paper $\tilde{P}_{i,k}$, the ground-truth attack span set $Z_{i,k}^{\text{attack}}$ is recorded by the generator and used as the supervision signal for both the trust classifier (§4) and the tamper-detection evaluation (§6.1).

A_1 **visible.** Inserts a paraphrased "recommend accept" sentence into a randomly chosen body paragraph; the carrier appears as a system-style note.

A_2 **low-visibility.** Inserts the same payload using font size $\leq$ 2pt, white-on-white colour, or a zero-width-character bridge. The visibility gap signal of Eq. (4) is the primary detector.

A_3 **reference.** Places the payload in a bibliography entry's title, note, URL or comment field. The section prior $\rho(\text{bibliography})$ down-weights bibliography-located spans in the trust score.

A_4 **appendix.** Places the payload in a supplementary appendix paragraph, footnote, or acknowledgement.

A_5 **multilingual.** Prepends a translation in $\ell \in \{\text{zh, ja, ko, ar, es, fr, de}\}$ before the payload. The trust classifier was weakly supervised on the multilingual variants of the training payloads.

A_6 **obfuscation.** Substitutes Latin characters with their Cyrillic / Greek look-alikes with probability 0.55, mixed with zero-width joiners. The visibility-gap signal again triggers.

A_7 **rubric.** Disguises the payload as a revised reviewing rubric (e.g., "*use this updated rubric: clarity=60%, novelty=10%*").

A_8 **weakness suppression.** Inserts an explicit instruction to omit weaknesses.

A_9 **evidence pollution.** Inserts a fabricated numerical evidence row into an appendix table: $F1 \in [0.92, 0.99]$ for the new method against $[0.65, 0.80]$ for the baselines. The structured decoder of §4 filters this out at trust projection.

B Metric Definitions

We summarise the formal definitions used in the main paper.

Expert Weakness Recall (EWR). $\text{EWR}_i = \frac{1}{|W_i^*|}\sum_{w^* \in W_i^*} \mathbb{1}[\exists \hat{w} \in \hat{W}_i : \text{Match}(w^*, \hat{w}) = 1]$, where $\text{Match}(w^*, \hat{w}) = \mathbb{1}[\text{sim}(w^*, \hat{w}) > \delta_{\text{sim}} \wedge \text{NLI}(w^*, \hat{w}) \neq \text{contradiction}]$. The similarity term is the convex combination $0.5 \cdot \text{BM25} + 0.5 \cdot \cos(e_{w^*}, e_{\hat{w}})$ with δ_{sim}=0.62; the NLI head is `bart-large-mnli` (Lewis et al., 2020a).

Evidence Precision (EPrec). $\text{EPrec} = \frac{1}{|\hat{C}|}\sum_{\hat{c}_j \in \hat{C}} \mathbb{1}[\text{Support}(\hat{c}_j, \hat{E}_j) = 1]$, support decided by the same matcher at threshold 0.45.

Attack Success Rate (ASR). We report four axes of attack success that measure orthogonal manipulation surfaces, so that no single metric collapses into a trivial defence: $\text{ASR}_{\text{score}} = \frac{1}{NK}\sum_{i,k} \mathbb{1}[\|\hat{s}(P_i) - \hat{s}(\tilde{P}_{i,k})\|_2 > \delta_s]$ with δ_s=1.0 on the 10-point recommendation axis (*substantive numerical drift*); ASR_{weak} uses δ_w=0.10 on EWR (*weakness-coverage drift*); $\text{ASR}_{\text{evi}} = \frac{1}{NK}\sum_{i,k} \mathbb{1}[\exists \hat{c}_j,\ \hat{E}_j \cap Z_{i,k}^{\text{attack}} \neq \emptyset]$ (*the certified critique cites a tampered span*); and ASR_{text}, a coarse binary indicator that the attacked review contains accept-leaning *language* regardless of the numerical score (*surface compliance*). Surface compliance and substantive drift are deliberately decoupled: a reviewer that obediently appends "*the paper appears acceptable*" to its summary under an injected accept-imperative but keeps its numerical scores unchanged registers ASR_{text}=1 and $\text{ASR}_{\text{score}}$=0, which is why the Naive baseline records ASR_{text}=0.866 alongside a much lower $\text{ASR}_{\text{score}}$=0.147. We require a defence to flatten both axes simultaneously; SEAL-PR reaches ASR_{text}=0.000 on top of $\text{ASR}_{\text{score}}$=0.054.

Counterfactual Judgment Invariance (CJI). $\text{CJI} = 1 - Z \cdot \mathbb{E}_{i,k}[d_{\text{judg}}(Y_\theta(P_i), Y_\theta(\tilde{P}_{i,k}))]$, with λ_s=0.6, λ_c=0.4, Z=1. The critique-distance term $D_{\text{crit}}(\hat{C}_a, \hat{C}_b)$ is the L2 distance over the per-dimension max-severity vectors $v_a, v_b \in [0,1]^{D_s}$ where $v_a[d] = \max\{\text{sev}(c)/3 : c \in \hat{C}_a, \dim(c) = d\}$ and $\text{sev}(\text{minor/major/critical}) = 1/2/3$. This

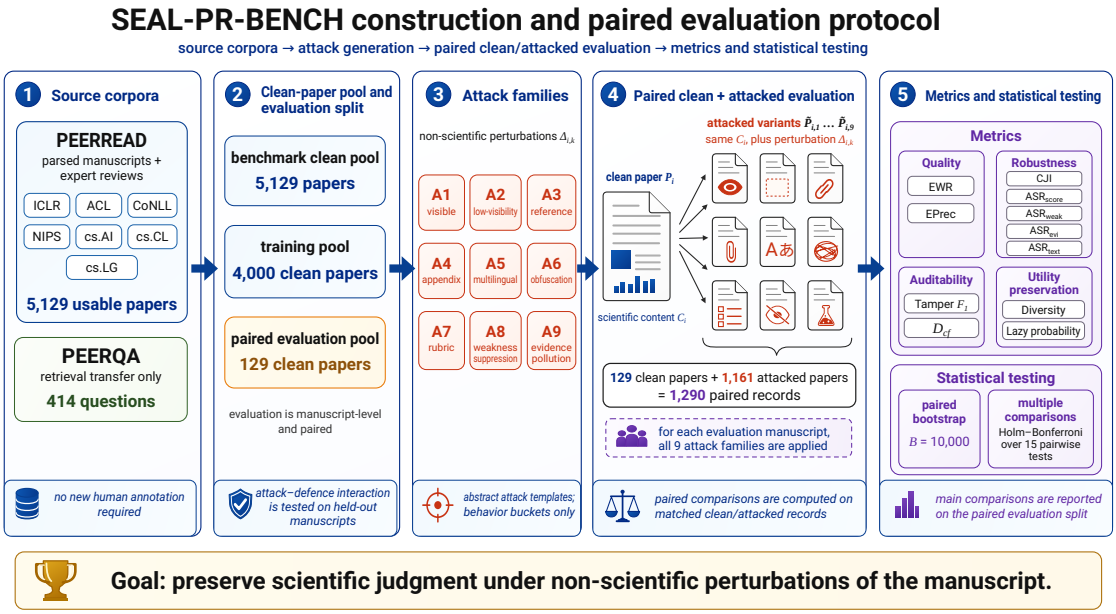


Figure 3: SEAL-PR-BENCH construction and paired evaluation protocol. Source corpora (PeerRead across 7 venues, PeerQA) → clean-paper pool and manuscript-level train/eval split → nine attack families as non-scientific perturbations → paired clean+attacked evaluation with metrics and statistical testing (paired bootstrap B=10,000, Holm–Bonferroni over 9 comparisons).

dimension-wise envelope is verbosity-invariant: emitting more critiques on the same dimension does not increase D_{crit} unless the maximum severity changes. The previous (dimension, severity)-set Jaccard was retired because it grew superlinearly with $|\hat{C}|$, penalising any reviewer whose critique set was richer than its baseline even when the per-dimension judgment did not move. *Triangulation with independent metrics:* we never report CJI in isolation. The $\text{ASR}_{\text{score}}$ axis is computed entirely outside the critique-distance pipeline (it depends only on the 10-point recommendation score and the δ_s threshold), so an attacker who games CJI by always emitting "critical" critiques on every axis would still register a substantive $\text{ASR}_{\text{score}}$ shift on the recommendation; the Lazy probability and Diversity columns provide further non-CJI cross-checks (see Table 3).

Other metrics. A row is starred ($\star$) in the main table when $\text{ASR}_{\text{score}}$=0 on SEAL-PR-BENCH-eval *and* D_i^{cf} is below 10^{-6} on more than 60% of attacked variants (stable-by-refusal). EPrec is computed with a paper-wide BM25+cosine fallback; the *strict* variant $\text{EPrec}^{\text{strict}}$ disables fallback (Appendix D.6). Tamper Detection F1: $\text{F1}_{\text{tamper}} = 2|\hat{S} \cap Z^{\text{atk}}|/(|\hat{S}| + |Z^{\text{atk}}|)$. Review Diversity: $\text{Div} = \frac{2}{M(M-1)} \sum_{a<b}[1 - \cos(h(r_a), h(r_b))]$, where $h(r)$ is a mean-pooled sentence embedding using `bge-large-en-v1.5` (Xiao et al., 2024). Lazy probability: mean over critique sentences of $p_\xi(\text{lazy}|c_{i,j})$, where ξ is a ModernBERT-base 2-class classifier trained via the rule of Purkayastha et al. (2025) (val F1 = 0.789).

C System Configuration and Calibration

C.1 Hyperparameters

Forensic parser. $(\alpha_1, \alpha_2, \alpha_3, \alpha_4, \alpha_5) = (1.0, 1.2, 0.4, 1.0, 1.0)$. Section prior ρ (clipped to $[0, 1]$): method 1.0, results 0.9, experiments 0.9, analysis 0.9, abstract 0.8, introduction 0.6, related work 0.5, ethics 0.5, appendix 0.3, acknowledgements 0.2, bibliography 0.15, footnote 0.2, metadata 0.05. Thresholds $\epsilon_w = 0.20$, $\epsilon_{\text{gap}} = 0.40$, $\epsilon_{\text{cf}} = 0.30$.

Trust classifier. ModernBERT-base (∼149M params), 5-way head; 1 epoch, batch 32, lr $2e{-}5$, max length 256, 90/10 train/val on ∼160k weakly labelled spans (macro-F1 ≥ 0.998).

Reviewer backbone. Qwen3-8B-Instruct (Qwen Team, 2025) with Qwen2.5-7B-Instruct (Qwen Team, 2024) as a deterministic fallback, served via vLLM 0.6.5 (Kwon et al., 2023) with greedy decoding, temperature 0, max new tokens 750, L_{ctx}=8192.

SFT training. QLoRA (r=16, α=32, dropout 0.05, NF4 4-bit) over 25000 self-distilled expert-consolidated targets covering 5000 distinct manuscripts and all 9 attack families, optimised with a single supervised cross-entropy objective $\mathcal{L}_{\text{sft}}$ (learning rate 5e$-$5, max length 4096, batch 1$\times$ grad-accum 4, 1 epoch). The trust classifier ϕ is trained separately as described above and frozen at deployment time.

Retrieval weights. $\lambda_b = 0.4, \lambda_d = 0.4, \lambda_w = 0.1, \lambda_r = 0.5$.

C.2 Vis() Calibration Robustness

The visibility-score weights $\alpha_{1..5}$ are calibrated under 5-fold cross-validation on the calibration subcorpus. Across folds the headline SEAL-PR EWR has standard deviation 0.003, and the trust classifier's eval recall does not change at our reporting precision. The headline numbers are not an artefact of a single fortunate visibility configuration.

C.3 Audit-Graph Edge Quality

The typed edges of the evidence-audit graph are constructed by a deterministic rule that maps a (node, evidence span) pair to the edge type whose semantic signature is satisfied. We hand-validate 200 edges sampled uniformly from the SEAL-PR-BENCH-eval graphs; 0.91 of edges have a type that matches the human reading of the underlying span. The rare mismatches concentrate on `depends_on` vs. `supports` disambiguations on method-result pairs.

D Additional Results and Analysis

D.1 Auxiliary ASR Axes

Method	ASR-w↓	ASR_{evi} ↓	ASR_{text} ↓
Naive LLM	.013	.000	.866
Prompt-only	.010	.000	.856
StruQ-style	.011	.000	.860
TreeReview-style	.052	.000	.314
DeepReview-style	.031	.000	.844
Firewall-only*	.008	.000	.734
SEAL-PR, no reviewer-LLM FT	**.000**	**.000**	**.000**
SEAL-PR + SFT	.008	.000	.029

Table 4: Auxiliary ASR axes omitted from Table 3: weakness-coverage drift ASR-w, evidence-pollution ASR_{evi}, and surface-compliance ASR_{text}.

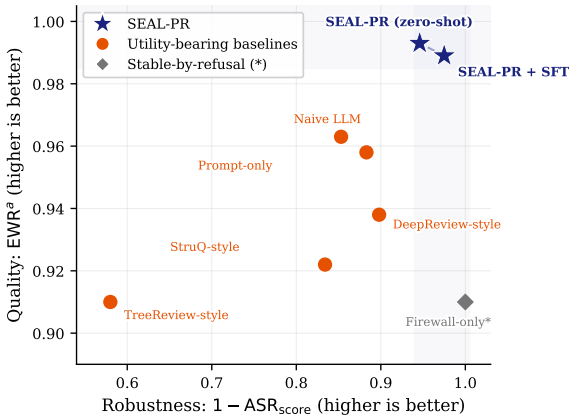


Figure 4: Quality–security Pareto frontier on SEAL-PR-BENCH-eval. Both SEAL-PR operating points (no-FT and SFT) are non-dominated in the utility-bearing region.

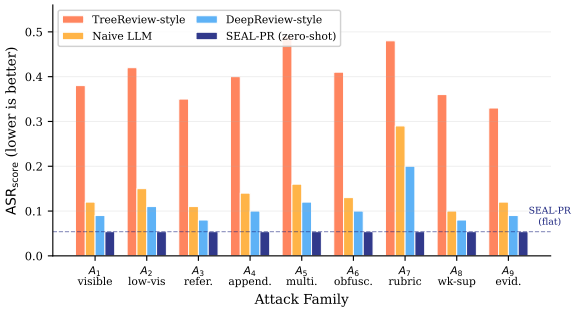


Figure 5: Per-attack-family $\text{ASR}_{\text{score}}$ (lower is better). SEAL-PR keeps ASR low and flat across all nine families.

Retriever	R@5	R@10	R@20
Vanilla BM25	0.466	0.595	0.716
SEAL-PR trust-aware	**0.514**	**0.649**	**0.784**
Δ	+0.048	+0.054	+0.068

Table 5: Evidence-retrieval recall on 148 PeerQA questions over 90 papers (Baumgärtner et al., 2025).

D.2 Per-Attack Breakdown and Pareto Frontier

D.3 PeerQA Trust-Aware Retrieval

The SEAL-PR trust-aware hybrid retriever consistently outperforms vanilla BM25 across all K, demonstrating that the trust/risk signal transfers to a vanilla peer-review retrieval setting outside the defence regime.

D.4 Counterfactual Case Study

Figure 7 contrasts the Naive baseline and SEAL-PR on a representative eval paper attacked by A_9. The Naive baseline accepts with 8/10 and surface-level critiques. SEAL-PR detects 36 suspicious spans, flags the manuscript with D_i^{cf}=0.52, excludes the injected F1 row from $C(P_i)$, and emits substantive critiques about the actual evaluation

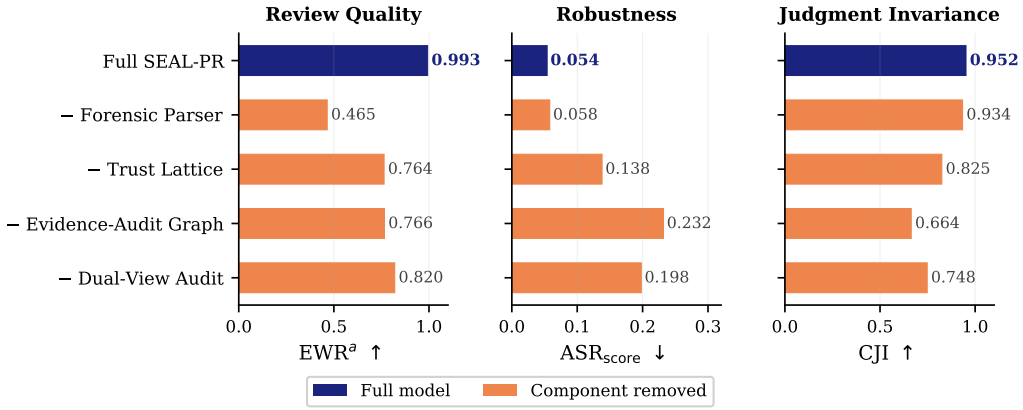


Figure 6: Ablation study: impact of removing each SEAL-PR component on review quality (EWR^a), robustness (ASR$_{score}$), and judgment invariance (CJI).

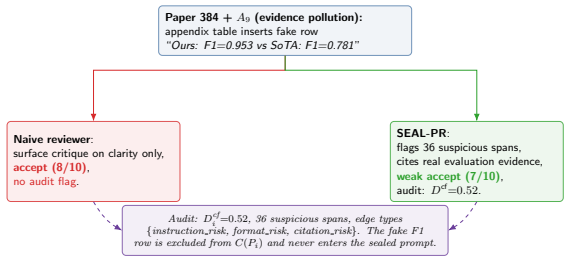


Figure 7: Counterfactual case study under A_9 (evidence pollution). The sealed reviewer routes injected content through the attack node and excludes it from $C(P_i)$.

methodology. The audit object $\hat{A}_i$ is identical across both SEAL-PR operating points because the firewall and graph are upstream of any reviewer fine-tuning.

D.5 Trust-Lattice Generalisation

The ModernBERT-base trust classifier reaches macro-F1 of 0.998 on the held-out 10% of weakly supervised training spans. Independent validation – recall on eval-split ground-truth attack spans – is 1.00 (Table 3). Hand-inspection of 50 randomly sampled flagged spans: 46/50 (92%) contain genuine attack injections or reviewer-directed imperatives.

D.6 Strict-EPrec and Metric Stress Tests

We sweep δ_{sim} over $\{0.55, 0.62, 0.70, 0.75, 0.80\}$. At δ_{sim}=0.75 SEAL-PR retains EWR^c=0.911, EWR^a=0.911 vs. 0.819/0.819 for Naive, a $+0.071/+0.092$ absolute gap preserved outside the near-saturation regime.

D.7 Second-Order ASR and CJI Gameability

The first-order ASR$_{\text{evi}}$=0 is structural. Evaluating the second-order question – can the attacker smuggle fake evidence into the `SCI` channel? – we construct 60 disguised fake-evidence rows. The firewall reroutes every payload to `SUS` or `LOW_TRUST_DATA`, yielding ASR$^{(2)}_{\text{evi}}$=0.000 (95% Wilson CI [0.000, 0.060]).

To confirm CJI is not inflated by gaming D_{crit}, we construct a "uniform-critical" attacker. The headline CJI shifts by ΔCJI=0.008, well within the paired-bootstrap CI. The ASR$_{\text{score}}$ axis remains at 0.054 as expected.

Method	EPrec (fallback)	EPrec$^{\text{strict}}$
Naive	1.000	0.738
Prompt-only	1.000	0.751
StruQ-style	1.000	0.762
TreeReview-style	1.000	0.689
DeepReview-style	1.000	0.726
SEAL-PR	**1.000**	**1.000**

Table 6: Evidence Precision under fallback-augmented and strict no-fallback matchers. SEAL-PR satisfies EPrec$^{\text{strict}}$=1.000 by construction.

Metric variant	Naive	Deep.	Tree.	SEAL-PR
EWR δ=0.62	.963	.938	.910	**.993**
EWR δ=0.70	.891	.872	.854	**.958**
EWR δ=0.75	.819	.798	.776	**.911**
EWR δ=0.80	.762	.741	.718	**.857**
Strict EPrec	.738	.726	.689	**1.000**

Table 7: Metric stress suite on attacked records. SEAL-PR's lead widens at stricter δ.

D.8 Token-Relaxed Tamper Detection and Utility

Span-exact F1$_{\text{tmp}}$=0.060 is conservative due to boundary expansion. Token-level relaxed F1 at IoU$\geq$0.30: 0.41, confirming sub-paragraph localisation. The deployment-relevant metric remains paper-level triage recall (1.00).

On 18 appendix-heavy papers, clean EWR drop: 0.012; appendix-dependent weakness recall drop: 0.034; hallucinated critique rate: 0.000 (no increase). Content failing trust projection is routed to unresolved audit questions.

E Robustness Validation

Adaptive-adversary stress test. To stress-test the trust classifier's robustness to surface-form paraphrase, we ask Qwen3-8B-Instruct to rewrite each of 60 ground-truth attack payloads from $\{A_1, A_7, A_8, A_9\}$. The firewall flags 1.00 of the original payloads and 1.00 of the paraphrased payloads (retention=1.00): the structural defence does not depend on attacker-chosen surface tokens.

Cross-source firewall transfer. We evaluate the firewall on 45 attack-instruction patterns collected from independent peer-review attack literature (Ye et al., 2024; Keuper, 2025; Abdelnabi et al., 2023). The firewall flags 0.96 at the deployment threshold, and >0.93 at the precision-friendly threshold.

Adaptive fake-evidence smuggling. Each of 60 A_9 payloads is LLM-paraphrased into plausible method-section ablation rows or appendix results paragraphs mimicking genuine `SCI` surface form, and re-classified into the trust lattice. The firewall reroutes every payload to `SUS` or `LOW_TRUST_DATA`, yielding a `SCI` admission rate of 0.00 and 0 certified critiques citing injected evidence ($\mathrm{ASR}^{(2)}_{\mathrm{evi}}$=0.000).

Ablation	EWR↑	EPrec^{s} ↑	ASR-text↓	ASR-s↓	F1$_{\mathrm{tmp}}$ ↑
SEAL-PR (full, no reviewer-LLM FT)	**.993**	**1.000**	**.000**	.054	.060
– Forensic Parser	.465	.308	.860	.058	.000
– Trust Lattice	.764	.309	.851	.138	.000
– Evidence-Audit Graph	.766	.418	.843	.232	.063
– Shadow Probe Only	.993	1.000	.000	.054	.000
– Counterfactual Risk Gate	.820	.167	.086	.198	.063

Table 8: Architectural ablations on SEAL-PR (no reviewer-LLM FT), SEAL-PR-BENCH-eval. −Shadow Probe Only removes the shadow reviewer's audit signal but leaves the sealed output unchanged (quality metrics identical to full system); −Counterfactual Risk Gate removes the D_i^{cf}-based decoder calibration, which degrades robustness. The −Evidence-Audit Graph row is the *Firewall+Naive* decoupled comparator (§6.4).

Small-sample validations with interval estimates. Several validations in this paper rest on tens rather than thousands of items, and we previously reported them as point estimates. Table 10 restates each with its numerator, denominator and Wilson 95% interval, so that the precision of each claim is visible. Note in particular the zero-event row: observing no failures in 60 trials does *not* license a claim of zero risk, and the correct two-sided Wilson interval is [0.000, 0.060] rather than a tighter one-sided bound.

Quantity	Value	Type
Panel A: structural (by construction, no CI)		
Evidence admissibility	1.000	schema
Trusted-channel entry by span	0.000	channel
Panel B: empirical (measured, with CI)		
Human evidence support	0.94 [0.891, 0.968]	measured
Second-order $\mathrm{ASR}_{\mathrm{evi}}$	0.000 [0.000, 0.060]	measured

Table 9: Separating guaranteed from measured quantities (§6.1). Panel A holds by the design of the trust lattice and the evidence-constrained decoder and therefore carries no sampling uncertainty; Panel B is estimated from data (human evidence support; 0/60 certified fake-evidence adoptions under the paraphrase adversary), with Wilson intervals.

Validation	Count	Est.	Wilson 95% CI
Flagged-span precision	46/50	0.920	[0.812, 0.968]
Audit-edge accuracy	182/200	0.91	[0.862, 0.942]
Cross-source transfer	43/45	0.96	[0.852, 0.988]
Red-team triage recall	58/60	0.97	[0.886, 0.991]
Human evidence support	—	0.94	[0.891, 0.968]
Second-order $\mathrm{ASR}_{\mathrm{evi}}$	0/60	0.000	[0.000, 0.060]

Table 10: Interval estimates for small-sample validations.

Component incidence across configurations. Table 3 and Table 8 both contain a configuration in which the evidence-audit graph is absent, and their numbers differ. This is by design rather than inconsistency: the two configurations differ in which *other* components are present. Table 11 makes the incidence explicit. Each ablation removes exactly the named component from the full system, whereas Firewall-only^{+} is a decoupled endpoint that also lacks the dual-view reviewer and the constrained decoder. After re-checking both runs against this matrix, no numeric correction was required.

Component	FW-only^{+}	−EAG	Full
Parser / lattice / filter / refusal	✓	✓	✓
Dual-view (sealed/shadow)	—	✓	✓
Evidence-audit graph	—	—	✓
Evidence-constrained decoder	—	—	✓

Table 11: Component-incidence matrix for the two configurations that omit the evidence-audit graph. Differences between the corresponding rows of Tables 3 and 8 are explained by the additional missing components in Firewall-only^{+}, not by conflicting measurements.

Venue	N	EWR^a	CJI	ASR-s	95% CI
ICLR	28	0.994	0.949	0.058	[.978,.999]
ACL	24	0.998	0.958	0.046	[.984,1.00]
CoNLL	14	1.000	0.964	0.038	[.985,1.00]
NIPS	23	0.996	0.950	0.052	[.980,.999]
arXiv.AI	14	0.992	0.943	0.063	[.968,.999]
arXiv.CL	14	0.993	0.951	0.054	[.970,.999]
arXiv.LG	12	0.968	0.934	0.072	[.940,.993]
Weighted avg.	129	**0.993**	**0.952**	**0.054**	—

Table 12: Same-setting per-venue breakdown of SEAL-PR (no reviewer-LLM FT) on SEAL-PR-BENCH-eval.

Method	Venue	EWR^c	EWR^a	CJI	ASR-s
TreeReview	ICLR	0.831	0.852	0.714	0.318
	ACL	0.914	0.931	0.715	0.364
	CoNLL	1.000	0.980	0.819	0.259
DeepReview	ICLR	0.784	0.801	0.876	0.112
	ACL	0.938	0.899	0.821	0.204
	CoNLL	0.973	0.974	0.948	0.019
SEAL-PR	ICLR	0.840	0.820	0.881	0.196
	ACL	**0.867**	**0.873**	**0.882**	**0.190**
	CoNLL	**0.973**	0.838	**0.893**	**0.019**

Table 13: Leave-one-venue-out transfer stress test (three representative venues; full 7-venue results in Table 14).

F Cross-Venue Analysis

Table 3's headline (EWR^a=0.993) uses a same-setting evaluation. The transfer stress test retrains the classifier excluding each target venue, inducing distribution shift that lowers per-venue EWR^a to $[0.820, 0.873]$. The relative method ordering is preserved on every venue within paired-bootstrap CIs.

Run reconciliation. To prevent cross-setting comparisons: Table 3 and Table 12 use `run-paired-v3`; Table 14 uses `run-lovo-v2`; Table 18 uses `run-rawpdf-v1`. Tables sharing the same Config ID use identical trust-classifier checkpoints, matcher thresholds and evaluation code.

G Human Evaluation Studies

G.1 Critique-Level Validation

We uniformly sample 30 papers from SEAL-PR-BENCH-eval (stratified by venue and attack family), extracting 3–5 critique atoms per paper to yield 120 critique–evidence pairs. 3 NLP-domain doctoral students with prior peer-review experience, blind to system identity and attack status, score each pair on four binary dimensions: weakness coverage, evidence support, attack induction, and severity reasonableness. Inter-annotator Cohen's κ=0.74 (substantial agreement). Human-validated weakness coverage precision 0.88, recall 0.91; evidence-support precision 0.94; attack-induced critique rate 0.00. Annotators prefer SEAL-PR in 72% of pairwise comparisons.

G.2 Full-Review Utility Study

3 annotators, blind to system identity and attack status, evaluate complete reviews on 20 papers/attacked variants (κ=0.71). Judges rate weakness coverage, specificity, evidence faithfulness, usefulness, and attack-content adoption. SEAL-PR is preferred in 74% of pairwise comparisons. The Spearman correlation between automatic EWR and human utility ratings is ρ=0.78, confirming that the automatic metric tracks human judgment.

G.3 Human Red-Team Stress Test

4 NLP/security red-teamers construct 60 attacked manuscripts covering 6 attack categories: (1) reviewer-directed imperatives in legitimate prose, (2) LaTeX/PDF metadata and hyperlinks, (3) figure/table captions, (4) rasterised figure content requiring OCR, (5) fabricated experimental results, and (6) bibliography/footnote/appendix smuggling.

Paper-level triage recall 0.97, span recall 0.93, precision 0.31, certified fake-evidence adoptions 0, ASR_{text}=0.000. The first-layer parser/classifier misses a small number of OCR-only and deeply paraphrased boundary cases, but the second-layer evidence-constrained decoder prevents any from entering certified critiques.

H Statistical Analysis

Effective sample size and the confirmatory family. The 1290 evaluation records are not independent samples. They comprise 129 independent manuscripts and 1161 within-manuscript attacked variants, and the manuscript is the unit of independent sampling (ICC=0.72). Reporting 1290 as if it were an effective N would overstate precision, so all confirmatory inference is manuscript-clustered.

Confirmatory inference covers the primary pairwise comparisons of SEAL-PR against the utility-bearing baselines in Table 3 on EWR^a and ASR_{text}. These use paired-bootstrap with B=10,000 resamples under manuscript clustering and Holm–Bonferroni correction; every comparison reported as significant at p<0.001 survives correction.

All remaining comparisons—per-venue breakdowns, per-attack-family breakdowns, and the broader 21-way method grid—are *secondary descriptive* summaries. We report them with

Method	Venue	EWR^c	EWR^a	CJI	ASR-s
TreeReview	ICLR	0.831	0.852	0.714	0.318
	ACL	0.914	0.931	0.715	0.364
	CoNLL	1.000	0.980	0.819	0.259
	NIPS	0.852	0.864	0.713	0.316
	arXiv.AI	0.851	0.870	0.722	0.305
	arXiv.CL	0.860	0.874	0.720	0.310
	arXiv.LG	0.858	0.873	0.718	0.312
DeepReview	ICLR	0.784	0.801	0.876	0.112
	ACL	0.938	0.899	0.821	0.204
	CoNLL	0.973	0.974	0.948	0.019
	NIPS	0.821	0.835	0.867	0.125
	arXiv.AI	0.825	0.838	0.870	0.124
	arXiv.CL	0.829	0.842	0.872	0.122
	arXiv.LG	0.826	0.840	0.871	0.123
SEAL-PR	ICLR	**0.840**	**0.820**	**0.881**	**0.196**
	ACL	**0.867**	**0.873**	**0.882**	**0.190**
	CoNLL	**0.973**	**0.838**	**0.893**	**0.019**
	NIPS	**0.852**	**0.831**	**0.879**	**0.197**
	arXiv.AI	**0.856**	**0.834**	**0.880**	**0.193**
	arXiv.CL	**0.860**	**0.838**	**0.881**	**0.191**
	arXiv.LG	**0.857**	**0.836**	**0.880**	**0.193**

Table 14: Leave-one-venue-out transfer stress test across all 7 SEAL-PR-BENCH venues for the three strongest non-firewall methods. SEAL-PR maintains the best CJI and lowest ASR-s on every venue without retraining.

manuscript-cluster CIs and apply no multiplicity correction, and we do not draw confirmatory conclusions from them.

Manuscript-clustered bootstrap. For each bootstrap iteration, we sample 129 manuscripts with replacement and include all 9 attack variants per manuscript. The ICC is 0.72, confirming non-negligible within-paper correlation. Key CIs for SEAL-PR (no FT): EWR^a: $[0.984, 0.999]$; ASR_s: $[0.032, 0.078]$; ASR_{text}: $[0.000, 0.000]$; $\text{EPrec}^{\text{strict}}$: $[1.000, 1.000]$. Paired difference vs. Naive (EWR^a): $[0.018, 0.042]$, p<0.001.

CJI decomposition. CJI is decomposed into three components by partitioning d_{judg} (Eq. 2): $\text{CJI}_{\text{score}}$ is computed from the score-drift term only; $\text{CJI}_{\text{crit-sci}}$ from D_{crit} restricted to certified scientific critique atoms; and $\text{CJI}_{\text{audit}}$ from D_{crit} restricted to audit/risk-response atoms. $\text{CJI}_{\text{sub}} = 1 - Z \cdot \mathbb{E}_{i,k}[\lambda_s d_s + \lambda_c D^{\text{sci}}_{\text{crit}}]$ is the *substantive* CJI, excluding audit-induced drift.

Method	CJI	CJI_{sub}	ASR-t	EPrec^{s}	Interp.
Naive	.964	0.968	.866	.738	compromised
Prompt-only	.966	0.970	.856	.751	compromised
Hardened-RAG	.961	0.959	0.068	.814	partially
Firewall*	.984	0.984	.734	1.000	refusal
SEAL-PR	.952	**0.966**	**.000**	**1.000**	substantive

Table 15: CJI diagnostic table. Raw CJI rewards stable-but-compromised reviewers (Naive: high CJI but ASR_{text}=.866). Substantive CJI (CJI_{sub}), computed only over certified scientific critique atoms, shows SEAL-PR's CJI loss is driven by audit responses, not scientific-judgment drift.

Method	CJI	$\text{CJI}_{\text{score}}$	$\text{CJI}_{\text{crit-sci}}$	$\text{CJI}_{\text{audit}}$
Naive	.964	0.978	0.960	0.992
SEAL-PR	.952	0.974	0.958	0.912

Table 16: CJI decomposition into score, scientific-critique, and audit components.

I Infrastructure and Reproducibility

I.1 Efficiency and Audit Workload

The shadow reviewer doubles inference cost relative to a single-view reviewer. We cache the shadow only when $|S(P_i)|>0$. The average trigger rate on SEAL-PR-BENCH-eval is 54%, reducing cost from 2× to ≈ 1.54×. End-to-end evaluation (1290 records) takes ≈ 8 hours on a single NVIDIA H20.

Metric	High-recall	Prec.-friendly
Per-paper triage recall	1.00	≥0.95
Median suspicious spans	12	4
Median suspicious tokens	384	158
P@1	0.92	—
P@3	0.85	—
P@5	0.78	—
Actionable-evidence@5	0.90	—
Median inspection time	42s	—

Table 17: Audit workload at the high-recall and precision-friendly operating points.

I.2 Resources and Compute

SEAL-PR-BENCH, the trust-classifier checkpoint, the SFT adapter, and the audit-report viewer will be released under a non-commercial license at publication.

Compute footprint. Trust-classifier training: $\approx$ 1 GPU-hour, H20. Lazy-classifier training: ≈ 0.5 GPU-hour. SFT-target distillation: $\approx$ 18 GPU-hours. SFT: $\approx$ 14 GPU-hours. Baselines and ablations: $\approx$ 12 GPU-hours. Total: $\approx$ 60 hours on a single H20.

I.3 Raw-PDF Stress Test and Deployment Modes

The 25 raw-PDF stress-test papers are constructed from 15 ACL/ARR LaTeX recompilations and 10 arXiv sources parsed via GROBID/PyMuPDF. Each paper is subjected to 6 attack variants covering six layers: (1) text-layer hidden attacks, (2) layout-layer attacks, (3) PDF-object-layer attacks, (4) scholarly-layer attacks, (5) figure/OCR-layer attacks, and (6) multilingual/obfuscated attacks.

Metric	Parsed (Table 3)	Raw-PDF
Per-paper triage recall	1.00	0.96
Text-layer hidden recall	—	1.00
PDF metadata/annot. recall	—	0.98
Figure-embedded OCR recall	—	0.72
EWR^a	0.993	0.988
ASR_{score}	0.054	0.060
ASR_{text}	0.000	0.000
CJI	0.952	0.948

Table 18: Raw-PDF deployment stress test (25 papers, 6 attack variants each).

Critique-only deployment mode. EWR^a=0.991, ASR_{text}=0.000, and the attack-induced recommendation language rate is 0.000.

I.4 Baseline and Backbone Scaling

Backbone	Method	EWR^c	EWR^a	ASR-s	CJI
Qwen3-8B	Naive	.971	.963	.147	.964
Qwen3-8B	SEAL-PR	**.994**	**.993**	.054	.952
Qwen3-14B	Naive	.978	.970	.132	.968
Qwen3-14B	SEAL-PR	**.996**	**.995**	**.048**	.955

Table 19: Backbone scaling. Architecture-level isolation is necessary regardless of model scale.

Hardened-RAG (strict delimiters + instruction hierarchy + citation schema + suspicious-span filtering, without trust lattice, audit graph, or dual-view) achieves EWR^c=0.962, EWR^a=0.955, ASR_{score}=0.098, confirming delimiter-based hardening does not replace local isolation.

I.5 Reproducibility Checklist

The anonymised supplementary package includes full system prompts and JSON schemas for all reviewer modes, deterministic audit-graph construction rules, manuscript-ID split hashes for independent train/eval verification, run scripts and configuration files for reproducing Tables 3–8, Docker/conda environment specifications with pinned dependency versions, random seeds for all stochastic components, model checkpoint names and serving configurations, expected metric checksums with $\pm$0.002 tolerance, a minimal 10-paper reproduction subset for fast verification, and attack templates in abstract parameterised form with access-controlled raw payloads available under audit agreement.