Fighting Low-Effort LLM-Assisted Reviews Requires Continued Investigation of Prompt Injection

Xuanli He[1,2*] **Weijun Li**[1*] **Timothy Chard**[1] **Mark Dras**[1] **Qiongkai Xu**[1†]

[1]Macquarie University, Sydney, NSW, Australia
[2]University College London, London, UK
xuanli.he@ucl.ac.uk weijun.li1@hdr.mq.edu.au qiongkai.xu@mq.edu.au

Abstract

The rapid growth of submissions to AI venues has intensified concerns about the integrity and scalability of peer review. As large language models (LLMs) become increasingly capable, reviewers may rely on automated evaluation with limited verification. We refer to the resulting incentive structure as *low-effort rational reviewing*: situations in which automated assistance substantially reduces review effort while accountability remains difficult to verify. Such practices may weaken expert accountability and create opportunities for adversarial manipulation. In response, the community has begun exploring countermeasures, exemplified by the ICML 2026 experiment embedding hidden prompt injections to detect unauthorized LLM use. This position paper argues that preserving review integrity may benefit from complementing *deterrence-centered governance*, focused on misuse detection and enforcement, with a *compliance-by-design* perspective that addresses the incentives and workflows surrounding LLM-assisted reviewing. While existing countermeasures primarily aim to identify unauthorized use after the fact, such approaches may face practical limitations, including unreliable triggering, low-cost circumvention, and uncertain enforcement. More broadly, we argue that the challenge should be viewed not solely as a detection problem, but as a mechanism-design problem: how can review workflows be structured such that unverified delegation to LLMs becomes less attractive than careful human evaluation? To this end, we explore how ideas from prompt-injection security can support workflow-centered interventions. In particular, we discuss mechanisms such as distributed prompt injection, watermarking, and extraction friction that may reduce the practical advantages of unverified LLM reviewing while remaining transparent to human-grounded evaluation. We conclude by outlining a research agenda for compliance-by-design approaches that can complement existing governance mechanisms and help sustain rigorous peer review in the age of increasingly capable AI systems.

1 Introduction

Peer review is the central mechanism for validating scientific progress, yet it is under growing strain in modern machine learning venues. Major AI conferences now receive on the order of 10,000 submissions, placing substantial pressure on reviewers and consequently contributing to highly variable review quality [Kim et al., 2025]. Empirical studies show that reviews often vary widely in depth and correctness; some are shallow or weakly grounded, and even a small number of low-quality

*Equal contribution.
†Corresponding author.

Preprint, June 2026.

reviews can substantially affect acceptance outcomes [Ross et al., 2006, Centeno et al., 2015, Gao et al., 2019, Russo, 2021, Shah, 2022].

This fragility is driven by structural factors, including rapid growth in submissions [Huang et al., 2023, Maslej et al., 2024], lower barriers to manuscript writing recently amplified by Large Language Models (LLMs) [Liang et al., 2024a,b], and academic incentives that reward frequent publication [van Dalen and Henkens, 2012, Stephan, 2012]. At the same time, the pace of research introduces knowledge gaps even for experts [Wu et al., 2021], while weak accountability allows low-quality reviews to persist. Recent evidence suggests that LLM-assisted reviewing may already be occurring, with observable shifts in review patterns [Liang et al., 2024a], and existing detection tools remain unreliable [Elkhatat et al., 2023].

The emergence of LLMs amplifies existing challenges in peer-review governance. While they can assist with legitimate review tasks, they also enable the generation of coherent and plausible reviews with substantially reduced effort, creating opportunities for *judgment replacement*, in which model-generated evaluations substitute for independent human assessment. Recent empirical work suggests that LLM-generated reviews exhibit systematic weaknesses, including excessive within-paper agreement and limited ability to distinguish genuine scientific merit from superficial stylistic changes [Baumann et al., 2026]. These findings raise concerns that unverified reliance on LLM-generated evaluations may undermine review quality and render the process more vulnerable to gaming. Recognizing these risks, the community has begun exploring countermeasures, exemplified by the ICML 2026 experiment embedding hidden prompt injections in submissions to detect unauthorized LLM use [Kamath, 2026]. However, as we argue, such deterrence-centered approaches may be limited by their reactive nature.

Position. Peer-review integrity cannot rely solely on deterrence and post-hoc detection; it requires a shift from deterrence-centered governance toward compliance-by-design. Existing governance mechanisms largely focus on identifying or penalizing undesirable behavior after it occurs. However, LLM-assisted reviewing changes the underlying cost structure of review generation, making it easier to produce plausible evaluations without proportional human effort. As long as this cost asymmetry persists, detection alone may leave unverified delegation attractive. Integrity mechanisms must therefore be embedded into the review environment, shaping the conditions under which reviewing occurs rather than relying only on retrospective enforcement. This perspective aligns with policy design and regulatory theory [Lessig, 1999, Thaler and Sunstein, 2008, Cavoukian, 2009], which emphasizes that system architecture can itself function as a regulatory mechanism.

The key issue is not LLM assistance itself, but *judgment replacement*: coherent model-generated evaluations come to substitute for independent human assessment despite compromised reliability. We term one manifestation of this failure mode **low-effort rational reviewing**: situations in which reviewers rely primarily on unverified LLM outputs because doing so substantially reduces effort while accountability remains difficult to verify.

We explore compliance-by-design mechanisms that introduce friction to diminish the utility of unverified LLM reviewing. These mechanisms, including distributed prompt injection, watermarking, and extraction friction, degrade unverified LLM outputs while remaining non-disruptive to human evaluation, making low-effort reviewing less appealing than enforced human review.

In summary, we (1) argue that deterrence-centered governance alone may be insufficient because it does not alter the underlying incentives that make low-effort rational reviewing attractive; and (2) explore compliance-by-design mechanisms that shift the equilibrium toward verification-based review.

2 Deterrence-Centered Governance: Mechanisms and Limitations

Recent responses to concerns about unverified LLM use in peer review have largely focused on deterrence-centered governance. Rather than changing how reviews are produced, these approaches seek to preserve review integrity through detection and enforcement. Their central premise is that sufficiently reliable detection, combined with meaningful penalties, can discourage unauthorized automation. The ICML 2026 review process represents the most visible large-scale implementation of this idea. In this section, we use ICML 2026 as a representative case study to examine both the mechanisms and limitations of deterrence-centered governance.

2.1 The ICML 2026 Experiment: A Case Study

ICML 2026 operationalized this deterrence-centered approach through a two-policy framework governing LLM use in reviewing [Kamath, 2026]. Policy A prohibited all LLM assistance, while Policy B permitted limited and disclosed use of LLMs. Enforcement focused on identifying violations of Policy A. To this end, ICML embedded hidden prompt injections in submitted PDFs and monitored reviews for evidence that the LLM had followed these instructions. Reviews exhibiting triggered signals were manually verified, and confirmed violations resulted in substantial penalties: papers authored by reciprocal reviewers who violated the policy were desk rejected without review. At scale, the mechanism identified 795 violating reviews (approximately 1% of all reviews), resulting in 497 desk rejections—roughly 2% of total submissions. In deterrence-theoretic terms, the intervention sought to increase the expected cost of unauthorized LLM use by combining automated detection with meaningful sanctions.

The current approach relied on a dictionary-based watermarking scheme [Rao et al., 2025], in which two randomly sampled phrases were embedded as instructions "hidden" in the footer of each PDF—invisible to human readers but visible to LLMs—instructing the model to include both phrases in generated reviews. Reviews were scanned for the presence of both watermark phrases, with the detection threshold calibrated to achieve a family-wise error rate of 0.0001 [Kamath, 2026], corresponding to a very low probability of producing any false positive among submitted reviews. According to the pre-deployment experiments conducted by the organizers, frontier LLMs followed the injected instructions with success rates exceeding 80% for most models, though success varied by ingestion method and model.

Here, we argue that deterrence-centered approaches face limitations that undermine their effectiveness. We examine these limitations from two perspectives: (1) *incentive structure*, where deterrence fails to alter the cost–benefit equilibrium that makes low-effort reviewing attractive; and (2) *technical robustness*, where detection mechanisms prove fragile under realistic reviewing workflows.

2.2 Incentive Fragility: The Cost-Benefit Calculus

Deterrence-centered governance assumes that the threat of punishment will discourage violations. However, this assumption overlooks a critical factor: reviewers engage in rational calculation, weighing punishment risk against effort savings. To formalize this, we model reviewing strategies along two dimensions: **effort cost** C (time and cognitive load) and **output reliability** R (alignment with well-informed human evaluation). A low-effort regime exists when sufficiently high perceived R can be achieved at low C.

While deterrence mechanisms introduce an expected punishment cost, they do not fundamentally restructure the core cost (C) and reliability (R) dynamics of the underlying workflow. Low-effort reviewing with LLMs remains efficient: reviewers can generate plausible outputs in a fraction of the time required for rigorous evaluation, preserving low C. Meanwhile, LLM outputs appear coherent and well-structured despite compromising actual quality, maintaining perceived high R. The punishment risk manifests as an uncertain, probabilistic cost rather than a guaranteed barrier.

From the reviewer's perspective, this creates a gamble: accept minimal detection risk in exchange for substantial time savings. Under realistic time constraints, when reviewers face mounting workloads and tight deadlines, the immediate benefit of reduced effort can outweigh the uncertain prospect of future punishment. Deterrence does not eliminate the appeal of low-effort reviewing; it merely adds a calculated risk, leaving the strategy rationally viable for those operating under severe time pressure.

Consequently, deterrence-centered approaches function as **forensic filters**: they enable post hoc identification of certain violations but do not eliminate the low-effort equilibrium. As long as C remains low and perceived R remains acceptable, unverified LLM reviewing persists as a rational strategy under time constraints.

2.3 Technical Fragility: Evidence from Controlled Experiments

Beyond the incentive structure problem, deterrence-centered detection faces another fundamental limitation: technical fragility under realistic reviewing workflows. The current injection mechanism assumes that LLMs will reliably process and follow those embedded instructions. However, this assumption is brittle due to the multi-channel and model-dependent nature of LLM-assisted reviewing.

Table 1: Detection robustness under realistic reviewing workflows. Success rate is the percentage of 20 papers for which both watermark phrases appeared verbatim in generated reviews. Time is wall-clock seconds per review (mean $\pm$ std); cost is USD per review (mean $\pm$ std).

Model	**PDF Upload**	**PDF to Text**	**Main-Body Extraction**	**Screenshot + OCR**
Success Rate (%)				
GPT-5.5	0	60	0	0
Claude Sonnet 4.5	100	100	0	0
Gemini 3 Flash	100	100	0	0
Time (seconds)				
GPT-5.5	28.4 ± 3.3	25.0 ± 2.1	26.0 ± 2.7	33.1 ± 4.8
Claude Sonnet 4.5	49.0 ± 20.1	28.0 ± 1.8	28.2 ± 1.6	28.7 ± 1.9
Gemini 3 Flash	11.4 ± 2.0	7.8 ± 1.2	8.7 ± 1.1	13.6 ± 0.9
Cost (USD)				
GPT-5.5	0.206 ± 0.093	0.159 ± 0.059	0.144 ± 0.050	0.053 ± 0.002
Claude Sonnet 4.5	0.209 ± 0.103	0.102 ± 0.041	0.093 ± 0.034	0.028 ± 0.001
Gemini 3 Flash	0.008 ± 0.003	0.017 ± 0.007	0.014 ± 0.005	0.005 ± 0.000

To assess detection robustness, we conducted a controlled experiment replicating the ICML watermarking design [Rao et al., 2025]. We prepared 20 papers with invisible footer prompts and tested detection over four plausible reviewing workflows: direct PDF upload, full-text extraction, main-body extraction excluding footer regions, and screenshot-based input with OCR. We used three frontier models (GPT-5.5 [OpenAI, 2026], Claude Sonnet 4.5 [Anthropic, 2025], Gemini 3 Flash [Google DeepMind, 2025]) and measured trigger success, *i.e.,* whether both target phrases appeared verbatim in generated reviews. Table 1 reports success rates alongside timing and cost metrics.

The results reveal fragility across two dimensions. First, detection is model-dependent: GPT-5.5 failed to trigger even with PDF upload, while Claude Sonnet 4.5 and Gemini 3 Flash succeeded. This heterogeneity implies that a reviewer switching models can evade detection without altering the workflow. Second, minor input modifications eliminate detection: main-body extraction and screenshot-based input produced 0% detection success across all tested models, despite requiring comparable time and even lower cost. These evasion strategies are particularly practical because they impose negligible overhead to low-effort reviewing; removing footer content or submitting screenshots can effectively bypass detection with minimal inconvenience.

These findings demonstrate that footer-based watermarking, while effective under narrow conditions (direct PDF ingestion with compliant models), fails to provide robust coverage of realistic reviewing workflows. Together with the incentive structure limitations identified in Section 2.2, this evidence underscores the inadequacy of deterrence-centered governance for ensuring review integrity.

2.4 The Failure Mode: Enforcement Leaves Intact

We now abstract from the preceding incentive and technical limitations to identify the behavioral failure mode that viable intervention should address.

Threat Model. The preceding analysis points to a more general threat model than any particular enforcement mechanism: under time pressure, reviewers can transform a paper into machine-consumable input, delegate core evaluative judgments to an LLM, and lightly post-edit the result into a plausible review. We refer to this failure mode as **low-effort rational reviewing**.

Cost–Reliability Regimes. Figure 1 illustrates three review regimes under the proposed cost–reliability framework. Regime I (Human Review) occupies the high-C, high-R region: a careful review often requires several hours or more and provides judgments grounded in expert reading and verification. Regime II (Verified LLM Assistance) may reduce drafting and organization costs while remaining a moderate-C, high-R regime because the reviewer must still read the paper, check claims, and validate generated judgments. Regime III (Low-Effort LLM Reviewing) has low C and low actual R, while preserving sufficient superficial quality to appear acceptable. This regime can produce coherent, well-structured reviews with less effort because it bypasses the reading and cross-checking that dominate review cost.

Behavioral Boundary. The relevant boundary is therefore behavioral rather than tool-based. A review is not problematic simply because an LLM was involved; it becomes problematic when model output *replaces*, rather than supports, human judgment. Replacement occurs when core review judgments—such as strengths, weaknesses, novelty assessments, and recommendations—are derived from LLM outputs without independent validation. Support, in contrast, preserves epistemic ownership: the reviewer may use an LLM to summarize, edit, or organize, but the final evaluation remains grounded in reading and verification by human.

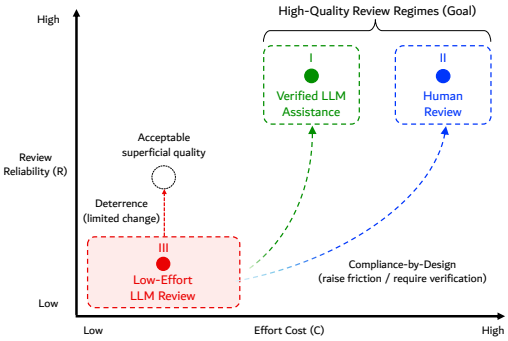

Figure 1: Three review regimes under the cost–reliability framework. Regime I (Human Review) and Regime II (Verified LLM Assistance) represent high-quality review regimes, while Regime III (Low-Effort LLM Review) relies on unverified judgment replacement. Compliance-by-design mechanisms aim to shift reviewing away from Regime III toward the high-quality regimes.

Mechanism-Level Implication. Low-effort rational reviewing is not tied to a single technical pathway. Reviewers may upload the full PDF, extract plain text, copy selected sections, use screenshots with OCR, or combine partial inputs across multiple prompts. These workflows differ in surface form, but share the same structure: the paper is converted into LLM input, and the resulting output becomes the basis of the review. This is why misuse detection and enforcement struggle to eliminate the shortcut. They may identify some artifacts after the fact, but they do not change the basic cost–reliability profile of the unverified workflow: the reviewer still obtains a plausible review cheaply, while enforcement remains uncertain and avoidable.

The goal of intervention is therefore not to eliminate traditional peer reviewing or verified forms of LLM assistance, but rather to reduce viability of unverified substitution of automatic LLM pipelines. This shifts attention from post hoc policing to the document-to-model interface where low-effort reviewing occurs. Effective mechanisms should ask not only whether misuse can be detected, but whether unverified LLM-mediated workflows remain cheap and plausible enough to be attractive. Concretely, they must either **increase the effort required for unverified workflows** (raising C) or **decrease the reliability of outputs produced without verification** (reducing R), while preserving the usability and integrity of legitimate, verification-based reviewing.

This threat model motivates the mechanism framework and design principles that follow. Because low-effort reviewing spans models, input channels, and workflows, viable mechanisms must be distributed, multi-channel, model-agnostic, and resistant to low-cost removal.

3 Position: Safeguard Review Integrity through Compliance-by-Design

Section 2 reframes the problem from detecting individual violations to the workflow enabling unverified LLM substitution. This leads to a design-oriented position: review integrity should be built into reviewing conditions, rather than enforced only after review submission.

We argue for a shift to **compliance-by-design**: review infrastructures should be designed so that verification-based reviewing is the path of least resistance, while unverified LLM replacement becomes costly, unreliable, or auditable during the workflow itself. This approach does not rely solely on reviewers' voluntary compliance or on ex-post punishment. Instead, it changes the practical conditions under which reviewing strategies are chosen.

Position

Peer-review integrity requires compliance-by-design mechanisms that make unverified LLM replacement lose its low-cost, high-plausibility advantage.

Under the cost–reliability model, any strategy that is cheaper than full evaluation, produces acceptable-looking outputs, and faces only uncertain enforcement can become stable even when its reliability is poor. Compliance-by-design aims to remove this shortcut by making unverified replacement less attractive than verification-based reviewing.

Operational Objective. The target is the third regime identified in Section 2: low-C reviewing with low actual reliability but sufficient perceived reliability. A viable intervention should move this regime out of the attractive region by raising C, lowering actual reliability or perceived plausibility, or adding auditability that becomes meaningful when combined with one of these effects.

The intervention point is the document-to-model interface. Low-effort reviewing occurs when a submission is converted into LLM-consumable input and model outputs become the basis for review judgments. Compliance-by-design mechanisms therefore operate on the conditions of machine consumption: how content is extracted, transformed, interpreted, and propagated into generated reviews. The objective is not to block access to papers, but to ensure that unverified automation is no longer significantly cheaper than human review.

Mechanism Roles. Compliance-by-design mechanisms can act on the low-effort regime in three complementary ways. First, **extraction friction raises** C by making clean unverified LLM ingestion harder without preventing legitimate access to the paper. Such mechanisms target copying, parsing, reformatting, OCR, and other transformations that construct model-ready inputs without verification.

Second, **review-integrity probes lower actual reliability or perceived plausibility** by introducing document-level signals that are inert under human reading but can affect LLM-mediated workflows. When model outputs substitute for independent judgment, these probes can introduce inconsistencies, distortions, or visible artifacts that make unverified outputs less reliable or less acceptable.

Third, **audit layers support accountability** by providing evidence that a review was derived from instrumented content. On their own, such mechanisms remain deterrence-centered; within compliance-by-design, they are most useful when combined with cost increase or reliability degradation.

Design Constraints. These mechanisms should satisfy three categories of constraints. First, they should **preserve human-grounded reviewing**: the paper's meaning, readability, accessibility, and legitimate verification-based LLM assistance should remain intact. Second, they should **cover adaptive machine consumption** through multi-channel robustness, distributed embedding across document regions and representations, and model-agnostic design that does not rely on specific parser or model behaviors. Third, deployment should be **governed by venues** rather than individual authors, with due process, low false-positive risk, and transparency of existence rather than specific instantiation. Detailed design principles are provided in Appendix A.

Together, these roles and constraints define the success condition: unverified workflows should lose their low-cost advantage while legitimate review remains intact. Bypassing the mechanism should require effort closer to human reading and verification; ignoring it should make unverified LLM-generated reviews less reliable, less acceptable, or more auditable.

4 Instantiations of Compliance-by-Design Mechanisms

We now sketch concrete mechanisms that instantiate the compliance-by-design position in Section 3, with detailed principles in Appendix A. These are not a single prescribed implementation; they illustrate how document-level interventions can *raise the cost of clean unverified ingestion, lower the reliability or plausibility of unverified outputs, and support auditability*.

Distributed Prompt Injection as Review-Integrity Probes Prompt injection is usually treated as a vulnerability: external content can influence model behavior because LLMs do not reliably separate instructions from data [Willison, 2022, Liu et al., 2025]. We repurpose this observation for review integrity: if papers are consumed by models during reviewing, the document-to-model interface can carry signals that are inert for human readers but active in unverified LLM-mediated workflows.

The key difference from adversarial prompt injection is intent and governance. The goal is not to exfiltrate information, manipulate reviewers, or alter papers. Venue-controlled probes instead introduce bounded perturbations or detectable traces when model outputs substitute for human judgment. Signals embedded in text, layout, or rendering may cause model-generated reviews to reveal artifacts, follow irrelevant instructions, or exhibit inconsistencies that a verifier could catch.

The failure modes in Section 2 motivate distribution. A footer-only signal can be removed through partial extraction, missed under image-based workflows, or ignored by some models. Review-integrity probes should therefore be spread across document regions and representations, so that upload, copy-

Table 2: Distributed prompt-injection trigger rate under *main-body extraction*, on the same 20 papers and three models as Table 1.

Model	Success Rate (%)	Time (s)	Cost (USD)
GPT-5.5	95	72.0 ± 10.9	0.235 ± 0.066
Claude Sonnet 4.5	95	41.0 ± 6.8	0.110 ± 0.039
Gemini 3 Flash	95	9.2 ± 1.2	0.017 ± 0.007

Table 3: Token-bomb impact under common LLM-assisted reviewing workflows. Cells report the median input-token count over 20 papers, with percent increase over the unmodified baseline (N=0) in brackets. N is the number of zero-width-space characters appended to each word-final glyph.

Model	PDF Upload			PDF to Text		
	N=0	N=1	N=5	N=0	N=1	N=5
GPT	22 852	31 748 (+33%)	42 348 (+75%)	18 150	26 458 (+43%)	37 254 (+99%)
Claude	21 582	21 586 (+0%)	21 588 (+0%)	20 252	38 552 (+85%)	79 786 (+273%)
Gemini	8 326	8 326 (+0%)	8 326 (+0%)	20 270	29 838 (+42%)	41 471 (+94%)

paste, text extraction, screenshots, and OCR do not all erase the mechanism. Robustness comes from diversity and redundancy, not any one prompt firing reliably.

As a preliminary check, we tested a distributed injection design on the same 20 papers and three models used in Table 1. Unlike the footer-only ICML-style version, the injection was distributed across page headers and inter-paragraph gaps so main-body extraction retained some signals. Table 2 shows that this restores high trigger reliability where the footer-only baseline scored 0%.

This mechanism targets reliability and perceived plausibility. Verified reviewers can ignore or correct injected artifacts because final judgment remains grounded in reading. Unverified workflows, by contrast, may produce less reliable, less acceptable, or auditable outputs.

Extraction Friction Through Representation Design Extraction friction targets cost. The aim is to make clean, model-ready input harder to obtain without making the paper harder for humans to read. At the representation level, document structure, layout, encoding, or visual organization can be designed so that naive transformations lose information, introduce inconsistencies, or require manual cleanup [Markwood et al., 2017].

For example, layout-dependent content may not survive plain-text extraction cleanly; partial copying may break references between claims, equations, and figures; and OCR-based workflows may require additional cleaning when signals are distributed across visual structure. These effects should not block access to the paper; they raise the marginal cost of unverified automation until constructing a clean input begins to resemble reading and checking the work.

As a concrete instance, a paper's font `/ToUnicode` tables can be augmented so selected glyphs extract as a letter followed by N zero-width-space (ZWSP) characters (`U+200B`), extending prior work on glyph-level document embedding and imperceptible Unicode perturbations [Xiao et al., 2018, Boucher et al., 2022]. Reusing original glyph outlines preserves visible rendering, while restricting padding to whitespace preserves accessibility. Only workflows that rely on extracted text (*e.g.,* copy-paste or PDF-to-Text conversion) see the padding; human readers and PDF renderers do not. Applied at every word-final position, this creates a *token bomb*: a visually unchanged document whose extracted text contains a large number of hidden padding characters, dramatically increasing token counts after tokenization. Table 3 reports inflation across two ingestion paths and three models.

These results are a lower bound on what the mechanism can produce: N is a free parameter and can be raised further per-paper, per-section, or per-glyph without changing the visible document. Increased token counts consume context-window budget and computational resources that would otherwise be available for the paper itself, thereby raising the cost of large-scale automated reviewing. The on-disk cost of raising N is negligible—because the padding lives in per-font `/ToUnicode` entries rather than the page content stream, the bomb adds only a few kilobytes to a multi-megabyte PDF, even at large N. The bomb survives the copy-paste path against all three providers, with Claude roughly $3\times$ as ZWSP-sensitive as GPT or Gemini. It has no measurable effect when the PDF is

uploaded directly to Claude or Gemini, indicating that providers handle uploaded PDFs differently. Thus, a signal embedded in one representation will not necessarily reach every model's tokenizer; representation-level friction should be paired with signals in other channels.

At the same time, extraction friction must satisfy the non-interference constraints in Appendix A: it should preserve meaning, readability, accessibility, and archival quality. The target is not human use of the document, but low-effort conversion into machine-consumable input.

Watermarking as an Audit Layer Watermarking provides a complementary audit mechanism. Prior work studies detectable statistical or lexical patterns in generated text [He et al., 2022a,b, Kirchenbauer et al., 2023]. In our setting, the submission itself can contain distributed provenance signals that persist across common forms of machine ingestion and re-generation.

Venue-controlled transformations may introduce subtle but trackable patterns across wording, formatting, citation structure, or visual layout. These signals need not affect readability, but fragments may reappear when a paper is summarized or rewritten through LLM pipelines. A review that reproduces unusual structural or lexical traces can provide evidence that it may have been derived from instrumented content.

Unlike review-integrity probes, watermarking does not directly perturb model behavior. Its role is evidentiary: supporting investigation and accountability for unverified workflows. However, prior work shows that watermark signals can degrade under paraphrasing, transformation, or adversarial modification [Liu et al., 2024]. More importantly, auditability alone does not change the cost–reliability profile of low-effort reviewing. Watermarking is therefore useful, but insufficient unless paired with mechanisms that raise effort cost or reduce output reliability.

Platform-Side Coordination as Amplification The mechanisms above can operate at the document level, but review platforms and model providers could amplify their effects when document-level signals are encountered during LLM-mediated workflows. Platforms may warn users, request confirmation, recommend verification, or preserve provenance metadata. Even lightweight warnings can add interaction steps, increase cognitive load, and interrupt fully automated reviewing.

Platform coordination primarily raises effort cost. If a reviewer follows a verification prompt, the workflow moves closer to human-grounded review. If the reviewer ignores it, the output remains subject to document-level probes, extraction friction, or audit evidence. We treat platform coordination as optional amplification: useful when available, but not necessary for the document-level framework.

Composition and Evaluation These mechanisms are intended to compose. Distributed probes can reduce reliability or reveal unverified model use; extraction friction can raise the cost of clean ingestion; watermarking can support auditability; and platform coordination can prompt verification. Together, they target the regime in which unverified LLM-assisted reviewing remains both cheap and plausible.

The key evaluation question is not whether any single mechanism is perfectly robust, but whether adaptation preserves the low-effort advantage. If bypassing distributed signals requires model switching, signal removal, representation comparison, manual cleanup, and output verification, the reviewer has incurred costs approaching the process the shortcut was meant to avoid. If skipping those steps yields compromised or less reliable reviews, the shortcut also loses value.

Our preliminary experiments follow this logic: they measure trigger reliability and token-cost inflation across models and ingestion pathways. Future evaluations should extend this measurement to human effort, output degradation, and adaptation cost.

5 Alternative Views and Scope

Compliance-by-design is not a replacement for existing approaches to peer-review reform. Incentives, policy enforcement, reviewer norms, and model-side safeguards all address real parts of the problem. Our claim is narrower: these approaches are incomplete when the reviewing workflow still admits a low-effort path for unverified LLM replacement. We therefore clarify how our position complements, rather than displaces, alternative views.

Why Incentives and Enforcement Are Not Enough. A prominent view treats the peer-review crisis as an incentive problem and proposes reviewer rewards, reputation systems, or bi-directional feedback between authors and reviewers [Kim et al., 2025]. Detection and policy enforcement offer another route, using hidden instructions, artifact-level signals, automated classifiers, or manual audits to identify possible misuse [Kamath, 2026]. These tools are valuable: review quality depends on motivation, accountability, and credible governance.

Their limitation is that they do not necessarily change the workflow that makes low-effort reviewing attractive. LLM-assisted workflows weaken the traditional link between effort and plausible output: a reviewer can produce an acceptable-looking review without investing the time required for careful evaluation. If evasion is cheap and punishment uncertain, incentives and enforcement may improve the broader review culture while leaving the low-effort shortcut intact. Compliance-by-design targets this different layer: the practical conditions that make unverified substitution cheap.

Why LLM Bans Fall Short. Another response is to sharply restrict LLM use in reviewing. Such policies may be appropriate in venues with serious confidentiality or fairness concerns. They also send a clear normative signal that reviewers remain responsible for their judgments.

The difficulty is enforcement and scope. LLM tools are widely available, and reviewers can use many input channels, including PDFs, extracted text, screenshots, and partial prompts. Strict bans may also suppress legitimate uses such as editing, summarization, or accessibility support. Our focus is therefore behavioral rather than tool-based: the target is unverified reliance on model outputs, not LLM assistance itself.

Why Model-Side Safeguards Fall Short. Model providers may improve review integrity through calibration, alignment, provenance tools, or safeguards against following irrelevant instructions. These advances could help, and platform coordination could amplify document-level mechanisms.

Nevertheless, peer-review systems cannot assume uniform model behavior. Reviewers may use different models, local tools, OCR pipelines, or intermediate transformations. A venue-level mechanism should therefore remain robust even when model-side safeguards are absent or inconsistent. Compliance-by-design treats model variability as part of the threat model rather than an external problem that must be solved first.

How Compliance-by-Design Avoids Overreach. A natural concern is that document instrumentation could itself become harmful: authors might try to manipulate reviewers, venues might over-penalize ambiguous signals, or mechanisms might burden accessibility and legitimate review. This concern is serious and motivates the governance constraints in Appendix A.

Compliance-by-design mechanisms should be venue-controlled, standardized, and auditable; they should preserve the paper's meaning and readability; and detection signals should support investigation rather than automatic punishment. The purpose is not to trick reviewers or privilege authors, but to make unverified automation less attractive while keeping human-grounded review intact.

The implication is not that conferences should abandon incentives, policies, or model-side safeguards. It is that these tools should be paired with workflow-level mechanisms that make unverified substitution harder to sustain. Compliance-by-design addresses this distinct layer: the document-to-model workflow that allows unverified LLM outputs to substitute for human judgment at low cost.

6 Conclusion

Peer review faces a new integrity challenge in the era of LLMs: unverified model outputs can substitute for human judgment while still producing coherent, acceptable-looking reviews. We call this failure mode *low-effort rational reviewing*: under time pressure, reviewers may rationally choose workflows that are cheap, plausible, and weakly constrained by enforcement.

Our position is that peer-review integrity cannot rest on deterrence alone. Detection and punishment support accountability, but they do not remove the low-effort shortcut. Safeguarding review quality requires *compliance-by-design*: review workflows should make verification-based reviewing the default path, while unverified LLM replacements lose their low-cost, high-plausibility advantage.

This shift implies a design agenda at the document-to-model interface. Mechanisms should raise the cost of clean unverified ingestion, reduce the reliability or perceived plausibility of unverified outputs,

and support auditability while preserving readability, accessibility, due process, and legitimate LLM assistance. Distributed probes, watermarking, extraction friction, and platform-side coordination illustrate this agenda, but their value lies in composition.

The goal is not to remove LLMs from peer review. It is to keep LLM assistance organized around human verification rather than replacing it. As LLMs become part of the scholarly infrastructure, compliance-by-design offers a complementary path forward: conferences should move beyond merely stating expectations for responsible use and begin building review workflows that make human-grounded verification the rational default.

References

Anthropic. Claude sonnet 4.5 system card, 2025. URL https://www-cdn.anthropic.com/963373e433e489a87a10c823c52a0a013e9172dd.pdf.

Joachim Baumann, Jiaxin Pei, Sanmi Koyejo, and Dirk Hovy. Stop automating peer review without rigorous evaluation. In *Forty-third International Conference on Machine Learning*, 2026. Spotlight.

Nicholas Boucher, Ilia Shumailov, Ross Anderson, and Nicolas Papernot. Bad characters: Imperceptible nlp attacks. In *2022 IEEE Symposium on Security and Privacy (SP)*, pages 1987–2004, 2022. doi: 10.1109/SP46214.2022.9833641. URL https://doi.org/10.1109/SP46214.2022.9833641.

Ann Cavoukian. Privacy by design: The 7 foundational principles. Technical Report 2009, Information and Privacy Commissioner of Ontario, Canada, 2009.

Roberto Centeno, Ramón Hermoso, and Maria Fasli. On the inaccuracy of numerical ratings: dealing with biased opinions in social networks. *Information Systems Frontiers*, 17(4):809–825, August 2015. ISSN 1572-9419. doi: 10.1007/s10796-014-9526-1. URL https://doi.org/10.1007/s10796-014-9526-1.

Ahmed M. Elkhatat, Khaled Elsaid, and Saeed Almeer. Evaluating the efficacy of AI content detection tools in differentiating between human and AI-generated text. *International Journal for Educational Integrity*, 19(1):17, September 2023. ISSN 1833-2595. doi: 10.1007/s40979-023-00140-5. URL https://doi.org/10.1007/s40979-023-00140-5.

Yang Gao, Steffen Eger, Ilia Kuznetsov, Iryna Gurevych, and Yusuke Miyao. Does my rebuttal matter? insights from a major NLP conference. In Jill Burstein, Christy Doran, and Thamar Solorio, editors, *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 1274–1290, Minneapolis, Minnesota, June 2019. Association for Computational Linguistics. doi: 10.18653/v1/N19-1129. URL https://aclanthology.org/N19-1129/.

Google DeepMind. Gemini 3 flash model card, 2025. URL https://storage.googleapis.com/deepmind-media/Model-Cards/Gemini-3-Flash-Model-Card.pdf.

Xuanli He, Qiongkai Xu, Lingjuan Lyu, Fangzhao Wu, and Chenguang Wang. Protecting intellectual property of language generation apis with lexical watermark. *Proceedings of the AAAI Conference on Artificial Intelligence*, 36(10):10758–10766, Jun. 2022a. doi: 10.1609/aaai.v36i10.21321. URL https://ojs.aaai.org/index.php/AAAI/article/view/21321.

Xuanli He, Qiongkai Xu, Yi Zeng, Lingjuan Lyu, Fangzhao Wu, Jiwei Li, and Ruoxi Jia. Cater: Intellectual property protection on text generation apis via conditional watermarks. In S. Koyejo, S. Mohamed, A. Agarwal, D. Belgrave, K. Cho, and A. Oh, editors, *Advances in Neural Information Processing Systems*, volume 35, pages 5431–5445. Curran Associates, Inc., 2022b. URL https://proceedings.neurips.cc/paper_files/paper/2022/file/2433fec2144ccf5fea1c9c5ebdbc3924-Paper-Conference.pdf.

Junjie Huang, Win bin Huang, Yi Bu, Qi Cao, Huawei Shen, and Xueqi Cheng. What makes a successful rebuttal in computer science conferences?: A perspective on social interaction. *Journal of Informetrics*, 17(3):101427, 2023. ISSN 1751-1577. doi: https://doi.org/10.1016/j.joi.2023.101427. URL https://www.sciencedirect.com/science/article/pii/S1751157723000524.

Gautam Kamath. On violations of llm review policies. https://blog.icml.cc/2026/03/18/on-violations-of-llm-review-policies/, 2026.

Jaeho Kim, Yunseok Lee, and Seulki Lee. Position: The AI conference peer review crisis demands author feedback and reviewer rewards. In *Forty-second International Conference on Machine Learning Position Paper Track*, 2025. URL https://openreview.net/forum?id=l8QemUZaIA.

John Kirchenbauer, Jonas Geiping, Yuxin Wen, Jonathan Katz, Ian Miers, and Tom Goldstein. A watermark for large language models. In Andreas Krause, Emma Brunskill, Kyunghyun Cho, Barbara Engelhardt, Sivan Sabato, and Jonathan Scarlett, editors, *Proceedings of the 40th International Conference on Machine Learning*, volume 202 of *Proceedings of Machine Learning Research*, pages 17061–17084. PMLR, 23–29 Jul 2023. URL https://proceedings.mlr.press/v202/kirchenbauer23a.html.

Lawrence Lessig. *Code and Other Laws of Cyberspace*. Basic Books, Inc., USA, 1999. ISBN 046503912X.

Weixin Liang, Zachary Izzo, Yaohui Zhang, Haley Lepp, Hancheng Cao, Xuandong Zhao, Lingjiao Chen, Haotian Ye, Sheng Liu, Zhi Huang, Daniel McFarland, and James Y. Zou. Monitoring AI-modified content at scale: A case study on the impact of chatGPT on AI conference peer reviews. In *Forty-first International Conference on Machine Learning*, 2024a. URL https://openreview.net/forum?id=bX3J7ho18S.

Weixin Liang, Yaohui Zhang, Zhengxuan Wu, Haley Lepp, Wenlong Ji, Xuandong Zhao, Hancheng Cao, Sheng Liu, Siyu He, Zhi Huang, Diyi Yang, Christopher Potts, Christopher D Manning, and James Y. Zou. Mapping the increasing use of LLMs in scientific papers. In *First Conference on Language Modeling*, 2024b. URL https://openreview.net/forum?id=YX7QnhxESU.

Aiwei Liu, Leyi Pan, Yijian Lu, Jingjing Li, Xuming Hu, Xi Zhang, Lijie Wen, Irwin King, Hui Xiong, and Philip Yu. A survey of text watermarking in the era of large language models. *ACM Comput. Surv.*, 57(2), November 2024. ISSN 0360-0300. doi: 10.1145/3691626. URL https://doi.org/10.1145/3691626.

Yi Liu, Gelei Deng, Yuekang Li, Kailong Wang, Zihao Wang, Xiaofeng Wang, Tianwei Zhang, Yepang Liu, Haoyu Wang, Yan Zheng, Leo Yu Zhang, and Yang Liu. Prompt injection attack against llm-integrated applications, 2025. URL https://arxiv.org/abs/2306.05499.

Ian Markwood, Dakun Shen, Yao Liu, and Zhuo Lu. PDF mirage: Content masking attack against Information-Based online services. In *26th USENIX Security Symposium (USENIX Security 17)*, pages 833–847, Vancouver, BC, August 2017. USENIX Association. ISBN 978-1-931971-40-9. URL https://www.usenix.org/conference/usenixsecurity17/technical-sessions/presentation/markwood.

Nestor Maslej, Loredana Fattorini, Raymond Perrault, Vanessa Parli, Anka Reuel, Erik Brynjolfsson, John Etchemendy, Katrina Ligett, Terah Lyons, James Manyika, Juan Carlos Niebles, Yoav Shoham, Russell Wald, and Jack Clark. Artificial intelligence index report 2024, 2024. URL https://arxiv.org/abs/2405.19522.

OpenAI. Gpt-5.5 system card, 2026. URL https://deploymentsafety.openai.com/gpt-5-5/introduction.

Vishisht Srihari Rao, Aounon Kumar, Himabindu Lakkaraju, and Nihar B Shah. Detecting llm-generated peer reviews. *PLoS One*, 20(9):e0331871, 2025. URL https://doi.org/10.1371/journal.pone.0331871.

Joseph S. Ross, Cary P. Gross, Mayur M. Desai, Yuling Hong, Augustus O. Grant, Stephen R. Daniels, Vladimir C. Hachinski, Raymond J. Gibbons, Timothy J. Gardner, and Harlan M. Krumholz. Effect of Blinded Peer Review on Abstract Acceptance. *JAMA*, 295(14):1675–1680, April 2006. ISSN 0098-7484. doi: 10.1001/jama.295.14.1675. URL https://doi.org/10.1001/jama.295.14.1675. _eprint: https://jamanetwork.com/journals/jama/articlepdf/202669/joc60040.pdf.

Alessio Russo. Some ethical issues in the review process of machine learning conferences, 2021. URL https://arxiv.org/abs/2106.00810.

Nihar B. Shah. Challenges, experiments, and computational solutions in peer review. *Commun. ACM*, 65(6):76–87, May 2022. ISSN 0001-0782. doi: 10.1145/3528086. URL https://doi.org/10.1145/3528086.

Paula Stephan. *How Economics Shapes Science*. Harvard University Press, 2012. ISBN 9780674049710. URL http://www.jstor.org/stable/j.ctt2jbqd1.

Richard H. Thaler and Cass R. Sunstein. *Nudge: Improving Decisions About Health, Wealth, and Happiness*. Yale University Press, New Haven, CT, 2008. ISBN 978-0-300-12223-7.

Hendrik P. van Dalen and Kène Henkens. Intended and unintended consequences of a publish-or-perish culture: A worldwide survey. *Journal of the American Society for Information Science and Technology*, 63(7):1282–1293, 2012. doi: https://doi.org/10.1002/asi.22636. URL https://onlinelibrary.wiley.com/doi/abs/10.1002/asi.22636.

Simon Willison. Prompt injection attacks against llms. https://simonwillison.net/2022/Sep/12/prompt-injection/, 2022. Blog post.

Ruihan Wu, Chuan Guo, Felix Wu, Rahul Kidambi, Laurens Van Der Maaten, and Kilian Weinberger. Making paper reviewing robust to bid manipulation attacks. In Marina Meila and Tong Zhang, editors, *Proceedings of the 38th International Conference on Machine Learning*, volume 139 of *Proceedings of Machine Learning Research*, pages 11240–11250. PMLR, 18–24 Jul 2021. URL https://proceedings.mlr.press/v139/wu21b.html.

Chang Xiao, Cheng Zhang, and Changxi Zheng. Fontcode: Embedding information in text documents using glyph perturbation. *ACM Trans. Graph.*, 37(2), February 2018. ISSN 0730-0301. doi: 10.1145/3152823. URL https://doi.org/10.1145/3152823.

A Detailed Principles for Compliance-by-Design

Compliance-by-design mechanisms must satisfy two requirements simultaneously: they must disrupt the low-effort advantage of unverified LLM-assisted reviewing while preserving the integrity of the review process itself. In practice, this means increasing the cost C of unverified workflows, reducing the reliability or plausibility of their outputs, or providing auditability that supports one of these effects, without interfering with legitimate human evaluation. We organize the resulting design constraints around three questions: what must remain invariant for human reviewers, what forms of document-to-model interaction must be covered, and how deployment should be governed.

Table 4: Summary of design principles for compliance-by-design mechanisms.

Principle	Role
Preserve human-grounded reviewing	
Human-legible invariance	Preserve meaning, readability, accessibility, and archival quality.
Support-preserving selectivity	Target replacement of judgment, not legitimate LLM assistance.
Cover adaptive machine consumption	
Multi-channel coverage	Cover PDFs, extracted text, screenshots, OCR, and hybrid workflows.
Distributed embedding	Avoid bypass through partial extraction or reformatting.
Model-agnostic robustness	Avoid reliance on one parser, model, or instruction-following behavior.
Adaptation resistance	Make bypass costly enough to reduce the low-effort advantage.
Bound deployment through governance	
Venue-controlled deployment	Prevent author-side manipulation and standardize deployment.
Due process; low false positives	Use signals for investigation rather than automatic penalties.
Existence transparency	Disclose the mechanism class without revealing removable details.

A.1 Preserve Human-Grounded Reviewing

The first constraint is **human-legible invariance**: mechanisms must preserve the paper's meaning, readability, accessibility, and archival quality. A human reviewer should be able to read, interpret, annotate, and cite the submission as usual. The mechanism should be invisible to the substance of human evaluation, even if it changes how the document behaves under machine ingestion.

This constraint also requires **support-preserving selectivity**. Verified LLM assistance should remain possible when it supports reading, editing, or organization rather than replacing judgment. Mechanisms should therefore target unverified workflows in which model outputs become substitutes for independent evaluation, not LLM involvement as such. Designs that impose costs on careful readers, change the meaning of the submission, or interfere with accessibility tools fail this requirement.

A.2 Cover Adaptive Machine Consumption

The second constraint is **multi-channel coverage**. As Section 2 shows, low-effort reviewing is not tied to a single ingestion pathway. Reviewers may upload PDFs, extract text, copy sections, use screenshots with OCR, or combine partial inputs across prompts and models. A mechanism that depends on one footer signal, one parser, or one model behavior is therefore too brittle.

Coverage requires **distributed embedding** across document regions and representations. Relevant signals or frictions should appear across text, layout, and visual structure, so that partial extraction or reformatting does not trivially remove them. It also requires **model-agnostic robustness**: effectiveness should not depend on a particular model reliably following a particular instruction or preserving a particular formatting artifact.

Finally, mechanisms must provide **adaptation resistance**. This does not mean that every signal must survive every transformation. Rather, bypass should require enough coordinated effort that the shortcut loses its advantage. If neutralizing the mechanism requires inspecting the paper, comparing representations, cleaning extracted content, and validating the resulting review, the reviewer has moved toward the cost profile of verification-based review. Robustness is therefore measured not by perfect prevention, but by whether adaptation destroys the low-effort regime.

A.3 Bound Deployment Through Governance

The third constraint is **venue-controlled deployment**. Compliance-by-design should not become a license for arbitrary author-side manipulation. Integrity mechanisms should be applied by the venue or review platform under standardized, auditable procedures, rather than by individual authors who may have incentives to influence reviewers. This protects authors from uneven deployment and protects reviewers from adversarially crafted submissions.

Governance also requires **due process and low false-positive risk**. Detection signals, when present, should support investigation rather than trigger automatic penalties. Watermarking and other detection layers are useful as audit mechanisms, but they should not be treated as standalone solutions. Because model behavior and document transformations vary, false positives can occur; audit evidence should be interpreted together with review content, workflow context, and policy commitments.

Finally, deployment should provide **transparency of existence, not instantiation**. Reviewers may be told that compliance-by-design mechanisms exist, but specific implementations should vary across papers and rounds to prevent trivial removal. Review platforms or LLM providers could amplify this approach through warnings, standardized handling of instrumented documents, or provenance support, but such coordination should remain an enhancement rather than a required assumption.