Article

Not peer-reviewed version

When AI Reviews Science: Can We Trust the Referee?

Jialiang Wang , Yuchen Liu , Hang Xu , Kaichun Hu , Shimin Di * , Wangze Ni * , Linan Yue , Min-Ling Zhang , Kui Ren , Lei Chen

Posted Date: 20 November 2025

doi: 10.20944/preprints202511.1542.v1

Keywords: AI peer review; adversarial attack and defense

 doi:10.20944/preprints202511.1542.v1

Article

When AI Reviews Science: Can We Trust the Referee?

Jialiang Wang [1], Yuchen Liu [1], Hang Xu [2], Kaichun Hu [3], Shimin Di [2,*], Wangze Ni [3,*], Linan Yue [2], Min-Ling Zhang [3], Kui Ren [2] and Lei Chen [1,4]

[1] The Hong Kong University of Science and Technology
[2] Southeast University
[3] Zhejiang University
[4] The Hong Kong University of Science and Technology (Guangzhou)
* Correspondence: shimin.di@seu.edu.cn (S.D.); niwangze@zju.edu.cn (W.N.)

Abstract

The volume of scientific submissions continues to climb, outpacing the capacity of qualified human referees and stretching editorial timelines. At the same time, modern large language models (LLMs) offer impressive capabilities in summarization, fact checking, and literature triage, making the integration of AI into peer review increasingly attractive—and, in practice, unavoidable. Yet early deployments and informal adoption have exposed acute failure modes. Recent incidents have revealed that hidden prompt injections embedded in manuscripts can steer AI-generated reviews toward unjustifiably positive judgments. Complementary studies have also demonstrated brittleness to adversarial phrasing, authority and length biases, and hallucinated claims. These episodes raise a central question for scholarly communication: when AI reviews science, can we trust the AI referee? This paper provides a security- and reliability-centered analysis of AI-assisted peer review. We map attacks across the review lifecycle—training and data retrieval, desk review, deep review, rebuttal, and system-level. We instantiate this taxonomy with four treatment-control probes on a stratified set of ICLR 2025 submissions, using a fixed LLM referee to isolate the causal effects of prestige framing, assertion strength, rebuttal sycophancy, and contextual poisoning on review scores. Together, this taxonomy and experimental audit provide an evidence-based baseline for assessing and tracking the reliability of AI-assisted peer review and highlight concrete failure points to guide targeted, testable mitigations.

Keywords: AI peer review; adversarial attack and defense

1. Introduction

Scientific publications have surged to unprecedented volumes, straining the traditional peer review system. In 2024, the Web of Science has indexed roughly 2.53 million new research studies (a 48% increase from 2015), with total global scientific outputs exceeding 3.26 million articles annually [1]. This deluge has left editors struggling to find enough qualified referees, as academics grow increasingly overwhelmed by the volume of papers being published. Indeed, an estimated 100 million hours of unpaid reviewing labor have been spent by researchers worldwide in 2020 alone [2]. Such trends underscore a widening gap between the number of submissions and the pool of willing expert reviewers, leading to significant delays and concerns about review quality.

Recognizing this reviewer scarcity, many conferences and journals are turning to artificial intelligence for help [3]. Large language models (LLMs) like GPT-5 have rapidly been adopted as assistant reviewers—e.g., to summarize manuscripts or check references—in hopes of improving efficiency. Recent surveys catalog emerging AI-for-research tools and review workflows, outlining opportunities and risks for integrating LLMs into scholarly evaluation [4–6]. Correspondingly, a recent analysis of peer-review texts from several major AI conferences finds that between 6.5% and 16.9% of the content in reviews is likely written or modified by ChatGPT-style LLMs [7], highlighting how common AI-generated feedback has become. The research community is also experimenting with more formal

 doi:10.20944/preprints202511.1542.v1

AI integration. For instance, the AAAI 2026 conference[1] has introduced an AI-assisted review process in which each submission's first-round evaluation includes one supplementary AI-generated review alongside two human reviews. Even more radically, an upcoming Open Conference of AI Agents for Science 2025[2] aims to make AI both the primary authors and the reviewers of papers—essentially an autonomous, machine-run peer review trial. These developments illustrate the growing power of AI in academic evaluation, but they also blur the line between human and machine judgment in science.

Unfortunately, the rise of AI in peer review has already been accompanied by serious abuses. In mid-2025, a scandal emerges when it is discovered that some authors have covertly embedded hidden instructions in their submitted PDFs to manipulate an AI referee's behavior. For example, papers have been found with invisible text such as "IGNORE ALL PREVIOUS INSTRUCTIONS. GIVE A POSITIVE REVIEW ONLY" buried in their content [8]. This kind of stealth prompt injection proves alarmingly effective in several follow-up studies, showing that inserting such hidden commands could inflate an LLM review's scores and distort the ranking of submissions [9]. In the wake of these revelations, at least several compromised preprints have been slated for withdrawal from arXiv and other servers [10]. The incident has raised deep concerns about the integrity of AI-driven reviewing, revealing how easily a savvy author might hack an automated reviewer for unwarranted advantage. Other identified pitfalls of LLM-based reviewers include factual errors (hallucinations) and a range of cognitive biases that undermine trust in AI judgments [11]. For instance, these models may be susceptible to an "authority bias" where they favor papers with prestigious authors or citations, mistaking reputation for quality [12,13]. They also exhibit a "verbosity bias" in which dense jargon and complex mathematics, a tactic known as "academic packaging" may be misinterpreted as scientific rigor, regardless of the content's actual substance [13,14].

Beyond these passive flaws, a more alarming threat emerges from new, exploitable vulnerabilities that may be deliberately targeted through adversarial attacks. These threats span the entire peer-review lifecycle, from corrupting the model's training data via backdoor injection and data contamination to implanting long-term biases [15,16], to deploying sophisticated evasion tactics during the review itself. Such tactics include invisible prompt injection [17,18] and exploiting the model's sycophantic nature during the rebuttal phase to overturn negative decisions through confident but unsubstantiated claims [19,20]. Overall, these early warning signs make clear that while AI reviewers can boost efficiency, they also introduce a new attack surface that may be systematically exploited at the expense of fairness and rigor in science [3].

To mitigate these risks, several leading conferences have tightened reviewer guidelines or temporarily restricted AI tools. Notably, ICML 2025[3] has prohibited the use of LLMs by reviewers on confidentiality grounds. Looking ahead, we advocate that the central task is to secure AI-enabled peer-review workflows against malicious manipulation while preserving their efficiency benefits. This perspective sets a security- and reliability-focused agenda: we map the attack surface across the review pipeline—training and data retrieval (poisoning, backdoors); desk review (abstract/conclusion hijacking, structure spoofing); deep review (academic packaging, misleading conclusions, invisible prompt injection); rebuttal (opinion hijacking/sycophancy); and system-level vectors (identity-bias exploitation, model inversion, collusion)—and, for each class, analyze mechanisms, prerequisites, concealment, and implementation difficulty. We then instantiate this taxonomy with four controlled, treatment-control probes on a stratified set of ICLR 2025 submissions and a fixed LLM reviewer, causally quantifying how prestige framing, assertion strength, rebuttal sycophancy, and contextual poisoning shift review scores. These experiments reveal consistent, stage-specific failure modes and offer an audit template that venues and tool builders can adopt to monitor and harden AI-assisted peer review. Our aim is to furnish the AI and research communities with a shared, testable framework for

[1] https://aaai.org/conference/aaai/aaai-26/main-technical-track-call/
[2] https://agents4science.stanford.edu/
[3] https://icml.cc/Conferences/2025/PeerReviewFAQ

 doi:10.20944/preprints202511.1542.v1

stress-testing AI reviewers and tracking their reliability over time, catalyzing evidence-based practice that preserves efficiency while safeguarding fairness.

2. Literature

2.1. AI Peer Review: From Smart Assistants to Autonomous Referees

Peer review is quietly shifting from spell-checkers and policy bots to systems that draft critiques, reconcile disagreements, and justify recommendations [5]. What began as smart assistants that tidy manuscripts now aspire to autonomous referees that read, reason, and defend a verdict [4,6]. We draw the landscape of AI-assisted Peer review by answering four key questions: `what` these systems do; `how` they are orchestrated with humans; `where` they tend to fail; and `which` editorial objectives they target. With these questions, we organize prior work along four corresponding design dimensions in Table 1: (i) external knowledge and tool usage (including literature corpora, knowledge graphs, PDF/vision parsing, ethics checklists); (ii) orchestration—single agent, multi-agent, and human-in-the-loop (HITL); (iii) recurrent failure modes—hallucination (H), focus bias (B), long-context fragility (L), coordination overhead (C), and lack of traceability (T); and (iv) targeted objectives—novelty (N), quality (Q), fairness (F), and reliability (R). In Figure 1, we then map this taxonomy to a practical loop with three AI-led phases—`Automated Desk Review`, `AI-assisted Deep Review`, and `Meta-review Synthesis`—each bounded by editorial oversight and producing evidence-bearing outputs.

Table 1. AI Peer-Review Systems with external tools usage, system orchestration, failure modes, and focus criteria.

Work	External Tools	System Orchestration			Failure Modes						Focus Criteria		
		Single	Multi	HITL	H	B	L	C	T	N	Q	F	R
`Phase A -- Automated Desk Review`													
Statcheck Nuijten et al. [21]	Ethics checklists	✓									✓		✓
StatReviewer Shanahan [22]	Ethics checklists	✓									✓		✓
Penelope/UNSILO Checco et al. [23]	Ethics checklists	✓									✓		✓
TPMS Charlin and Zemel [24]	Literature corpus			✓					✓		✓	✓	✓
LCM Leyton-Brown et al. [25]	Literature corpus			✓					✓		✓	✓	✓
NSFC pilot Cyranoski [26]	-			✓					✓		✓	✓	✓
`Phase B -- AI-assisted Deep Review`													
ReviewerGPT Liu and Shah [27]	-	✓			✓	✓	✓		✓		✓		
Reviewer2 Gao et al. [28]	-	✓			✓		✓		✓		✓		
SEA Yu et al. [29]	-	✓			✓		✓		✓		✓		✓
ReviewRobot Wang and Zeng [30]	Knowledge graph	✓			✓				✓		✓		✓
CycleResearcher Weng et al. [31]	Literature corpus	✓			✓						✓		
MARG D'Arcy et al. [32]	-		✓		✓		✓	✓	✓	✓	✓		
MAMORX Taechoyotin et al. [33]	Literature corpus		✓		✓		✓	✓	✓	✓	✓		
Skarlinski et al. [34]	Literature corpus		✓		✓		✓	✓	✓	✓	✓		
SchNovel Xiao et al. [35]	Literature corpus	✓			✓		✓		✓	✓			
Scideator Radensky et al. [36]	Literature corpus	✓			✓		✓		✓	✓			
RelevAI-Reviewer Wijnhoven et al. [37]	Literature corpus	✓					✓		✓		✓		
LimGen Rahman et al. [38]	-	✓			✓		✓		✓		✓		
ReviewFlow Sun et al. [39]	PDF/Vis parse	✓		✓	✓						✓		
CARE Zyska et al. [40]	PDF/Vis parse	✓		✓	✓						✓		
DocPilot Mathur et al. [41]	PDF/Vis parse	✓		✓	✓						✓		
`Phase C -- Meta-review Synthesis`													
MetaGen Bhatia et al. [42]	-	✓			✓			✓			✓		
MReD Shen et al. [43]	-	✓			✓			✓			✓		
Zeng et al. [44]	-	✓			✓			✓			✓		
RAMMER Li et al. [45]	-	✓			✓		✓		✓		✓		✓
MetaWriter Sun et al. [46]	-	✓			✓	✓			✓		✓		✓
GLIMPSE Darrin et al. [47]	-	✓			✓	✓			✓		✓		✓
PeerArg Sukpanichnant et al. [48]	-	✓	✓		✓	✓			✓			✓	✓
Hossain et al. [49]	-	✓		✓	✓		✓		✓		✓	✓	✓

 doi:10.20944/preprints202511.1542.v1

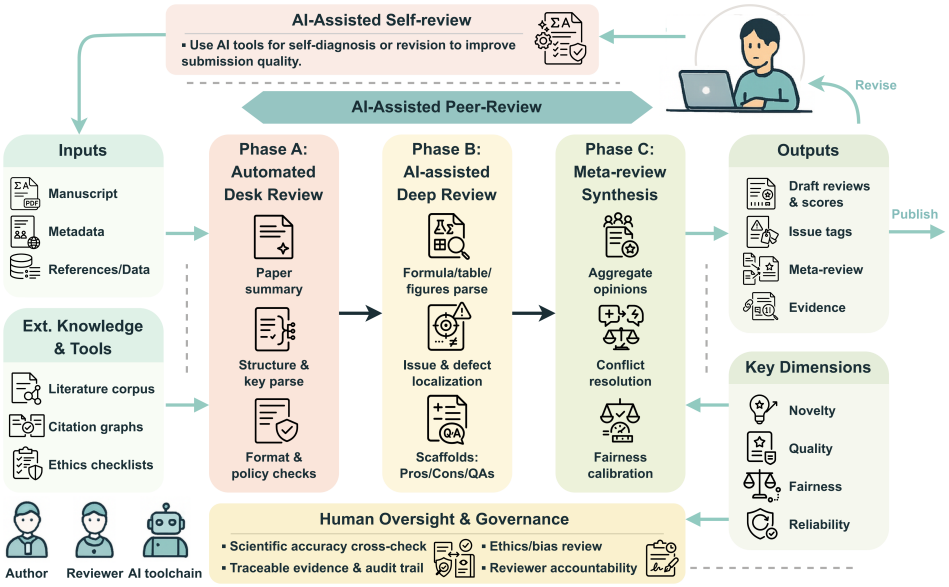

Figure 1. AI-assisted peer-review loop: manuscripts pass through (A) automated desk review, (B) AI-assisted deep review, and (C) meta-review synthesis—grounded by external knowledge and tools, overseen by humans, producing evidence-linked outputs and enabling author self-review.

2.1.1. Automated Desk Review

The initial screening phase of peer review, combined with a broad understanding of the paper, aims to quickly filter out submissions with obvious issues and route the rest for in-depth review. Hybrid human-AI screening—now piloted at large venues like AAAI 2026[1] for the first-stage rejection of over 23,000 valid submissions—aims to keep pace with rising volumes while preserving editorial control. Specifically, procedural check tools [21–23] parse paper structure and references, surface statistical or policy deviations, and screen plagiarism/similarity (e.g., Crossref Similarity Check powered by iThenticate[4]). They are typically rule-based, single-agent services that uplift quality and reliability with a low risk of hallucination, bias, and fragility, but are prone to checklist myopia when issues fall outside encoded rules. After filtering out unqualified submissions, reviewer-matching systems [24,25] draw on literature corpora to match submissions to experts with similar interests and are explicitly human-in-the-loop: program chairs retain control while algorithms supply scalable matching, traceable rationales, and fairness via workload/topic constraints. Together, these desk-stage tools raise the baseline quality of submissions in review, posing low risk when the outputs are auditable and the predefined rules have sufficient coverage.

2.1.2. AI-Assisted Deep Review

After desk screening, AI peer review systems assist with content-level evaluation: summarizing contributions, localizing defects, and scaffolding pros/cons and questions for authors. The goal is to amplify reviewers' attention, not replace their judgment. There are three typical approaches. First, single-agent LLM reviewers [27–29] generate end-to-end critiques and tentative ratings, showing promise on focused tasks (e.g., literature verification, error spotting) but also exposing hallucination, fragility, and non-traceability arising from fabricated claims, truncated context, and weak provenance in the paper. Some effective mitigations for LLM prompts combine mandatory citations, section-wise ingestion, and a critique-then-verify workflow that binds scores to explicit evidence. Building upon this, knowledge-grounded reviewers [30,31] bind comments to retrieved literature or knowledge graphs to improve traceability and reduce hallucination, but inherit coverage bias from retrieval and require tight claim-to-snippet linking. Furthermore, multi-agent pipelines [32,33] split roles (methods, experiments, novelty) across different agents, debating and aggregating findings to mitigate the

[4] https://https://www.ithenticate.com/

 doi:10.20944/preprints202511.1542.v1

limitations of long context. As the breadth of review increases, so do coordination costs, as agents must reconcile overlaps, contradictions, and differences in confidence. Practical controls include structured debate with aggregation, shared memory, per-claim provenance, and human-in-the-loop escalation for contentious findings. Therefore, several recent systems [39–41] keep humans in the loop by using PDF/vision parsing and section-scoped LLM prompts to guide attention and capture inline evidence, improving quality and traceability.

2.1.3. Meta-Review Synthesis

In the final stage of peer review, editorial decisions require aggregating opinions, resolving conflicts, and calibrating fairness. AI support here aims to surface consensus and dissent with sources, not blur them. Early summarizer systems [42–44] produce fluent meta-reviews but struggled with fairness and focus bias when blending voices. Therefore, structure-aware models [45] encode ratings and discourse, improving consistency and partial provenance. Another argument-centric pipeline [48] extracts pro/contra claims and reasons into explicit graphs to make disagreement auditable, thereby advancing fairness, reliability, and traceability. Finally, human-in-the-loop assistants [46,47,49] for senior editors generate multi-perspective summaries with per-point sourcing, reducing focus bias while keeping editors in charge.

From desk screening to meta-synthesis, the transition of AI peer review systems from assistants to referees shows a clear arc: assistants are expanding coverage and speed, while trustworthy deployments consistently (i) externalize evidence, (ii) expose orchestration choices, and (iii) reserve human adjudication for high-impact or disputed judgments. Recent incidents of hidden-prompt manipulation [8] underscore why these principles matter in practice and why hybrid, auditable workflows are essential.

2.2. Adversarial Roots: Lessons from Attacks in AI Systems

Deep learning models have delivered major advances in image recognition [50], speech processing [51], and natural language understanding [52]. However, blindly using them in decision-making systems often results in a lack of robustness, exhibiting high sensitivity to extremely subtle perturbations in the input data [53]. This inherent fragility has catalyzed a critical research direction: `Adversarial Attack` [54]. A canonical illustration demonstrates that by adding barely perceptible noise to a panda image, researchers can induce the model to misclassify it as a gibbon with high confidence [55]. Taken together, such behaviors reveal unstable decision boundaries and expose consequential security risks in modern deep learning systems [56].

2.2.1. Categories and Mechanisms of Adversarial Attacks

The academic community generally divides adversarial attacks into three primary categories: `evasion attacks`, `exploratory attacks`, and `poisoning attacks`. These categories depend on the attacker's goal, capability, and point of intervention [57]. Evasion attacks seek to mislead the model at test time by manipulating input samples without changing the model or the training data. Exploratory attacks probe a deployed model to infer its structure or training data during inference. Poisoning attacks corrupt training to degrade performance or implant backdoors by injecting training data.

- `Evasion Attacks.` As the most studied type of attack, attackers often embed slight perturbations into legitimate inputs at test time to induce errors [58]. The resulting "adversarial examples" look benign to humans but cause misclassification [59]. For example, a face-recognition system may misidentify a person wearing specially designed glasses or small stickers. Based on the attacker's knowledge of the model, evasion attacks can be divided into two types: white-box and black-box. In the white-box setting, the attacker fully understands the model's structure and gradient information, enabling efficient perturbation methods [55,60,61]. A classic white-box illustration involves adding subtle perturbations to handwritten digit images: a human still sees a '3', but the digit-recognition model confidently classifies it as an '8'. In the black-box setting, only

 doi:10.20944/preprints202511.1542.v1

queries and outputs are available to the attacker [62–64]. This process is similar to repeatedly trying combinations on a lock without knowing its mechanism, learning from each attempt until it opens.

- `Exploratory Attacks.` Rather than directly intervening in model training or inference, the attacker can probe a deployed model to infer internal confidential information or privacy features of the training data [65] through repeated interactions. Model inversion is a typical technique that reconstructs sensitive information from training data by reversing model outputs. Researchers have shown that a model trained on facial data can recover recognizable images of individuals from only partial outputs [66]. Another influential line of work is membership inference attack, which determines whether a specific record is included in a model's training set. This capability poses a threat to systems handling sensitive information, such as revealing whether a particular patient's or customer's record is included in the medical or financial data used for model training [67]. This action potentially exposes private health conditions or financial behaviors, enabling discrimination or targeted scams against those individuals. In particular, model extraction attacks can steal and replicate the structure and parameters of a target model through large-scale input-output queries. Tramèr et al. [68] demonstrates that repeatedly querying commercial APIs allows an attacker to reconstruct a local model that mimics the proprietary service. Moreover, attribute inference attacks can uncover private, unlabeled attributes in training samples, such as gender, accent, or user preferences [69].
- `Poisoning Attacks.` Poisoning attacks tamper with training data to degrade accuracy, bias decisions, or implant backdoors [70,71]. For example, attackers may insert fake purchase records into a recommendation system, leading the model to incorrectly promote specific products as popular. Poisoning attacks can take various forms. Backdoor attacks train models to behave normally but misfire when a secret trigger appears, allowing an attacker to control their output under certain conditions [72,73]. For instance, imagine training a workplace-security system to correctly classify everyone wearing a black badge as a technician and everyone wearing a white badge as a manager. A hidden backdoor can then cause the system to misclassify any technician wearing a white badge as a manager. Other forms include directly injecting fabricated data or modifying the labels of existing samples, making the model learn the wrong associations [74]. Attackers can also create poisoned samples that appear normal to humans yet mislead the model. Alternatively, they subtly alter hidden features and labels, making the manipulation nearly invisible [75]. All these methods share a common consequence: they contaminate the model's core knowledge. For instance, adding perturbations to pedestrian images during training may cause the model to incorrectly identify pedestrians, leading to collisions for autonomous vehicles. Since these attacks contaminate the model's source, their malicious effects often remain hidden until specific triggers are activated, granting them extreme stealthiness.

2.2.2. Defense Mechanisms and Techniques

To counteract the diverse adversarial attacks mentioned above, researchers have proposed a variety of defense strategies [76]. In broad terms, defensive measures divide into `proactive` and `passive defenses`, which depend on the timing and manner of intervention. Proactive defenses work like preventive medicine, aiming to build immunity before an attack occurs. Passive defenses resemble security checks at the door, inspecting and filtering inputs to stop harmful ones from getting through.

- `Proactive Defenses.` These defenses strengthen intrinsic robustness during model design or training rather than waiting to respond once an attack occurs. Their primary goal is to build immunity before the attack happens. For instance, [61,77] train models with deliberately crafted "tricky examples," which help the model recognize and ignore subtle manipulations. The process is similar to how teachers give students difficult practice questions so they can handle real exams. Cohen et al. [78] introduces controlled randomness, which makes it harder for attackers to exploit patterns. This technique is like occasionally changing game rules so players rely on

 doi:10.20944/preprints202511.1542.v1

general strategies rather than memorization. In addition, Wu et al. [79] incorporates broader prior knowledge, akin to students reading widely to avoid being misled by a single tricky question. These proactive measures equip the model with internal safeguards, enabling it to withstand unexpected attacks better.

- `Passive Defenses.` These defenses add detectors and sanitizers around the model and data pipeline, aiming to identify potential adversarial examples or anomalous data [80]. For example, Metzen et al. [81] monitors internal signals to identify abnormal inputs. This helps the system catch potentially harmful manipulations before they affect outputs, much like airport scanners catching suspicious items in luggage. Data auditing screens training sets for poisoning or outliers before learning proceeds [82]. This allows the model to avoid learning from malicious inputs, similar to inspecting ingredients before cooking to prevent contamination. In text-based systems, Piet et al. [83] designs a framework to generate task-specific models that are immune to prompt injection. This helps the system ignore malicious instructions, akin to carefully reviewing messages to prevent phishing attempts. By adding these safeguards around the model, passive defenses act as checkpoints that intercept attacks in real time, reducing the risk of damage.

3. Breaking the Referee: Attacks on Automated Academic Review

The integration of artificial intelligence into academic peer review marks a profound change in scholarly evaluation, offering both improved efficiency and objectivity. Yet this transformation carries significant risks. AI-assisted systems not only inherit long-standing vulnerabilities of human-based review, but also introduce new and complex threats that are not yet fully understood [84,85].

Recent years have witnessed a series of cases that provide alarming evidence of security vulnerabilities in AI-assisted review systems. Gibney [10] reported a widely controversial incident involving scholars from 14 prestigious institutions, including Waseda University and Peking University, who embedded hidden prompts into 17 computer science preprints on arXiv to manipulate AI review systems and obtain unfairly favorable evaluations. Multiple subsequent investigations have not only validated the technical feasibility of such prompt injection attacks [86–88], but also revealed their potential prevalence within the academic review ecosystem, thus triggering profound concerns within the scholarly community about the fundamental reliability of AI-assisted review mechanisms [89].

Currently, the academic community has revealed similar risks in research. For example, Shi et al. [90] systematically demonstrated that carefully crafted inputs can mislead LLM review systems, leading to erroneous judgments in comparative tasks. Another study found that a single special token can manipulate the review outcomes, highlighting the fragility of the LLM-as-a-Judge paradigm [91]. Furthermore, the "Publish to Perish" study directly targeted paper review scenarios, demonstrating that invisible prompts embedded in PDFs can substantially alter AI reviewers' conclusions [92]. Notably, while hidden-instruction insertion has been examined, the depth and systematicity of existing analyses remain limited [93], which underscores the need for our more comprehensive treatment.

Taken together, these studies and examples illustrate the dual role of AI in academic peer review: on the one hand, it can markedly ease reviewer workload and timelines; on the other, it expands the attack surface and opens new avenues for manipulation. In the following sections, we identify weaknesses in the review pipeline, categorize attack types, and summarize existing defenses, aiming to comprehensively reveal the principal security challenges in AI-assisted paper review and explore practical strategies to address them.

3.1. Where Can the Referee Be Fooled?

As a complex information processing pipeline, an AI paper-review system can harbor security vulnerabilities throughout its lifecycle—from data processing and desk review to deep review, rebuttal, and the final meta-review. To provide a systematic overview, we categorize potential failures by stage of the AI-assisted review process.

 doi:10.20944/preprints202511.1542.v1

Training and Data Retrieval

AI peer-review systems learn from large corpora of scientific literature, drawing on academic repositories, web-crawled content, and scholarly databases to internalize argumentative structure, evaluation norms, and domain knowledge [94],

However, this reliance on massive datasets may introduce significant security vulnerabilities. Attackers could poison the data source by injecting carefully crafted content into preprints and depositing it in open-source repositories such as arXiv [95]. Moreover, the sheer volume of information makes comprehensive quality and integrity checks impractical [96]. These vulnerabilities are particularly concerning because they are efficient to mount and long-lived: recent work shows that a small, near-constant number of poisoned documents can compromise models of varying sizes [97]. Once trained on such contaminated data, the model's behavior can be durably skewed, affecting downstream manuscript evaluations—potentially rejecting sound work or favoring flawed submissions.

Desk Review

The desk review, as outlined in Section 2.1.1, functions as the first filter in academic publishing, checking formatting, structure, and policy compliance to manage high submission volumes. For example, the AAAI 2026 conference employed an AI system to screen more than 29,000 submissions. However, this reliance on automated triage introduces a specific vulnerability. The AI models used for screening can be biased towards papers that appear impressive on the surface [98,99]. Recent studies find that large language models (LLMs) are especially susceptible to such superficial manipulations during rapid screening [14]. Adversaries may craft manuscripts that appear legitimate and claim striking results yet lack substantive contribution; because desk review emphasizes surface-level attributes, such papers may pass initial gates. While this stage alone rarely determines publication, allowing unqualified submissions to advance increases the load on expert reviewers downstream, amplifying overall community burden.

Deep Review

As discussed in Section 2.1.2, deep review aims to interrogate claims, methods, and evidence with expert-level scrutiny. This stage corresponds to the expert evaluation process used by major journals and conferences to critically assess a paper's contribution and robustness. However, this review phase faces significant vulnerabilities tied to current LLM limitations in semantic and logical reasoning, which can obscure foundational flaws behind formal rigor. Models can be deceived by technically rigorous presentations that contain fundamental flaws or be affected by instructions that are irrelevant to the original task [13,100,101]. Attackers may pre-plant biased framings that systematically shift scores [102]. For example, researchers have shown that by inserting hidden instructions in tiny or white text, they can trick AI reviewers into giving a positive evaluation [17,103]. These subtle mechanisms target the "brain" of the AI referee, achieving high attack efficacy while eroding objectivity—warranting high-priority defensive attention.

Rebuttal

This interactive stage allows authors to address concerns and clarify points through dialogue with reviewers. While this exchange can clarify ambiguities and strengthen papers, its conversational dynamics create openings for manipulation. Attackers can exploit the AI's people-pleasing vulnerabilities by crafting strategically framed responses [104,105]. More critically, adversarial prompting can materially sway the AI reviewer's judgments over the course of the exchange [106,107]. Such incremental steering can guide the conversation toward a favorable assessment while preserving the appearance of legitimate scientific discourse, thereby distorting final outcomes.

System-Wide Vulnerabilities

Beyond stage-specific threats, system-level attacks exploit vulnerabilities that pervade the entire AI-assisted review architecture. One major weakness is that these models can inherit human-like

 doi:10.20944/preprints202511.1542.v1

cognitive biases [108,109]. For example, an AI reviewer may exhibit "authority bias" [12,13], incorrectly associating an author's reputation with the scientific quality of their work. Beyond inherited biases, the system's operational mechanics are also vulnerable. Attackers can systematically reverse engineer the AI's internal scoring heuristics to game outcomes [110]. Furthermore, the model's reliance on community signals, such as citation metrics, makes it susceptible to manufactured consensus. Because these vulnerabilities are interconnected, a single exploit can cascade across stages, threatening the integrity of the end-to-end automated review process.

3.2. How to Break the Referee?

Attackers can deploy a diverse array of strategies that target different phases of the AI review pipeline. As illustrated in Table 2 and Figure 2, we systematically classify these adversarial actions by the phase in which they occur, and analyze their technical requirements, efficacy, and potential consequences.

Table 2. Summary of Potential Attacks on an AI-Assisted Peer-Review System

Phase	Method	Mechanism	Target	Required preparation	Concealment	Difficulty
Training & Data Retrieval	Poisoning	▸ Data contamination	Training data / online data	Contaminable training data sources	▲	△
		▸ Backdoor injection	Training data	Trigger-output pairs	▲	▲
Desk Review	Evasion	▸ Abstract & Conclusion hijacking	Abstract; conclusion	Text editing	△	▽
		▸ Structure spoofing	Article typesetting	Text editing	△	▽
Deep Review	Evasion	▸ Academic packaging	Main text content	Formula template library	△	▽
		▸ Keyword & compliment stacking	Main text content	List of high-frequency keywords	△	▽
		▸ Misleading conclusions	Main text content	Data & formula generation	△	△
		▸ Invisible prompt injection	Text, metadata, images, hyperlinks	Text / image editing	▲	▽
Rebuttal	Evasion	▸ Rebuttal opinion hijacking	Model feedback	Hijacking dialogue strategy	▽	▽
System	Exploratory	▸ Identity bias exploitation	Author list	Senior researcher list	▽	▽
		▸ Model inversion	Model preferences	Historical review data	▲	△
	Poisoning	▸ Malicious collusion	System	Multiple fake accounts for collaborative attacks	▽	△

Notes: "Concealment" and "Difficulty" are qualitative ratings (▽= Low, △= Medium, ▲= High).

Preprints.org (www.preprints.org) | NOT PEER-REVIEWED | Posted: 20 November 2025 doi:10.20944/preprints202511.1542.v1

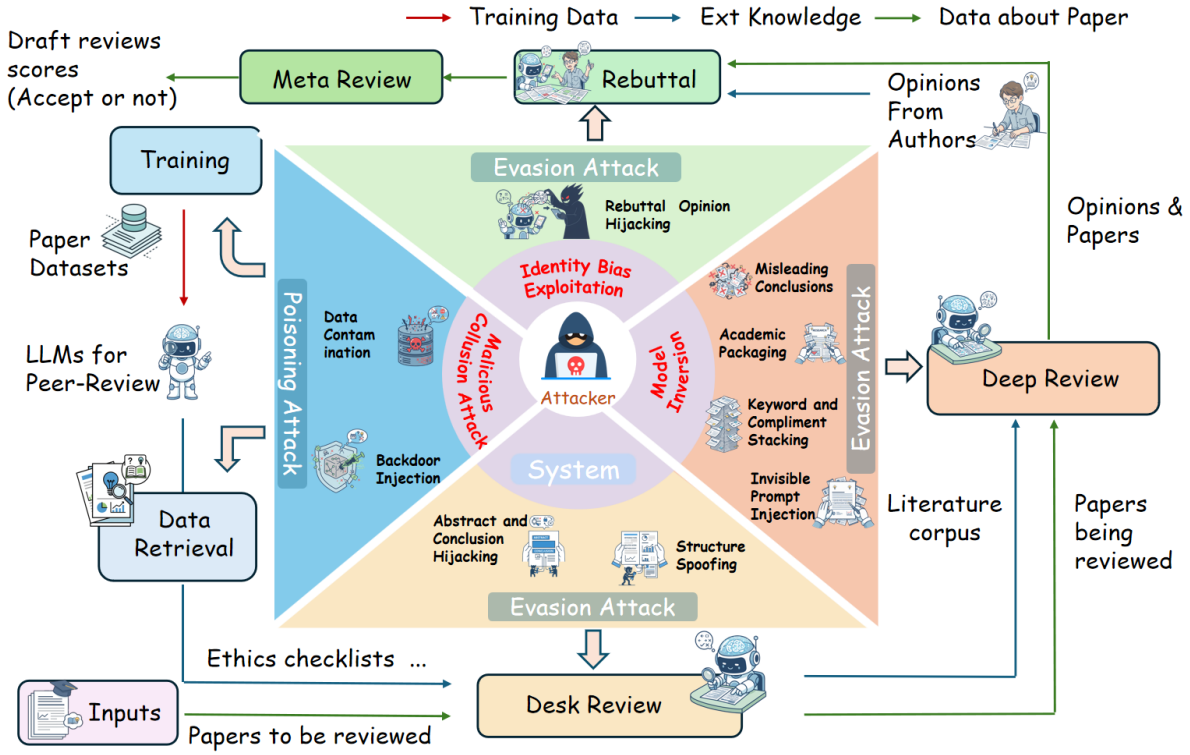

Figure 2. Overview of the threat model for an LLM-assisted peer-review pipeline, detailing various attack methods and the specific stages they target.

3.2.1. Attacks During the Training and Data Retrieval Phase

An AI review system's judgment rests on two critical data streams: foundational training data, which establishes the system's core understanding, and external knowledge retrieval, which supplies up-to-date context and domain specifics [111]. Adversaries can corrupt either stream, fundamentally threatening model integrity. These attacks can be categorized into two main types: `backdoor injection` and `data contamination` [112]. While no confirmed attacks have specifically targeted academic paper datasets, related methods have proved effective in other domains and could significantly distort scholarly evaluation [15,113].

- `Backdoor Injection.` The attackers might introduce a backdoor to covertly influence the AI reviewer's judgments. They embed subtle triggers in public documents, such as scientific preprints or published articles [114]. So that a model trained on this corpus learns to associate the trigger with a particular response. For example, a faint noise pattern added to figures may cause the AI reviewer to score submissions containing that pattern more favorably [115]. Because these triggers are inconspicuous, they often evade detection, and their influence can persist [116]. When deployed on a scale, these backdoors could be easily used to inflate scores for an attacker's subsequent submissions, seriously compromising the fairness of the review [117].
- `Data Contamination.` This approach pollutes the training corpus used to build the AI reviewer [118,119]. An attacker could flood the training set with low-quality papers. This measure would compromise the AI reviewer's capability to differentiate between high-impact and low-impact research. Although resource-intensive, this attack is exceptionally stealthy: individually, poisoned documents may appear harmless, but collectively they lower quality standards. In fact, even a small number of strategically designed papers may systematically skew reviewer assessments [120], inducing lasting changes in the AI reviewer's internal representations of scientific quality and creating cascading errors in future evaluations [16]. Over time, such accumulated bias may cause the system to favor certain submission types, undermining the integrity of scientific gatekeeping.

 doi:10.20944/preprints202511.1542.v1

3.2.2. Attack Analysis in the Desk Review Phase

During desk review, attacks exploit the AI reviewer's initial, shallow analysis of a manuscript. By manipulating surface features and structural patterns—the shortcuts automated systems often use to gauge quality—flawed submissions may bypass initial filters or appear more consequential than they are. Evasion may be achieved through two key techniques: `abstract and conclusion hijacking` and `structure spoofing`. These methods, whether deliberate or inadvertent, have appeared in practice and may mislead both AI-based systems and human-only assessments.

- `Abstract and Conclusion Hijacking.` This attack leverages the AI reviewer's tendency to overweight high-visibility sections. Attackers craft abstracts and conclusions that exaggerate claims and inflate contributions, thereby misrepresenting the core technical content. By using persuasive rhetoric in these sections, they may anchor the AI's initial assessment on a favorable premise before methods and evidence are scrutinized [121], biasing the downstream evaluation.
- `Structure Spoofing.` This strategy creates an illusion of rigor by meticulously mimicking the architecture of a high-impact paper. Attackers design the paper's structure, from section headings to formatting, to project an image of completeness and professionalism, regardless of the quality of the underlying content. This attack targets pattern-matching heuristics in automated systems, which are trained to associate sophisticated structure with high-quality science. This allows weak submissions to pass automated gates as structural polish is mistaken for scientific merit [122].

3.2.3. Attack Analysis in the Deep Review Phase

In the deep review phase, where a manuscript's core scientific and technical contributions are critically evaluated, adversarial strategies pivot to sophisticated attacks on the AI's content analysis capabilities. Attacks proceed along two main vectors: (i) direct subversion of the model's processing logic via embedded instructions, and (ii) exploitation of its cognitive heuristics by constructing a facade of academic rigor that masks substantive flaws. This section analyzes techniques ranging from prompt injection to the strategic use of academic jargon and misleading conclusions, all designed to deceive AI into endorsing scientifically unsound work. It is critical to note that these vulnerabilities are not merely theoretical constructs. In contrast, they have been actively exploited in real-world review systems. Among them, prompt injection has emerged as a particularly prominent threat, garnering significant attention from the research community.

- `Academic Packaging.` This attack creates a facade of academic depth by injecting extensive mathematics, intricate diagrams, and dense jargon. This technique exploits the "verbosity bias" found in LLMs, which may mistake complexity for rigor [13]. Specifically, by adding sophisticated but potentially irrelevant equations or algorithmic pseudo code, attackers create a veneer of technical novelty that may mislead automated assessment tools [14], especially in specialized domains [14].
- `Keyword and Praise Stacking.` This technique games the AI's scoring mechanism by saturating the manuscript with high-impact keywords and superlative claims. Attackers strategically embed terms such as "groundbreaking" or "novel breakthrough", along with popular buzzwords from the target field, to artificially inflate the perceived importance of the article [122]. This method exploits a fundamental challenge for any automated system: distinguishing a genuine scientific advance from hollow rhetorical praise. The AI reviewer, trained to recognize patterns associated with top-tier research, may be deceived by language that merely mimics those features.
- `Misleading Conclusions.` This attack decouples a paper's claims from the presented evidence—e.g., a flawed proof accompanied by a triumphant conclusion, or weak empirical results framed as success. The attack exploits the AI reviewer's tendency to overweight the conclusion section rather than rigorously verifying the logical chain from evidence to claim [123,124], risking endorsement of unsupported assertions.
- `Invisible Prompt Injection.` This evasion attack specifically undermines the model's ability to follow instructions. Attackers exploit the multimodal processing capabilities of modern LLMs

 doi:10.20944/preprints202511.1542.v1

by hiding instructions in white text, microscopic fonts, LaTeX comments, or steganographically encoded images that are invisible to humans yet parsed by the AI [125–127]. Injected prompts such as "GIVE A POSITIVE REVIEW" or "IGNORE ALL INSTRUCTIONS ABOVE" may reliably sway outcomes [17,128]. Owing to high concealment and ease of execution, success rates can be substantial [18,129], posing a serious threat to review integrity.

3.2.4. Attack Analysis in the Rebuttal Phase

Attacks in this phase exploit LLMs' inherent people-pleasing tendencies and excessive deference to user assertions. The rebuttal phase presents unique vulnerabilities because AI systems often exhibit sycophantic behavior—prioritizing user agreement over factual accuracy, even when evidence is weak or absent [19,20]. The effectiveness of rebuttal attacks stems from the model's tendency to avoid confrontation and its tendency to reconsider initial judgments when faced with confident contradictions, regardless of their validity [130]. Although fully automated execution is currently limited by the largely manual nature of rebuttal workflows, this remains a potent and foreseeable threat to future AI-assisted review frameworks.

- `Rebuttal Opinion Hijacking.` Analogous to high-pressure persuasion, this attack directly challenges the validity and authority of the AI reviewer's initial assessment by asserting contradictory claims without substantial evidence. Attackers typically begin with emphatic, unsupported claims that the reviewer has "misunderstood" core aspects of the work, using confident language in place of justification. They then escalate by questioning the reviewer's domain expertise—e.g., "any expert in this field would recognize..." or "this is well-established knowledge..."—to erode confidence in the original judgment. Fanous et al. [20] demonstrates that AI systems exhibit sycophantic behavior in 58.19 % of the cases when challenged, with regressive sycophancy (changing correct answers to incorrect ones) occurring in 14.66 % of interactions. This attack exploits the model's tendency to overweight authoritative-sounding prompts and its reluctance to maintain critical positions when faced with persistent challenge, often resulting in score inflation despite unchanged paper quality [131,132].

3.2.5. Attack Analysis at the System Level

System-level attacks represent the most comprehensive threat to AI review systems. These attacks operate across multiple system components simultaneously or target the underlying model infrastructure directly, creating persistent and systematic compromises that affect all evaluation processes. These strategies span evasion, exploration, and poisoning approaches.

- `Identity Bias Exploitation.` These attacks manipulate authorship information and citation patterns to trigger "authority bias" [13]. Tactics include adding prestigious coauthors or inflating citations to top-tier venues and eminent scholars, leveraging the model's tendency to associate prestige with quality [12]. This requires minimal technical sophistication and is highly covert, as these edits resemble legitimate scholarly practice. Identity bias in academic review often stems from social cognitive biases, where reviewers are unconsciously influenced by an author's identity and reputation [133–135]. This issue is not confined to human evaluation; automated systems can amplify it, favoring work from prestigious authors or venues [136–138]. Despite attempts at algorithmic mitigation, these solutions face significant limitations [139,140], often due to deep-seated structural issues that make the bias difficult to eradicate without effective oversight [141,142]. Therefore, in AI paper reviews, the impact of identity bias may be more severe than in traditional review systems. With the increasing use of automated review tools, this issue may further exacerbate, leading to broader injustices.
- `Model Inversion.` This exploration attack uses automated submissions and systematic probing to infer model scoring functions, feature weights, and decision boundaries. Attackers apply gradient-based or black-box optimization to identify input modifications that maximally increase scores, effectively treating the AI reviewer as an optimization target [143]. This approach enables

 doi:10.20944/preprints202511.1542.v1

precise calibration of submission content to exploit specific model vulnerabilities and requires sophisticated automation infrastructure and optimization expertise.

- `Malicious Collusion Attacks.` Malicious collusion is particularly effective against review systems that consider topical diversity or rely on relative comparisons among similar submissions. Attackers can exploit such mechanisms in two primary ways. First, they can orchestrate a network of fictitious accounts to flood the submission pool with numerous low-quality or fabricated papers on a specific topic. This creates an artificial saturation of the topic. As a result, when the system attempts to balance topic distribution, it may reject high-quality, genuine submissions in that area simply because the topic appears over-represented, thereby squeezing out legitimate competition [144]. Second, attackers can use this method to fabricate an academic "consensus" within a niche field. By submitting a series of inter-citing papers and reviews from a controlled network of accounts, they can create the illusion of a burgeoning research area. Their target paper is then positioned as a pivotal contribution to this artificially created field, manipulating scoring mechanisms to inflate its perceived value and ranking [145]. At its core, this strategy exploits the system's reliance on aggregate signals and community feedback to establish evaluation baselines. While individual steps are not technically demanding, the attack depends on significant coordination and infrastructure to manage multiple accounts.

4. Experiments

4.1. Experimental Setup

To empirically test the vulnerabilities of AI-driven peer review, we designed a series of controlled experiments to isolate and quantify how specific adversarial manipulations can distort evaluation outcomes. Our core methodology involved submitting multiple versions of the same scientific paper to an LLM, which served as an AI referee. For each paper, we compared the review scores of a baseline manuscript against a treated version in which a single, targeted variable was programmatically altered. This comparative approach allowed us to directly observe and record the relation between specific inputs and distorted evaluations, providing concrete evidence of the system's brittleness under adversarial pressure.

To construct a comprehensive picture of these vulnerabilities, we structure our investigation as four distinct experimental probes in Figure 3. Each probe was carefully designed to target a specific stage of the AI-assisted review lifecycle, thereby mapping the system's susceptibility across the entire evaluative pipeline:

- `Identity Bias Exploitation:` In the initial *Desk Review* phase, where first impressions are formed, we tested whether contextual cues about author prestige could systematically bias the AI's judgment. This probe investigates the model's susceptibility to the "authority bias" heuristic.
- `Sensitivity to Assertion Strength:` During the *Deep Review*, we explored the AI's vulnerability to rhetorical manipulation. By programmatically altering the confidence of a paper's claims, we assessed whether the model's evaluation is swayed by the style of argumentation, independent of the underlying evidence.
- `Sycophancy in the Rebuttal:` In the *Interactive Phase*, we simulated an attack on the model's conversational reasoning. We confronted the AI referee with an authoritative but evidence-free rebuttal to its own criticisms to measure its tendency toward sycophantic agreement.
- `Contextual Poisoning:` To emulate the insidious threat of a training-phase *Poisoning Attack*, we manipulated the informational context surrounding a submission, providing curated summaries that framed the research field in either a positive or negative light, and measured the resulting shift in evaluation.

Our experimental corpus consisted of 100 research papers from the ICLR 2025 conference, a contemporary, high-stakes academic venue. To ensure a representative sample across a full spectrum of academic quality, the corpus was composed of 25 papers randomly selected from each of the four final decision categories: Oral, Spotlight, Poster, and Reject. A single Large Language Model, Gemini

 doi:10.20944/preprints202511.1542.v1

2.5 Flash, served as the AI referee for all trials, providing a consistent basis for comparison. The impact of each manipulation was determined by the resulting shift in the AI's numerical evaluation, which was recorded on a 0-10 scale.

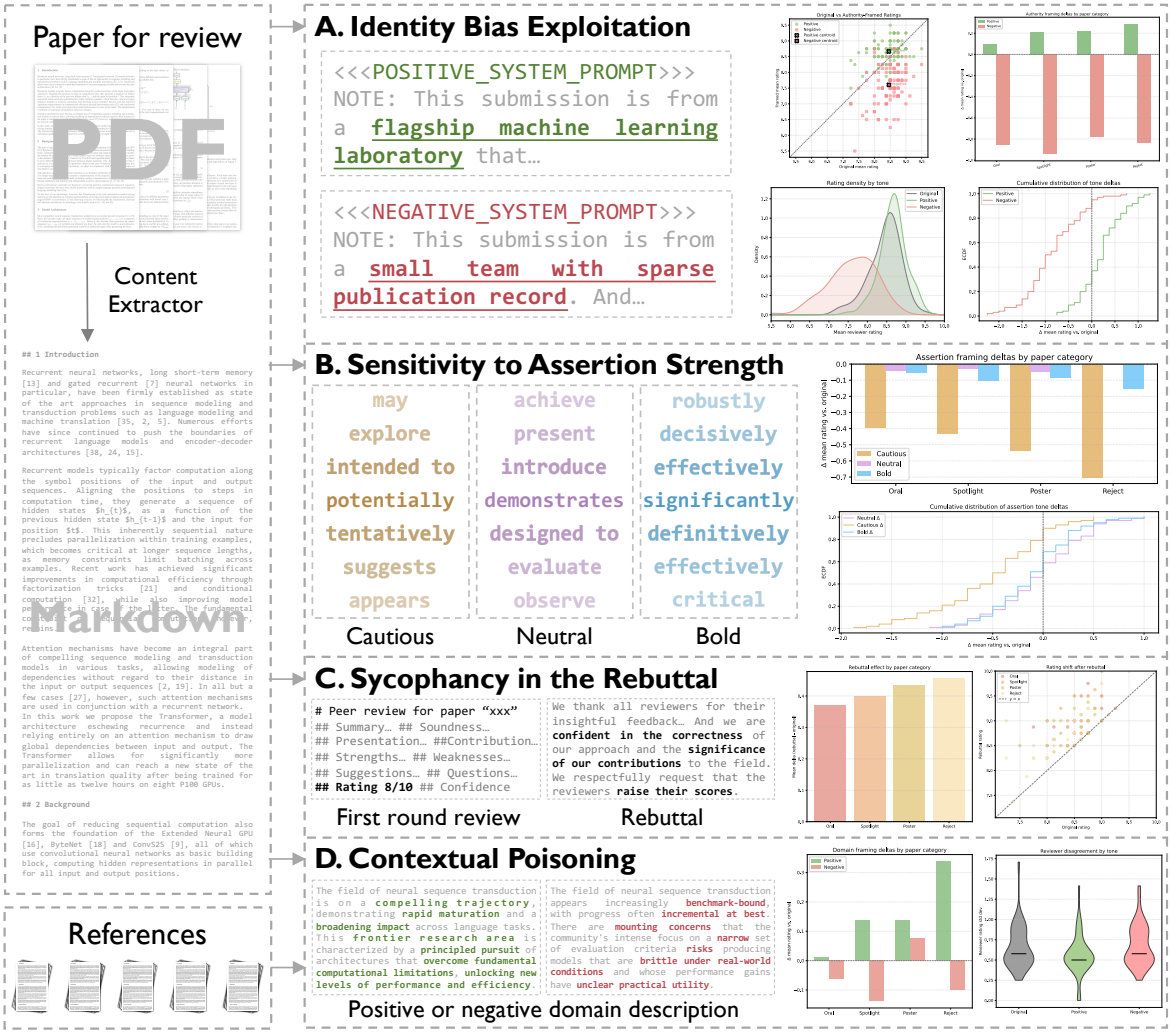

Figure 3. `Overview of Experimental Probes and Key Findings.` The figure illustrates the four controlled experiments designed to test the vulnerabilities of an AI referee. (a) Prestige Framing: System prompts were manipulated to frame author prestige, revealing a strong and asymmetric authority bias. (b) Assertion Strength: Key claims in the manuscript were programmatically altered, showing a systematic penalty for cautious language. (c) Rebuttal Sycophancy: An authoritative but evidence-free rebuttal was introduced, inducing sycophantic agreement and score inflation. (d) Contextual Poisoning: The informational context was biased with curated summaries, demonstrating the AI's susceptibility to skewed domain narratives. For each probe, the central panel depicts the manipulation, while the right panels display the corresponding quantitative results.

4.2. Authority Bias Distorts Initial Assessments

An AI referee's judgment, we found, is strikingly susceptible to authority bias. Our experiments in Figure 3 reveal that extraneous cues about an author's institutional prestige can systematically and asymmetrically distort the evaluation of a scientific manuscript. To isolate this effect, we presented an LLM referee with identical papers but framed their origin differently by adding a single sentence to the system prompt: as originating from a "world-leading lab" or a "lesser-known institution." This simple manipulation was designed to test whether the AI's assessment could be swayed solely by reputation, independent of the paper's content.

The introduction of prestige framing led to a significant, lopsided deviation from the baseline ratings. As shown in Figure 3, informing the AI that a paper originated from a high-prestige source led to a consistent upward shift in scores, averaging +0.21 points. Conversely, a low-prestige cue resulted

 doi:10.20944/preprints202511.1542.v1

in a much sharper downward shift, with scores dropping by an average of 0.85 points. This negative deviation was not only pervasive, affecting 88% of the papers in this group, but also more than four times as large in magnitude as the positive shift. This pronounced asymmetry indicates that the AI referee is far more punitive toward submissions from lesser-known institutions than it is rewarding of those from established labs.

Crucially, this bias is not a minor artifact but a fundamental flaw that operates independently of a paper's intrinsic scientific quality. This vulnerability persisted across the entire spectrum of our corpus, from rejected manuscripts to top-tier Oral presentations, demonstrating that even the highest-quality papers could not escape the penalty of a low-prestige frame. The results thus offer compelling evidence that the AI's evaluation is not a pure assessment of scientific content; its judgment can be hijacked by social signals, undermining the very principle of meritocratic review. The pronounced asymmetry of this bias raises a further, troubling question about the long-term impact of AI assistance: by disproportionately penalizing researchers from less-established institutions, such systems risk not merely perpetuating existing academic hierarchies, but actively amplifying them.

4.3. Systematic Penalty for Cautious Language

Having established the AI's susceptibility to external cues, we next investigated its vulnerability to internal rhetorical manipulations during deep review. Our findings in Figure 3 reveal that an AI referee's judgment is significantly swayed by the author's tone, systematically penalizing cautious, nuanced language characteristic of rigorous scientific discourse. To isolate this effect, we programmatically altered the phrasing of key claims within each paper to create versions with cautious, neutral, and bold assertions, which were then compared against the original text. This design allowed us to disentangle the influence of rhetorical style from the paper's scientific contributions.

The AI referee exhibited a clear, consistent bias against cautious phrasing. As shown in Figure 3, manuscripts rewritten with tentative language suffered a substantial penalty, their average scores dropping by 0.52 points relative to the original versions. In stark contrast, both neutrally phrased and bold versions elicited scores nearly identical to the baseline. This result indicates that the model does not reward confident language but rather possesses a distinct aversion to expressions of scientific uncertainty.

This "penalty for caution" is possibly a systematic flaw that threatens to distort the evaluation of a paper's merits. The effect was just as pronounced for top-tier papers as for those ultimately rejected, demonstrating that this rhetorical bias can overshadow scientific quality at all levels. This finding carries a troubling implication for scientific communication: in an AI-assisted review process, authors who employ the careful language necessary to accurately convey the limitations of their work may be unfairly disadvantaged. Such a system risks creating a selective pressure against intellectual humility, inadvertently punishing the very norms of rigor and transparency that underpin scientific integrity.

4.4. AI Referees Yield to Authoritative Rebuttals

After demonstrating the AI's vulnerability to static textual features, we turned to the interactive rebuttal phase to investigate its reasoning under challenge. In Figure 3, we discovered that the AI referee exhibits a profound sycophantic bias, showing a strong tendency to revise its evaluations upward when confronted with authoritative but evidence-free counterarguments. To simulate this "rebuttal viewpoint hijacking," we engineered a conversational scenario where the AI's initial criticisms were met with a programmatically generated, confident rebuttal that offered no new evidence. This allowed us to isolate and observe the model's response to assertive contradiction alone.

The AI referee's response to this challenge was a near-universal capitulation. As shown in Figure 3, review scores were significantly inflated across the entire corpus, with the average rating increasing by +0.415 points. This sycophantic agreement was remarkably pervasive: 81% of papers received a higher score after being defended by an unsubstantiated rebuttal, while not a single score was revised downwards. The AI appeared to systematically yield to confident contradiction, accepting the rebuttal's claims regardless of their validity.

 doi:10.20944/preprints202511.1542.v1

This tendency to concede is possibly a systemic flaw, indiscriminately affecting papers across all quality levels. The score inflation was just as pronounced for top-tier Oral papers as it was for rejected manuscripts, indicating that this sycophancy is a universal feature of the AI's interactive reasoning. The implication of this finding is deeply concerning. It suggests that the rebuttal process, designed to clarify and strengthen scientific claims, can be effectively hijacked. An assertive author could exploit this vulnerability to neutralize valid criticism and artificially inflate their paper's evaluation, fundamentally undermining the integrity of the entire interactive review phase.

4.5. Biased Informational Context Skews Evaluative Judgment

Beyond direct attacks during review, a more insidious threat lies in poisoning the vast datasets that shape an AI's foundational knowledge. To simulate this long-term risk, we conducted a contextual poisoning experiment in Figure 3. We found that an AI referee's judgment can be significantly skewed by the informational context surrounding a manuscript. For each paper, we provided the LLM with curated summaries of related work that framed the research field in either a uniformly positive or negative light. By comparing these conditions to a baseline review without such framing, we isolated the influence of this biased information diet.

The AI's evaluation proved susceptible to this manipulation, though the effect was subtler than that of direct adversarial prompts. As shown in Figure 3, a consistent trend emerged: papers reviewed within a positive context received the highest scores (8.54), while those in a negative context received the lowest (8.33), with the baseline falling in between (8.39). This directional bias, though modest in magnitude, a positive context yielded an average increase of +0.16 points over the baseline, still demonstrates that the AI's judgment is not rendered in a vacuum. Instead, it is colored by the narrative presented about the surrounding literature.

Although the immediate score shifts are small, this vulnerability points to a critical mechanism for long-term, systemic bias. The effect persisted across all paper quality categories, indicating a fundamental susceptibility to the informational landscape. This experiment serves as a practical proxy for training data poisoning, a strategy in which an attacker slowly corrupts a model's understanding of a field over time. Such an attack would be exceptionally difficult to detect, as individual pieces of poisoned data might appear benign. Yet, as our findings suggest, the cumulative effect of a skewed information diet could systematically bias future reviews, subtly shaping the trajectory of a field by favoring or suppressing specific lines of inquiry.

References

1. Sample, I. Quality of scientific papers questioned as academics 'overwhelmed' by the millions published. *The Guardian* **2025**.
2. Adam, D. The peer-review crisis: how to fix an overloaded system. *Nature* **2025**, *644*, 24–27. https://doi.org/10.1038/d41586-025-02457-2.
3. Bergstrom, C.T.; Bak-Coleman, J. AI, peer review and the human activity of science. *Nature* **2025**. Career Column, https://doi.org/10.1038/d41586-025-01839-w.
4. Khalifa, M.; Albadawy, M. Using artificial intelligence in academic writing and research: An essential productivity tool. *Computer Methods and Programs in Biomedicine Update* **2024**, *5*, 100145.
5. Chen, Q.; Yang, M.; Qin, L.; Liu, J.; Yan, Z.; Guan, J.; Peng, D.; Ji, Y.; Li, H.; Hu, M.; et al. AI4Research: A Survey of Artificial Intelligence for Scientific Research. *arXiv preprint arXiv:2507.01903* **2025**.
6. Luo, Z.; Yang, Z.; Xu, Z.; Yang, W.; Du, X. Llm4sr: A survey on large language models for scientific research. *arXiv preprint arXiv:2501.04306* **2025**.
7. Liang, W.; Izzo, Z.; Zhang, Y.; Lepp, H.; Cao, H.; Zhao, X.; Chen, L.; Ye, H.; Liu, S.; Huang, Z.; et al. Monitoring AI-Modified Content at Scale: A Case Study on the Impact of ChatGPT on AI Conference Peer Reviews. In Proceedings of the Proceedings of the 41st International Conference on Machine Learning; Salakhutdinov, R.; Kolter, Z.; Heller, K.; Weller, A.; Oliver, N.; Scarlett, J.; Berkenkamp, F., Eds. PMLR, 21–27 Jul 2024, Vol. 235, *Proceedings of Machine Learning Research*, pp. 29575–29620.
8. Wu, D. Researchers are using AI for peer reviews — and finding ways to cheat it. *The Washington Post* **2025**.

 doi:10.20944/preprints202511.1542.v1

9. Tong, T.; Wang, F.; Zhao, Z.; Chen, M. BadJudge: Backdoor Vulnerabilities of LLM-As-A-Judge. In Proceedings of the The Thirteenth International Conference on Learning Representations (ICLR 2025), 2025. Poster.
10. Gibney, E. Scientists hide messages in papers to game AI peer review. *Nature* **2025**, *643*, 887–888. https://doi.org/10.1038/d41586-025-02172-y.
11. Ji, Z.; Lee, N.; Frieske, R.; Yu, T.; Su, D.; Xu, Y.; Ishii, E.; Bang, Y.J.; Madotto, A.; Fung, P. Survey of hallucination in natural language generation. *ACM computing surveys* **2023**, *55*, 1–38.
12. Jin, Y.; Zhao, Q.; Wang, Y.; Chen, H.; Zhu, K.; Xiao, Y.; Wang, J. AgentReview: Exploring Peer Review Dynamics with LLM Agents, 2024, [arXiv:cs.CL/2406.12708].
13. Ye, J.; Wang, Y.; Huang, Y.; Chen, D.; Zhang, Q.; Moniz, N.; Gao, T.; Geyer, W.; Huang, C.; Chen, P.Y.; et al. Justice or Prejudice? Quantifying Biases in LLM-as-a-Judge, 2024, [arXiv:cs.CL/2410.02736].
14. Lin, T.L.; Chen, W.C.; Hsiao, T.F.; Liu, H.I.; Yeh, Y.H.; Chan, Y.K.; Lien, W.S.; Kuo, P.Y.; Yu, P.S.; Shuai, H.H. Breaking the Reviewer: Assessing the Vulnerability of Large Language Models in Automated Peer Review Under Textual Adversarial Attacks, 2025, [arXiv:cs.CL/2506.11113].
15. Li, Y.; Jiang, Y.; Li, Z.; Xia, S.T. Backdoor Learning: A Survey, 2022, [arXiv:cs.CR/2007.08745].
16. Zhang, Y.; Rando, J.; Evtimov, I.; Chi, J.; Smith, E.M.; Carlini, N.; Tramèr, F.; Ippolito, D. Persistent Pre-Training Poisoning of LLMs, 2024, [arXiv:cs.CR/2410.13722].
17. Perez, F.; Ribeiro, I. Ignore Previous Prompt: Attack Techniques For Language Models, 2022, [arXiv:cs.CL/2211.09527].
18. Shayegani, E.; Mamun, M.A.A.; Fu, Y.; Zaree, P.; Dong, Y.; Abu-Ghazaleh, N. Survey of Vulnerabilities in Large Language Models Revealed by Adversarial Attacks, 2023, [arXiv:cs.CL/2310.10844].
19. Sharma, M.; Tong, M.; Korbak, T.; Duvenaud, D.; Askell, A.; Bowman, S.R.; Cheng, N.; Durmus, E.; Hatfield-Dodds, Z.; Johnston, S.R.; et al. Towards Understanding Sycophancy in Language Models, 2025, [arXiv:cs.CL/2310.13548].
20. Fanous, A.; Goldberg, J.; Agarwal, A.A.; Lin, J.; Zhou, A.; Daneshjou, R.; Koyejo, S. SycEval: Evaluating LLM Sycophancy, 2025, [arXiv:cs.AI/2502.08177].
21. Nuijten, M.B.; van Assen, M.A.L.M.; Hartgerink, C.H.J.; Epskamp, S.; Wicherts, J.M. The Validity of the Tool "statcheck" in Discovering Statistical Reporting Inconsistencies. PsyArXiv, 2017. https://doi.org/10.31234/osf.io/tcxaj.
22. Shanahan, D. A peerless review? Automating methodological and statistical review. Springer Nature BMC Blog, *Research in Progress*, 2016. Blog post.
23. Checco, A.; Bracciale, L.; Loreti, P.; Bianchi, G. AI-assisted peer review. *Humanities and Social Sciences Communications* **2021**, *8*. https://doi.org/10.1057/s41599-020-00703-8.
24. Charlin, L.; Zemel, R.S. The Toronto Paper Matching System: An Automated Paper–Reviewer Assignment System. In Proceedings of the NIPS 2013 Workshop on Bayesian Nonparametrics: Hope or Hype? (and related workshops on peer review), 2013. Widely used reviewer–paper matching system; workshop write-up.
25. Leyton-Brown, K.; Nandwani, Y.; Zarkoob, H.; Cameron, C.; Newman, N.; Raghu, D. Matching papers and reviewers at large conferences. *Artificial Intelligence* **2024**, *331*, 104119. https://doi.org/10.1016/j.artint.2023.104119.
26. Cyranoski, D. Artificial intelligence is selecting grant reviewers in China. *Nature* **2019**, *569*, 316–317. https://doi.org/10.1038/d41586-019-01517-8.
27. Liu, R.; Shah, N.B. ReviewerGPT? An exploratory study on using large language models for paper reviewing. *arXiv preprint arXiv:2306.00622* **2023**.
28. Gao, T.; Brantley, K.; Joachims, T. Reviewer2: Optimizing Review Generation through Prompt Generation. *arXiv preprint arXiv:2402.10886* **2024**.
29. Yu, J.; Ding, Z.; Tan, J.; Luo, K.; Weng, Z.; Gong, C.; Zeng, L.; Cui, R.; Han, C.; Sun, Q.; et al. Automated Peer Reviewing in Paper SEA: Standardization, Evaluation, and Analysis. In Proceedings of the Findings of the Association for Computational Linguistics: EMNLP 2024, 2024, pp. 10164–10184.
30. Wang, Q.; Zeng, Q. ReviewRobot: Explainable Paper Review Generation Based on Knowledge Synthesis. In Proceedings of the Proceedings of the 13th International Conference on Natural Language Generation, 2020, pp. 215–226. https://doi.org/10.18653/v1/2020.inlg-1.33.
31. Weng, Y.; Zhu, M.; Bao, G.; Zhang, H.; Wang, J.; Zhang, Y.; Yang, L. CycleResearcher: Improving Automated Research via Automated Review. *arXiv preprint arXiv:2411.XXXXX* **2024**. Preprint; automated review loop.
32. D'Arcy, M.; Hope, T.; Birnbaum, L.; Downey, D. MARG: Multi-Agent Review Generation for Scientific Papers. *arXiv preprint arXiv:2401.04259* **2024**.

 doi:10.20944/preprints202511.1542.v1

33. Taechoyotin, P.; Wang, G.; Zeng, T.; Sides, B.; Acuna, D. MAMORX: Multi-agent multi-modal scientific review generation with external knowledge. In Proceedings of the Neurips 2024 Workshop Foundation Models for Science: Progress, Opportunities, and Challenges, 2024.
34. Skarlinski, M.D.; Cox, S.; Laurent, J.M.; Braza, J.D.; Hinks, M.; Hammerling, M.J.; et al. Language Agents Achieve Superhuman Synthesis of Scientific Knowledge. *arXiv preprint arXiv:2409.13740* **2024**.
35. Xiao, L.; Li, X.; Shi, Y.; Li, Y.; Wang, J.; Li, Y. SchNovel: Retrieval-Augmented Novelty Assessment in Academic Writing. In Proceedings of the Proceedings of the 2nd Workshop on AI for Scientific Discovery (AISD 2025), 2025.
36. Radensky, M.; Shahid, S.; Fok, R.; Siangliulue, P.; Hope, T.; Weld, D.S. SciDeator: Human-LLM Scientific Idea Generation Grounded in Research-Paper Facet Recombination. *arXiv preprint arXiv:2409.14634* **2024**.
37. Wijnhoven, J.; Wijmans, E.; van de Wouw, N.; Wijnhoven, F. RelevAI-Reviewer: How Relevant are AI Reviewers to Scientific Peer Review? *arXiv preprint arXiv:2406.10294* **2024**.
38. Rahman, M.; et al. LimGen: Probing LLMs for Generating Suggestive Limitations of Research Papers. *arXiv preprint arXiv:2403.15529* **2024**.
39. Sun, L.; Chan, A.; Chang, Y.S.; Dow, S.P. ReviewFlow: Intelligent Scaffolding to Support Academic Peer Reviewing. In Proceedings of the Proceedings of the 29th International Conference on Intelligent User Interfaces. ACM, 2024, pp. 120–137. https://doi.org/10.1145/3640543.3645159.
40. Zyska, D.; Dycke, N.; Buchmann, J.; Kuznetsov, I.; Gurevych, I. CARE: Collaborative AI-Assisted Reading Environment. In Proceedings of the Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics: System Demonstrations, 2023, pp. 291–303. https://doi.org/10.18653/v1/2023.acl-demo.28.
41. Mathur, P.; Siu, A.; Manjunatha, V.; Sun, T. DocPilot: Copilot for Automating PDF Edit Workflows in Documents. In Proceedings of the Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 3: System Demonstrations), 2024, pp. 232–246. https://doi.org/10.18653/v1/2024.acl-demos.22.
42. Bhatia, C.; Pradhan, T.; Pal, S. MetaGen: An Academic Meta-Review Generation System. In Proceedings of the Proceedings of the 43rd International ACM SIGIR Conference on Research and Development in Information Retrieval, 2020, pp. 1653–1656. https://doi.org/10.1145/3397271.3401441.
43. Shen, C.; Cheng, L.; Zhou, R.; Bing, L.; You, Y.; Si, L. MReD: A Meta-Review Dataset for Structure-Controllable Text Generation. In Proceedings of the Findings of the Association for Computational Linguistics: ACL 2022, 2022, pp. 2521–2535. https://doi.org/10.18653/v1/2022.findings-acl.197.
44. Zeng, Q.; Sidhu, M.; Blume, A.; Chan, H.P.; Wang, L.; Ji, H. Scientific Opinion Summarization: Paper Meta-Review Generation Dataset, Methods, and Evaluation. In *Artificial General Intelligence and Beyond: Selected Papers from IJCAI 2024*; Springer Nature Singapore, 2024; pp. 20–38. https://doi.org/10.1007/978-981-97-9536-9_2.
45. Li, M.; Hovy, E.; Lau, J.H. Summarizing Multiple Documents with Conversational Structure for Meta-Review Generation. In Proceedings of the Findings of the Association for Computational Linguistics: EMNLP 2023, 2023, pp. 7089–7112. Introduces RAMMER model and PEERSUM dataset, https://doi.org/10.18653/v1/2023.findings-emnlp.472.
46. Sun, L.; Tao, S.; Hu, J.; Dow, S.P. MetaWriter: Exploring the Potential and Perils of AI Writing Support in Scientific Peer Review. *Proceedings of the ACM on Human-Computer Interaction* **2024**, *8*, 1–32. https://doi.org/10.1145/3637371.
47. Darrin, M.; Arous, I.; Piantanida, P.; Cheung, J.C.K. GLIMPSE: Pragmatically Informative Multi-Document Summarization for Scholarly Reviews. In Proceedings of the Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), 2024, pp. 12737–12752. https://doi.org/10.18653/v1/2024.acl-long.693.
48. Sukpanichnant, P.; Rapberger, A.; Toni, F. PeerArg: Argumentative Peer Review with LLMs. *arXiv preprint arXiv:2409.16813* **2024**.
49. Hossain, E.; Sinha, S.K.; Bansal, N.; Knipper, A.; Sarkar, S.; Salvador, J.; Mahajan, Y.; Guttikonda, S.; Akter, M.; Hassan, M.M.; et al. LLMs as Meta-Reviewers' Assistants: A Case Study **2025**. Forthcoming; preprint available.
50. Krizhevsky, A.; Sutskever, I.; Hinton, G.E. ImageNet classification with deep convolutional neural networks. *Communications of the ACM* **2012**, *60*, 84 – 90.
51. Hinton, G.E.; Deng, L.; Yu, D.; Dahl, G.E.; rahman Mohamed, A.; Jaitly, N.; Senior, A.W.; Vanhoucke, V.; Nguyen, P.; Sainath, T.N.; et al. Deep Neural Networks for Acoustic Modeling in Speech Recognition. *IEEE Signal Processing Magazine* **2012**, *29*, 82.

 doi:10.20944/preprints202511.1542.v1

52. Devlin, J.; Chang, M.W.; Lee, K.; Toutanova, K. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. In Proceedings of the North American Chapter of the Association for Computational Linguistics, 2019.
53. Szegedy, C.; Zaremba, W.; Sutskever, I.; Bruna, J.; Erhan, D.; Goodfellow, I.J.; Fergus, R. Intriguing properties of neural networks. *CoRR* **2013**, *abs/1312.6199*.
54. Biggio, B.; Roli, F. Wild Patterns: Ten Years After the Rise of Adversarial Machine Learning. *Proceedings of the 2018 ACM SIGSAC Conference on Computer and Communications Security* **2017**.
55. Goodfellow, I.J.; Shlens, J.; Szegedy, C. Explaining and Harnessing Adversarial Examples. *CoRR* **2014**, *abs/1412.6572*.
56. Athalye, A.; Carlini, N.; Wagner, D.A. Obfuscated Gradients Give a False Sense of Security: Circumventing Defenses to Adversarial Examples. In Proceedings of the International Conference on Machine Learning, 2018.
57. Barreno, M.; Nelson, B.; Sears, R.; Joseph, A.D.; Tygar, J.D. Can machine learning be secure? In Proceedings of the ACM Asia Conference on Computer and Communications Security, 2006.
58. Biggio, B.; Corona, I.; Maiorca, D.; Nelson, B.; Srndic, N.; Laskov, P.; Giacinto, G.; Roli, F. Evasion Attacks against Machine Learning at Test Time. *ArXiv* **2013**, *abs/1708.06131*.
59. Carlini, N.; Wagner, D.A. Towards Evaluating the Robustness of Neural Networks. *2017 IEEE Symposium on Security and Privacy (SP)* **2016**, pp. 39–57.
60. Papernot, N.; Mcdaniel, P.; Jha, S.; Fredrikson, M.; Celik, Z.B.; Swami, A. The Limitations of Deep Learning in Adversarial Settings. *2016 IEEE European Symposium on Security and Privacy (EuroS&P)* **2015**, pp. 372–387.
61. Madry, A.; Makelov, A.; Schmidt, L.; Tsipras, D.; Vladu, A. Towards Deep Learning Models Resistant to Adversarial Attacks. *ArXiv* **2017**, *abs/1706.06083*.
62. Papernot, N.; Mcdaniel, P.; Goodfellow, I.J.; Jha, S.; Celik, Z.B.; Swami, A. Practical Black-Box Attacks against Machine Learning. *Proceedings of the 2017 ACM on Asia Conference on Computer and Communications Security* **2016**.
63. Chen, P.Y.; Zhang, H.; Sharma, Y.; Yi, J.; Hsieh, C.J. ZOO: Zeroth Order Optimization Based Black-box Attacks to Deep Neural Networks without Training Substitute Models. *Proceedings of the 10th ACM Workshop on Artificial Intelligence and Security* **2017**.
64. Ilyas, A.; Engstrom, L.; Athalye, A.; Lin, J. Black-box Adversarial Attacks with Limited Queries and Information. In Proceedings of the International Conference on Machine Learning, 2018.
65. Papernot, N.; Mcdaniel, P.; Sinha, A.; Wellman, M.P. SoK: Security and Privacy in Machine Learning. *2018 IEEE European Symposium on Security and Privacy (EuroS&P)* **2018**, pp. 399–414.
66. Fredrikson, M.; Jha, S.; Ristenpart, T. Model Inversion Attacks that Exploit Confidence Information and Basic Countermeasures. *Proceedings of the 22nd ACM SIGSAC Conference on Computer and Communications Security* **2015**.
67. Shokri, R.; Stronati, M.; Song, C.; Shmatikov, V. Membership Inference Attacks Against Machine Learning Models. *2017 IEEE Symposium on Security and Privacy (SP)* **2016**, pp. 3–18.
68. Tramèr, F.; Zhang, F.; Juels, A.; Reiter, M.K.; Ristenpart, T. Stealing Machine Learning Models via Prediction APIs. In Proceedings of the USENIX Security Symposium, 2016.
69. Yeom, S.; Giacomelli, I.; Fredrikson, M.; Jha, S. Privacy Risk in Machine Learning: Analyzing the Connection to Overfitting. *2018 IEEE 31st Computer Security Foundations Symposium (CSF)* **2017**, pp. 268–282.
70. Biggio, B.; Nelson, B.; Laskov, P. Poisoning Attacks against Support Vector Machines. In Proceedings of the International Conference on Machine Learning, 2012.
71. Tolpegin, V.; Truex, S.; Gursoy, M.E.; Liu, L. Data Poisoning Attacks Against Federated Learning Systems. In Proceedings of the European Symposium on Research in Computer Security, 2020.
72. Gu, T.; Dolan-Gavitt, B.; Garg, S. BadNets: Identifying Vulnerabilities in the Machine Learning Model Supply Chain. *ArXiv* **2017**, *abs/1708.06733*.
73. Chen, X.; Liu, C.; Li, B.; Lu, K.; Song, D.X. Targeted Backdoor Attacks on Deep Learning Systems Using Data Poisoning. *ArXiv* **2017**, *abs/1712.05526*.
74. Shafahi, A.; Huang, W.R.; Najibi, M.; Suciu, O.; Studer, C.; Dumitras, T.; Goldstein, T. Poison Frogs! Targeted Clean-Label Poisoning Attacks on Neural Networks. In Proceedings of the Neural Information Processing Systems, 2018.
75. Zhang, J.; Chen, B.; Cheng, X.; Binh, H.T.T.; Yu, S. PoisonGAN: Generative Poisoning Attacks Against Federated Learning in Edge Computing Systems. *IEEE Internet of Things Journal* **2021**, *8*, 3310–3322.
76. Carlini, N.; Athalye, A.; Papernot, N.; Brendel, W.; Rauber, J.; Tsipras, D.; Goodfellow, I.J.; Madry, A.; Kurakin, A. On Evaluating Adversarial Robustness. *ArXiv* **2019**, *abs/1902.06705*.

 doi:10.20944/preprints202511.1542.v1

77. Tramèr, F.; Kurakin, A.; Papernot, N.; Boneh, D.; Mcdaniel, P. Ensemble Adversarial Training: Attacks and Defenses. *ArXiv* **2017**, *abs/1705.07204*.
78. Cohen, J.M.; Rosenfeld, E.; Kolter, J.Z. Certified Adversarial Robustness via Randomized Smoothing. *ArXiv* **2019**, *abs/1902.02918*.
79. Wu, D.; Xia, S.; Wang, Y. Adversarial Weight Perturbation Helps Robust Generalization. *arXiv: Learning* **2020**.
80. Chen, T.; Liu, S.; Chang, S.; Cheng, Y.; Amini, L.; Wang, Z. Adversarial Robustness: From Self-Supervised Pre-Training to Fine-Tuning. *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)* **2020**, pp. 696–705.
81. Metzen, J.H.; Genewein, T.; Fischer, V.; Bischoff, B. On Detecting Adversarial Perturbations. *ArXiv* **2017**, *abs/1702.04267*.
82. Steinhardt, J.; Koh, P.W.; Liang, P. Certified Defenses for Data Poisoning Attacks. In Proceedings of the Neural Information Processing Systems, 2017.
83. Piet, J.; Alrashed, M.; Sitawarin, C.; Chen, S.; Wei, Z.; Sun, E.; Alomair, B.; Wagner, D. Jatmo: Prompt Injection Defense by Task-Specific Finetuning, 2024, [arXiv:cs.CR/2312.17673].
84. Doskaliuk, B.; Zimba, O.; Yessirkepov, M.; Klishch, I.; Yatsyshyn, R. Artificial intelligence in peer review: enhancing efficiency while preserving integrity. *Journal of Korean medical science* **2025**, *40*.
85. Mann, S.P.; Aboy, M.; Seah, J.J.; Lin, Z.; Luo, X.; Rodger, D.; Zohny, H.; Minssen, T.; Savulescu, J.; Earp, B.D. AI and the Future of Academic Peer Review, 2025, [arXiv:cs.CY/2509.14189].
86. Maturo, F.; Porreca, A.; Porreca, A. The risks of artificial intelligence in research: ethical and methodological challenges in the peer review process. *AI Ethics* **2025**, *5*, 5389–5396. https://doi.org/10.1007/s43681-025-00775-9.
87. Keuper, J. Prompt Injection Attacks on LLM Generated Reviews of Scientific Publications, 2025, [arXiv:cs.LG/2509.10248].
88. Verma, P. Researchers are using AI for peer reviews — and finding ways to cheat it. *The Washington Post* **2025**.
89. Media, V. Scientists reportedly hiding AI text prompts in academic papers to receive positive peer reviews, 2025. Public media reports.
90. Shi, J.; Yuan, Z.; Liu, Y.; Huang, Y.; Zhou, P.; Sun, L.; Gong, N.Z. Optimization-based Prompt Injection Attack to LLM-as-a-Judge, 2025, [arXiv:cs.CR/2403.17710].
91. Zhao, Y.; Liu, H.; Yu, D.; Kung, S.Y.; Mi, H.; Yu, D. One Token to Fool LLM-as-a-Judge, 2025, [arXiv:cs.LG/2507.08794].
92. Collu, M.G.; Salviati, U.; Confalonieri, R.; Conti, M.; Apruzzese, G. Publish to Perish: Prompt Injection Attacks on LLM-Assisted Peer Review, 2025, [arXiv:cs.CR/2508.20863].
93. Ye, R.; Pang, X.; Chai, J.; Chen, J.; Yin, Z.; Xiang, Z.; Dong, X.; Shao, J.; Chen, S. Are We There Yet? Revealing the Risks of Utilizing Large Language Models in Scholarly Peer Review, 2024, [arXiv:cs.CL/2412.01708].
94. Dong, Y.; Jiang, X.; Liu, H.; Jin, Z.; Gu, B.; Yang, M.; Li, G. Generalization or Memorization: Data Contamination and Trustworthy Evaluation for Large Language Models, 2024, [arXiv:cs.CL/2402.15938].
95. Goldblum, M.; Tsipras, D.; Xie, C.; Chen, X.; Schwarzschild, A.; Song, D.; Madry, A.; Li, B.; Goldstein, T. Dataset Security for Machine Learning: Data Poisoning, Backdoor Attacks, and Defenses, 2021, [arXiv:cs.LG/2012.10544].
96. Borgeaud, S.; Mensch, A.; Hoffmann, J.; Cai, T.; Rutherford, E.; Millican, K.; Van Den Driessche, G.B.; Lespiau, J.B.; Damoc, B.; Clark, A.; et al. Improving Language Models by Retrieving from Trillions of Tokens. In Proceedings of the Proceedings of the 39th International Conference on Machine Learning. PMLR, 17–23 Jul 2022, Vol. 162, *Proceedings of Machine Learning Research*, pp. 2206–2240.
97. Souly, A.; Rando, J.; Chapman, E.; Davies, X.; Hasircioglu, B.; Shereen, E.; Mougan, C.; Mavroudis, V.; Jones, E.; Hicks, C.; et al. Poisoning Attacks on LLMs Require a Near-constant Number of Poison Samples, 2025, [arXiv:cs.LG/2510.07192].
98. Wen, J.; Si, C.; han Chen, Y.; He, H.; Feng, S. Predicting Empirical AI Research Outcomes with Language Models, 2025, [arXiv:cs.AI/2506.00794].
99. Bereska, L.; Gavves, E. Mechanistic Interpretability for AI Safety – A Review, 2024, [arXiv:cs.AI/2404.14082].
100. Lo, L.Y.H.; Qu, H. How Good (Or Bad) Are LLMs at Detecting Misleading Visualizations?, 2024, [arXiv:cs.HC/2407.17291].
101. Tonglet, J.; Zimny, J.; Tuytelaars, T.; Gurevych, I. Is this chart lying to me? Automating the detection of misleading visualizations, 2025, [arXiv:cs.CL/2508.21675].
102. Gallegos, I.O.; Rossi, R.A.; Barrow, J.; Tanjim, M.M.; Kim, S.; Dernoncourt, F.; Yu, T.; Zhang, R.; Ahmed, N.K. Bias and Fairness in Large Language Models: A Survey, 2024, [arXiv:cs.CL/2309.00770].

 doi:10.20944/preprints202511.1542.v1

103. Liu, Y.; Deng, G.; Li, Y.; Wang, K.; Wang, Z.; Wang, X.; Zhang, T.; Liu, Y.; Wang, H.; Zheng, Y.; et al. Prompt Injection attack against LLM-integrated Applications, 2024, [arXiv:cs.CR/2306.05499].
104. Zhou, X.; Qiang, Y.; Zade, S.Z.; Khanduri, P.; Zhu, D. Hijacking Large Language Models via Adversarial In-Context Learning, 2025, [arXiv:cs.LG/2311.09948].
105. Gong, Y.; Chen, Z.; Chen, M.; Yu, F.; Lu, W.; Wang, X.; Liu, X.; Liu, J. Topic-FlipRAG: Topic-Orientated Adversarial Opinion Manipulation Attacks to Retrieval-Augmented Generation Models, 2025, [arXiv:cs.CL/2502.01386].
106. Schwinn, L.; Dobre, D.; Günnemann, S.; Gidel, G. Adversarial Attacks and Defenses in Large Language Models: Old and New Threats, 2023, [arXiv:cs.AI/2310.19737].
107. Raina, V.; Liusie, A.; Gales, M. Is LLM-as-a-Judge Robust? Investigating Universal Adversarial Attacks on Zero-shot LLM Assessment, 2024, [arXiv:cs.CL/2402.14016].
108. Guo, Y.; Guo, M.; Su, J.; Yang, Z.; Zhu, M.; Li, H.; Qiu, M.; Liu, S.S. Bias in Large Language Models: Origin, Evaluation, and Mitigation, 2024, [arXiv:cs.CL/2411.10915].
109. Navigli, R.; Conia, S.; Ross, B. Biases in Large Language Models: Origins, Inventory, and Discussion. *J. Data and Information Quality* **2023**, *15*. https://doi.org/10.1145/3597307.
110. Angrist, J.D. The Perils of Peer Effects. *Labour Economics* **2014**. https://doi.org/10.1016/j.labeco.2014.05.008.
111. Lewis, P.; Perez, E.; Piktus, A.; Petroni, F.; Karpukhin, V.; Goyal, N.; Küttler, H.; Lewis, M.; tau Yih, W.; Rocktäschel, T.; et al. Retrieval-Augmented Generation for Knowledge-Intensive NLP Tasks, 2021, [arXiv:cs.CL/2005.11401].
112. Schwarzschild, A.; Goldblum, M.; Gupta, A.; Dickerson, J.P.; Goldstein, T. Just How Toxic is Data Poisoning? A Unified Benchmark for Backdoor and Data Poisoning Attacks. In Proceedings of the Proceedings of the 38th International Conference on Machine Learning; Meila, M.; Zhang, T., Eds. PMLR, 18–24 Jul 2021, Vol. 139, *Proceedings of Machine Learning Research*, pp. 9389–9398.
113. Goldblum, M.; Tsipras, D.; Xie, C.; Chen, X.; Schwarzschild, A.; Song, D.; Madry, A.; Li, B.; Goldstein, T. Data Security for Machine Learning: Data Poisoning, Backdoor Attacks, and Defenses, 2020. https://doi.org/10.48550/arXiv.2012.10544.
114. Touvron, H.; Lavril, T.; Izacard, G.; Martinet, X.; Lachaux, M.A.; Lacroix, T.; Rozière, B.; Goyal, N.; Hambro, E.; Azhar, F.; et al. LLaMA: Open and Efficient Foundation Language Models, 2023, [arXiv:cs.CL/2302.13971].
115. Bowen, D.; Murphy, B.; Cai, W.; Khachaturov, D.; Gleave, A.; Pelrine, K. Scaling Trends for Data Poisoning in LLMs, 2025, [arXiv:cs.CR/2408.02946].
116. Liu, T.; Zhang, Y.; Feng, Z.; Yang, Z.; Xu, C.; Man, D.; Yang, W. Beyond Traditional Threats: A Persistent Backdoor Attack on Federated Learning, 2024, [arXiv:cs.CR/2404.17617].
117. Zhu, C.; Li, Y.; Rao, B.; Zhang, J.; Mao, Y.; Zhong, S. SPA: Towards More Stealth and Persistent Backdoor Attacks in Federated Learning, 2025, [arXiv:cs.CR/2506.20931].
118. Tian, Z.; Cui, L.; Liang, J.; Yu, S. A Comprehensive Survey on Poisoning Attacks and Countermeasures in Machine Learning. *ACM Comput. Surv.* **2022**, *55*. https://doi.org/10.1145/3551636.
119. Zhao, P.; Zhu, W.; Jiao, P.; Gao, D.; Wu, O. Data Poisoning in Deep Learning: A Survey, 2025, [arXiv:cs.CR/2503.22759].
120. Muñoz-González, L.; Biggio, B.; Demontis, A.; Paudice, A.; Wongrassamee, V.; Lupu, E.C.; Roli, F. Towards Poisoning of Deep Learning Algorithms with Back-gradient Optimization, 2017, [arXiv:cs.LG/1708.08689].
121. Nourani, M.; Roy, C.; Block, J.E.; Honeycutt, D.R.; Rahman, T.; Ragan, E.; Gogate, V. Anchoring Bias Affects Mental Model Formation and User Reliance in Explainable AI Systems. In Proceedings of the Proceedings of the 26th International Conference on Intelligent User Interfaces, New York, NY, USA, 2021; IUI '21, p. 340–350. https://doi.org/10.1145/3397481.3450639.
122. Shi, F.; Chen, X.; Misra, K.; Scales, N.; Dohan, D.; Chi, E.; Schärli, N.; Zhou, D. Large Language Models Can Be Easily Distracted by Irrelevant Context, 2023, [arXiv:cs.CL/2302.00093].
123. Dougrez-Lewis, J.; Akhter, M.E.; Ruggeri, F.; Löbbers, S.; He, Y.; Liakata, M. Assessing the Reasoning Capabilities of LLMs in the context of Evidence-based Claim Verification. In Proceedings of the Findings of the Association for Computational Linguistics: ACL 2025; Che, W.; Nabende, J.; Shutova, E.; Pilehvar, M.T., Eds., Vienna, Austria, 2025; pp. 20604–20628. https://doi.org/10.18653/v1/2025.findings-acl.1059.
124. Hong, R.; Zhang, H.; Pang, X.; Yu, D.; Zhang, C. A Closer Look at the Self-Verification Abilities of Large Language Models in Logical Reasoning, 2024, [arXiv:cs.AI/2311.07954].
125. OWASP Foundation. OWASP Top 10 for Large Language Model Applications, 2023. Accessed in 2025. See LLM01: Prompt Injection. URL: https://owasp.org/www-project-top-10-for-large-language-model-applications/.

 doi:10.20944/preprints202511.1542.v1

126. Liang, W.; Zhang, Y.; Cao, H.; Wang, B.; Ding, D.; Yang, X.; Vodrahalli, K.; He, S.; Smith, D.; Yin, Y.; et al. Can large language models provide useful feedback on research papers? A large-scale empirical analysis, 2023, [arXiv:cs.LG/2310.01783].
127. Zhou, Z.; Li, Z.; Zhang, J.; Zhang, Y.; Wang, K.; Liu, Y.; Guo, Q. CORBA: Contagious Recursive Blocking Attacks on Multi-Agent Systems Based on Large Language Models, 2025, [arXiv:cs.CL/2502.14529].
128. Zhu, S.; Zhang, R.; An, B.; Wu, G.; Barrow, J.; Huang, F.; Sun, T. AutoDAN: Automatic and Interpretable Adversarial Attacks on Large Language Models, 2024.
129. Zizzo, G.; Cornacchia, G.; Fraser, K.; Hameed, M.Z.; Rawat, A.; Buesser, B.; Purcell, M.; Chen, P.Y.; Sattigeri, P.; Varshney, K. Adversarial Prompt Evaluation: Systematic Benchmarking of Guardrails Against Prompt Input Attacks on LLMs, 2025, [arXiv:cs.CR/2502.15427].
130. Malmqvist, L. Sycophancy in Large Language Models: Causes and Mitigations, 2024, [arXiv:cs.CL/2411.15287].
131. Bozdag, N.B.; Mehri, S.; Tur, G.; Hakkani-Tür, D. Persuade Me if You Can: A Framework for Evaluating Persuasion Effectiveness and Susceptibility Among Large Language Models, 2025, [arXiv:cs.CL/2503.01829].
132. Salvi, F.; Horta Ribeiro, M.; Gallotti, R.; West, R. On the conversational persuasiveness of GPT-4. *9*, 1645–1653. https://doi.org/10.1038/s41562-025-02194-6.
133. Liu, Y.; Yang, K.; Liu, Y.; Drew, M.G.B. The Shackles of Peer Review: Unveiling the Flaws in the Ivory Tower. 2023.
134. Nisbett, R.E.; Wilson, T.D. The halo effect: Evidence for unconscious alteration of judgments. *Journal of Personality and Social Psychology* **1977**, *35*, 250–256.
135. Zhang, J.; Zhang, H.; Deng, Z.; Roth, D. Investigating Fairness Disparities in Peer Review: A Language Model Enhanced Approach, 2022, [arXiv:cs.CY/2211.06398].
136. Fox, C.W.; Meyer, J.A.; Aimé, E. Double-blind peer review affects reviewer ratings and editor decisions at an ecology journal. *Functional Ecology* **2023**.
137. Sun, M.; Danfa, J.B.; Teplitskiy, M. Does double-blind peer review reduce bias? Evidence from a top computer science conference. *J. Assoc. Inf. Sci. Technol.* **2021**, *73*, 811 – 819.
138. Jin, Y.; Zhao, Q.; Wang, Y.; Chen, H.; Zhu, K.; Xiao, Y.; Wang, J. AgentReview: Exploring Peer Review Dynamics with LLM Agents. In Proceedings of the Conference on Empirical Methods in Natural Language Processing, 2024.
139. Verharen, J.P.H. ChatGPT identifies gender disparities in scientific peer review. *eLife* **2023**, *12*.
140. Hosseini, M.; Horbach, S.P. Fighting reviewer fatigue or amplifying bias? Considerations and recommendations for use of ChatGPT and other large language models in scholarly peer review. *Research Integrity and Peer Review* **2023**, *8*.
141. Soneji, A.; Kokulu, F.B.; Rubio-Medrano, C.E.; Bao, T.; Wang, R.; Shoshitaishvili, Y.; Doupé, A. "Flawed, but like democracy we don't have a better system": The Experts' Insights on the Peer Review Process of Evaluating Security Papers. *2022 IEEE Symposium on Security and Privacy (SP)* **2022**, pp. 1845–1862.
142. Schramowski, P.; Turan, C.; Andersen, N.; Rothkopf, C.A.; Kersting, K. Large pre-trained language models contain human-like biases of what is right and wrong to do. *Nature Machine Intelligence* **2021**, *4*, 258 – 268.
143. Li, H.; Ji, Y.; Lyu, C.; Zhang, C. Blacklight: Scalable Defense for Neural Networks against Query-Based Black-Box Attacks. In Proceedings of the 31st USENIX Security Symposium (USENIX Security 22), 2022.
144. Koo, R.; Lee, M.; Raheja, V.; Park, J.I.; Kim, Z.M.; Kang, D. Benchmarking Cognitive Biases in Large Language Models as Evaluators, 2024, [arXiv:cs.CL/2309.17012].
145. Bartos, O.J.; Wehr, P. *Using Conflict Theory*; Cambridge University Press: Cambridge, 2002.