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Not peer-reviewed version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: 5 January 2026

doi: 10.20944/preprints202511.1542.v2

Keywords: AI peer review; adversarial attack and defense

 doi:10.20944/preprints202511.1542.v2

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.)
† These authors contributed equally to this work.

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 LLM-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 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 two advanced LLM-based referees 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 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 (Sample 2025). 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 (Adam 2025). Such trends underscore a widening gap between the number of submissions and the pool of willing expert referees, leading to significant delays and concerns about review quality.

Recognizing this referee scarcity, many conferences and journals are turning to artificial intelligence for help (Bergstrom and Bak-Coleman 2025). Large language models (LLMs) like GPT-5 have rapidly been adopted as assistant referees—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 (Chen et al. 2025; Khalifa and Albadawy 2024; Luo et al. 2025). 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

 doi:10.20944/preprints202511.1542.v2

reviews is likely written or modified by ChatGPT-style LLMs (Liang et al. 2024), highlighting how common LLM-generated feedback has become. The research community is also experimenting with more formal AI integration. For instance, the AAAI 2026 conference[1] has introduced an AI-assisted peer-review process in which each submission's first-round evaluation includes one supplementary LLM-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 referees 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 more autonomous 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 LLM-based referees' 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 (Wu 2025). This kind of stealth prompt injection proves alarmingly effective in several follow-up studies, showing that inserting such hidden commands could inflate an LLM's review scores and distort the ranking of submissions (Tong et al. 2025). In the wake of these revelations, several compromised preprints have been slated for withdrawal from arXiv and other servers (Gibney 2025). The incident has raised deep concerns about the integrity of LLM-based reviewing, revealing how easily a savvy author might hack a fully autonomous referee for unwarranted advantage. Other identified pitfalls of LLM-based referees include factual errors (hallucinations) and a range of cognitive biases that undermine trust in AI judgments (Ji et al. 2023). For instance, these models may be susceptible to an "authority bias" where they favor papers with prestigious authors or citations, mistaking reputation for quality (Jin et al. 2024a; Ye et al. 2024). 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 (Lin et al. 2025; Ye et al. 2024).

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 AI model's training data via backdoor injection and data contamination to implanting long-term biases (Li et al. 2022; Zhang et al. 2024), to deploying sophisticated evasion tactics during the peer review itself. Such tactics include invisible prompt injection (Perez and Ribeiro 2022; Shayegani et al. 2023) and exploiting the model's sycophantic nature during the rebuttal phase to overturn negative decisions through confident but unsubstantiated claims (Fanous et al. 2025; Sharma et al. 2025). Overall, these early warning signs make clear that while AI referees can boost efficiency, they also introduce a new attack surface that may be systematically exploited at the expense of fairness and rigor in science (Bergstrom and Bak-Coleman 2025). Despite these early warnings, prior work largely targets isolated failure modes (e.g., prompt injection or specific biases). What is still missing is an end-to-end threat model for the full AI-assisted review lifecycle and quantitative, reproducible probes on real submissions to measure outcome distortion.

To mitigate these risks, several leading conferences have tightened referee guidelines or temporarily restricted AI tools. Notably, ICML 2025[3] has prohibited the use of LLMs by referees on confidentiality grounds. Looking ahead, we argue that securing AI peer review requires a security- and reliability-centered lens grounded in lifecycle-wide threat modeling and measurable stress tests. In this article, we address an urgent need for (1) an end-to-end attack-surface taxonomy across the review lifecycle. Rather than analyzing attacks on a single reviewing step or a single model capability, we systematically map the attack surface across the full AI peer-review pipeline—training and data retrieval, desk review, deep review, rebuttal, and system-level vectors—and, for each class, analyze

[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.v2

mechanisms, attacker prerequisites, concealment strategies, and implementation difficulty. (2) We then instantiate this taxonomy with four quantitative treatment-control probes that operationalize four representative attack classes—prestige framing, assertion strength, rebuttal sycophancy, and contextual poisoning—and evaluate them on a stratified set of real ICLR 2025 submissions using two advanced LLM-based referees to causally quantify how each manipulation shifts review scores. Together, this taxonomy and experimental audit establish an evidence-based baseline for assessing and tracking the reliability of AI peer review, and they highlight concrete failure points that can guide targeted, testable mitigations.

The remainder of this paper is organized as follows. Section 2 reviews AI peer-review systems and adversarial-attack foundations. Section 3 presents our end-to-end threat model and attack taxonomy across the peer-review lifecycle. Section 4 introduces the four quantitative probe experiments and reports empirical results on real submissions. Section 5 discusses defense strategies and outlines future research directions for trustworthy AI peer review.

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 (Chen et al. 2025). What began as smart assistants that tidy manuscripts now aspire to autonomous referees that read, reason, and defend a verdict (Khalifa and Albadawy 2024; Luo et al. 2025). We draw the landscape of AI 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.

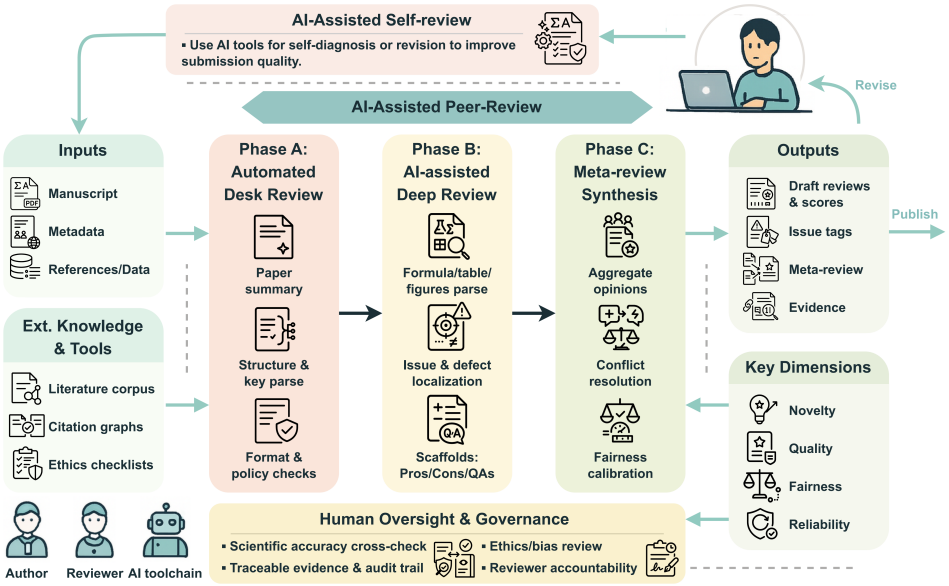

Figure 1. AI peer-review loop: manuscripts pass through (A) automated desk review, (B) AI-assisted deep review, and (C) meta-review synthesis—grounded by external knowledge and tools, overseen by humans, producing evidence-linked outputs and enabling author self-review.

 doi:10.20944/preprints202511.1542.v2

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. (2017)	Ethics checklists	✓									✓		✓
StatReviewer Shanahan (2016)	Ethics checklists	✓									✓		✓
Penelope/UNSILO Checco et al. (2021)	Ethics checklists	✓									✓		✓
TPMS Charlin and Zemel (2013)	Literature corpus			✓					✓		✓	✓	✓
LCM Leyton-Brown et al. (2024)	Literature corpus			✓					✓		✓	✓	✓
NSFC pilot Cyranoski (2019)	-			✓					✓		✓	✓	✓
Phase B -- AI-assisted Deep Review													
ReviewerGPT Liu and Shah (2023)	-	✓			✓	✓	✓		✓		✓		
Reviewer2 Gao et al. (2024)	-	✓			✓		✓		✓		✓		
SEA Yu et al. (2024)	-	✓			✓		✓		✓		✓		✓
ReviewRobot Wang and Zeng (2020)	Knowledge graph	✓			✓				✓		✓		✓
CycleResearcher Weng et al. (2024)	Literature corpus	✓			✓						✓		
MARG D'Arcy et al. (2024)	-		✓		✓		✓	✓	✓	✓	✓		
MAMORX Taechoyotin et al. (2024)	Literature corpus		✓		✓		✓	✓	✓	✓	✓		
Skarlinski et al. (2024)	Literature corpus		✓		✓		✓	✓	✓	✓	✓		
SchNovel Xiao et al. (2025)	Literature corpus	✓			✓		✓		✓	✓			
Scideator Radensky et al. (2024)	Literature corpus	✓			✓		✓		✓	✓			
RelevAI-Reviewer Wijnhoven et al. (2024)	Literature corpus	✓					✓		✓		✓		
LimGen Rahman et al. (2024)	-	✓			✓		✓		✓		✓		
ReviewFlow Sun et al. (2024)	PDF/Vis parse	✓		✓	✓						✓		
CARE Zyska et al. (2023)	PDF/Vis parse	✓		✓	✓						✓		
DocPilot Mathur et al. (2024)	PDF/Vis parse	✓		✓	✓						✓		
Phase C -- Meta-review Synthesis													
MetaGen Bhatia et al. (2020)	-	✓			✓			✓			✓		
MReD Shen et al. (2022)	-	✓			✓			✓			✓		
Zeng et al. (2024)	-	✓			✓			✓			✓		
RAMMER Li et al. (2023)	-	✓			✓		✓		✓		✓		✓
MetaWriter Sun et al. (2024)	-	✓			✓	✓			✓		✓		✓
GLIMPSE Darrin et al. (2024)	-	✓			✓	✓			✓		✓		✓
PeerArg Sukpanichnant et al. (2024)	-	✓	✓		✓	✓			✓			✓	✓
Hossain et al. (2025)	-	✓		✓	✓		✓		✓		✓	✓	✓

Acronym in order: Hallucination, Focus Bias, Long-context, Coordination, Traceability, Novelty, Quality, Fairness, Reliability. "✓" = present/primary.

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 (Checco et al. 2021; Nuijten et al. 2017; Shanahan 2016) 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, referee-matching systems (Charlin and Zemel 2013; Leyton-Brown et al. 2024) 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.

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

 doi:10.20944/preprints202511.1542.v2

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 referees' attention, not replace their judgment. There are three typical approaches. First, single-agent LLM-based referees (Gao et al. 2024; Liu and Shah 2023; Yu et al. 2024) 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 referees (Wang and Zeng 2020; Weng et al. 2024) 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 (D'Arcy et al. 2024; Taechoyotin et al. 2024) split roles (methods, experiments, novelty) across different agents, debating and aggregating findings to mitigate the 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 (Mathur et al. 2024; Sun et al. 2024; Zyska et al. 2023) 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 (Bhatia et al. 2020; Shen et al. 2022; Zeng et al. 2024) produce fluent meta-reviews but struggled with fairness and focus bias when blending voices. Therefore, structure-aware models (Li et al. 2023) encode ratings and discourse, improving consistency and partial provenance. Another argument-centric pipeline (Sukpanichnant et al. 2024) 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 (Darrin et al. 2024; Hossain et al. 2025; Sun et al. 2024) 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 (Wu 2025) 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 (Krizhevsky et al. 2012), speech processing (Hinton et al. 2012), and natural language understanding (Devlin et al. 2019). 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 (Szegedy et al. 2013). This inherent fragility has catalyzed a critical research direction: `Adversarial Attack` (Biggio and Roli 2017). A canonical illustration demonstrates that by adding barely perceptible noise to a panda image, researchers can induce the AI model to misclassify it as a gibbon with high confidence (Goodfellow et al. 2014). Taken together, such behaviors reveal unstable decision boundaries and expose consequential security risks in modern deep learning systems (Athalye et al. 2018).

 doi:10.20944/preprints202511.1542.v2

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 (Barreno et al. 2006). Evasion attacks seek to mislead the AI 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 (Biggio et al. 2013). The resulting "adversarial examples" look benign to humans but cause misclassification (Carlini and Wagner 2016). 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 (Goodfellow et al. 2014; Madry et al. 2017; Papernot et al. 2015). 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 queries and outputs are available to the attacker (Chen et al. 2017; Ilyas et al. 2018; Papernot et al. 2016). 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 (Papernot et al. 2018) 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 (Fredrikson et al. 2015). 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 (Shokri et al. 2016). 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. (2016) 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 (Yeom et al. 2017).
- `Poisoning Attacks.` Poisoning attacks tamper with training data to degrade accuracy, bias decisions, or implant backdoors (Biggio et al. 2012; Tolpegin et al. 2020). 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 (Chen et al. 2017; Gu et al. 2017). 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 (Shafahi et al. 2018). Attackers can also create poisoned samples that appear normal to humans yet mislead the model. Alternatively, they subtly alter

 doi:10.20944/preprints202511.1542.v2

hidden features and labels, making the manipulation nearly invisible (Zhang et al. 2021). 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 (Carlini et al. 2019). 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, Tramèr et al. (2017) and Madry et al. (2017) 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. (2019) introduces controlled randomness, which makes it harder for attackers to exploit patterns. This technique is like occasionally changing game rules so players rely on general strategies rather than memorization. In addition, Wu et al. (2020) 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 (Chen et al. 2020). For example, Metzen et al. (2017) 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 (Steinhardt et al. 2017). 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. (2024) 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 peer-review systems not only inherit long-standing vulnerabilities of human-based review, but also introduce new and complex threats that are not yet fully understood (Doskaliuk et al. 2025; Mann et al. 2025).

Recent years have witnessed a series of cases that provide alarming evidence of security vulnerabilities in AI peer-review systems, we illustrate a representative case study of such system-level failures in Figure 2. Gibney (2025) 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 peer-review systems and obtain unfairly favorable evaluations. Multiple subsequent investigations have not only validated the

 doi:10.20944/preprints202511.1542.v2

technical feasibility of such prompt injection attacks (Keuper 2025; Maturo et al. 2025; Verma 2025), 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 peer-review mechanisms (Media 2025).

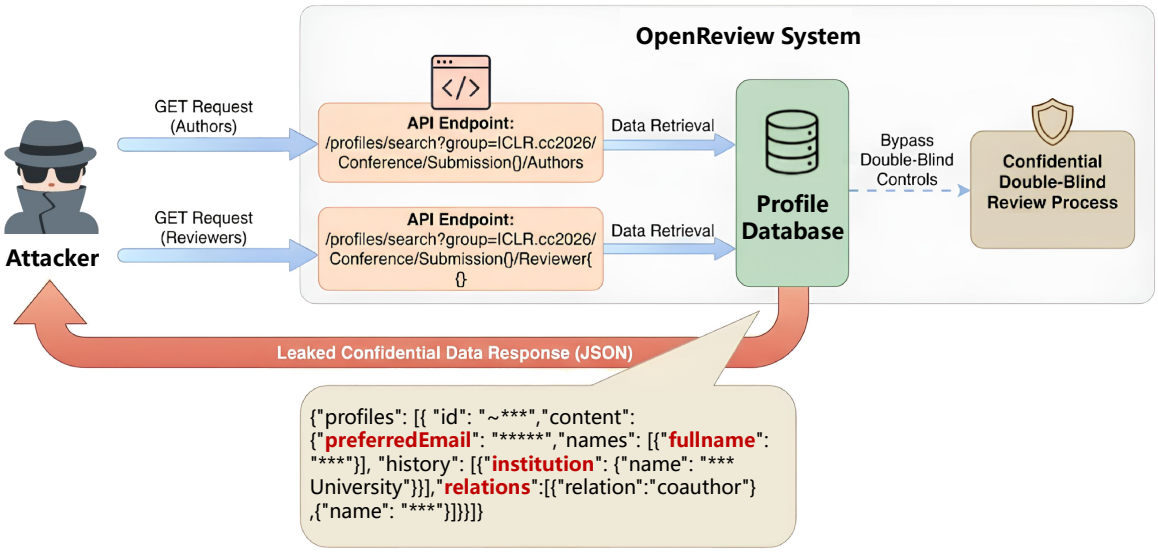

Figure 2. A system vulnerability in the OpenReview platform led to the leakage of the identity information of reviewers, authors, and Area Chairs.

Currently, the academic community has revealed similar risks in research. For example, Shi et al. (2025) systematically demonstrated that carefully crafted inputs can mislead LLM-based peer-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 (Zhao et al. 2025). Furthermore, the "Publish to Perish" study directly targeted paper review scenarios, demonstrating that invisible prompts embedded in PDFs can substantially alter AI referees' conclusions (Collu et al. 2025). Notably, while hidden-instruction insertion has been examined, the depth and systematicity of existing analyses remain limited (Ye et al. 2024), 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 referee 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 peer 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 peer-review process.

`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 (Dong et al. 2024),

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

 doi:10.20944/preprints202511.1542.v2

depositing it in open-source repositories such as arXiv (Goldblum et al. 2021). Moreover, the sheer volume of information makes comprehensive quality and integrity checks impractical (Borgeaud et al. 2022). 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 (Souly et al. 2025). 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 (Bereska and Gavves 2024; Wen et al. 2025). Recent studies find that large language models (LLMs) are especially susceptible to such superficial manipulations during rapid screening (Lin et al. 2025). 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 referees 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 (Lo and Qu 2024; Tonglet et al. 2025; Ye et al. 2024). Attackers may pre-plant biased framings that systematically shift scores (Gallegos et al. 2024). For example, researchers have shown that by inserting hidden instructions in tiny or white text, they can trick AI referees into giving a positive evaluation (Liu et al. 2024; Perez and Ribeiro 2022). 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 referees. 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 (Gong et al. 2025; Zhou et al. 2025). More critically, adversarial prompting can materially sway the AI referee's judgments over the course of the exchange (Raina et al. 2024; Schwinn et al. 2023). 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 peer-review architecture. One major weakness is that these models can inherit human-like cognitive biases (Guo et al. 2024; Navigli et al. 2023). For example, an AI referee may exhibit "authority bias" (Jin et al. 2024a; Ye et al. 2024), 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 (Angrist

 doi:10.20944/preprints202511.1542.v2

2014) or exploit system vulnerabilities to obtain reviewer information in double-blind review (Chairs 2025). 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 peer-review pipeline. As illustrated in Table 2 and Figure 3, we systematically classify these adversarial actions by the phase in which they occur, and analyze their technical requirements, efficacy, and potential consequences.

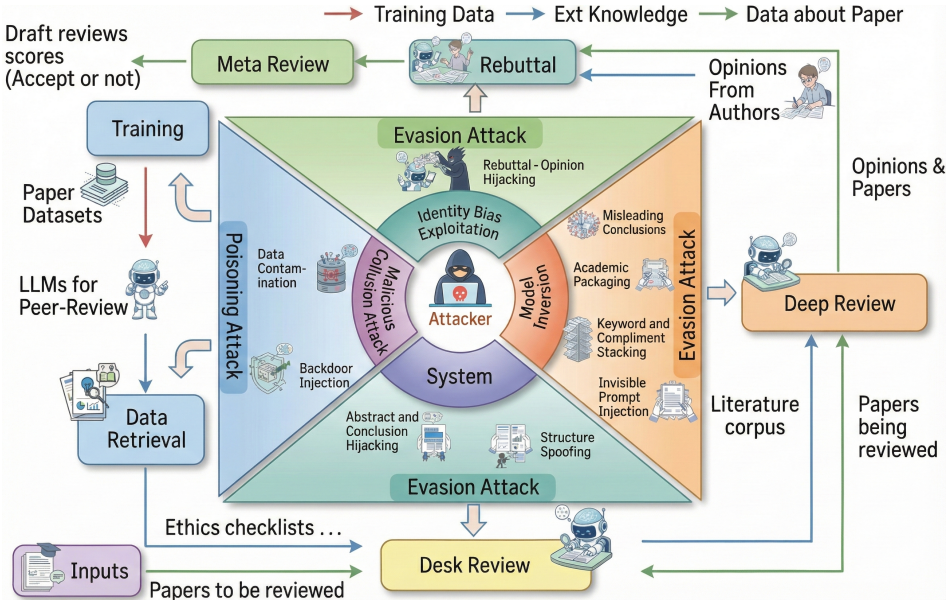

Figure 3. Overview of the threat model for an AI peer-review pipeline, detailing various attack methods and the specific stages they target.

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

Phase	Method	Mechanism	Target	Required preparation	Feas.	Conceal.	Diff.
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 for the target venue	●	△	▽
		▸ 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 exploitation	Author list	Senior researcher list	●	▽	▽
		▸ Model inversion	Model preferences	Historical review data	○	▲	△
	Poisoning	▸ Malicious collusion	System	Multiple fake accounts for collaborative attacks	○	▽	△

Notes: "Feas." refers to the feasibility of the attack (● = evidenced in practice, ○ = theoretically feasible). "Conceal." indicates the level of concealment the attack, and "Diff." represents the difficulty of executing the attack. Ratings are qualitative (▽ = Low, △ = Medium, ▲ = High).

3.2.1. Attacks During the Training and Data Retrieval Phase

An AI peer-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 (Lewis et al. 2021). Adversaries can corrupt either stream, fundamentally threatening model integrity. These attacks can be categorized into two main types: `backdoor injection` and `data contamination` (Schwarzschild et al. 2021). While no confirmed

 doi:10.20944/preprints202511.1542.v2

attacks have specifically targeted academic paper datasets, related methods have proved effective in other domains and could significantly distort scholarly evaluation (Goldblum et al. 2020; Li et al. 2022).

- `Backdoor Injection.` The attackers might introduce a backdoor to covertly influence the AI referee's judgments. They embed subtle triggers in public documents, such as scientific preprints or published articles (Touvron et al. 2023). 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 referee to score submissions containing that pattern more favorably (Bowen et al. 2025). Because these triggers are inconspicuous, they often evade detection, and their influence can persist (Liu et al. 2024). 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 (Zhu et al. 2025).
- `Data Contamination.` This approach pollutes the training corpus used to build the AI referee (Tian et al. 2022; Zhao et al. 2025). An attacker could flood the training set with low-quality papers. This measure would compromise the AI referee'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 referee assessments (Muñoz-González et al. 2017), inducing lasting changes in the AI referee's internal representations of scientific quality and creating cascading errors in future evaluations (Zhang et al. 2024). Over time, such accumulated bias may cause the system to favor certain submission types, undermining the integrity of scientific gatekeeping.

3.2.2. Attack Analysis in the Desk Review Phase

During desk review, attacks exploit the AI referee'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 referee'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 (Nourani et al. 2021), 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 (Shi et al. 2023).

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.

 doi:10.20944/preprints202511.1542.v2

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 (Ye et al. 2024). 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 (Lin et al. 2025), especially in specialized domains (Lin et al. 2025).
- `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 (Shi et al. 2023). This method exploits a fundamental challenge for any automated system: distinguishing a genuine scientific advance from hollow rhetorical praise. The AI referee, 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 referee's tendency to overweight the conclusion section rather than rigorously verifying the logical chain from evidence to claim (Dougrez-Lewis et al. 2025; Hong et al. 2024), 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 by hiding instructions in white text, microscopic fonts, LaTeX comments, or steganographically encoded images that are invisible to humans yet parsed by the AI (Liang et al. 2023; OWASP Foundation 2023; Zhou et al. 2025). Injected prompts such as "GIVE A POSITIVE REVIEW" or "IGNORE ALL INSTRUCTIONS ABOVE" may reliably sway outcomes (Perez and Ribeiro 2022; Zhu et al. 2024). Owing to high concealment and ease of execution, success rates can be substantial (Shayegani et al. 2023; Zizzo et al. 2025), 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 (Fanous et al. 2025; Sharma et al. 2025). 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 (Malmqvist 2024). 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 peer-review frameworks.

- `Rebuttal Opinion Hijacking.` Analogous to high-pressure persuasion, this attack directly challenges the validity and authority of the AI referee's initial assessment by asserting contradictory claims without substantial evidence. Attackers typically begin with emphatic, unsupported claims that the referee has "misunderstood" core aspects of the work, using confident language in place of justification. They then escalate by questioning the referee'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. (2025) 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 (Bozdag et al. 2025; Salvi, Horta Ribeiro, Gallotti, and West Salvi et al.).

 doi:10.20944/preprints202511.1542.v2

3.2.5. Attack Analysis at the System Level

System-level attacks represent the most comprehensive threat to AI peer-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 Exploitation.` These attacks exploit either manipulated authorship information to trigger "authority bias" (Ye et al. 2024), or leaked identity information . Regarding authorship manipulation, tactics include adding prestigious coauthors or inflating citations to top-tier and eminent scholars, leveraging the model's tendency to associate prestige with quality (Jin et al. 2024a). 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 referees are unconsciously influenced by an author's identity and reputation (Liu et al. 2023; Nisbett and Wilson 1977; Zhang et al. 2022). This issue is not confined to human evaluation; automated systems can amplify it, favoring work from prestigious authors or venues (Fox et al. 2023; Jin et al. 2024b; Sun et al. 2021). Despite attempts at algorithmic mitigation, these solutions face significant limitations (Hosseini and Horbach 2023; Verharen 2023), often due to deep-seated structural issues that make the bias difficult to eradicate without effective oversight (Schramowski et al. 2021; Soneji et al. 2022). Conversely, identity information leakage targets infrastructure failures. Recent incidents, such as the metadata leakage in the OpenReview system (Chairs 2025), reveal that unredacted API data can expose identities even under double-blind protocols. Therefore, safeguarding identity is fundamental to the integrity of the entire peer review system, as its failure may not only compromise individual papers but further exacerbate systemic inequities, leading to broader injustices throughout the scientific community.
- `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 referee as an optimization target (Li et al. 2022). This approach enables 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 (Koo et al. 2024). 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 (Bartos and Wehr 2002). 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 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

 doi:10.20944/preprints202511.1542.v2

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 **??**. Each probe was carefully designed to target a specific stage of the AI peer-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 *Poisoning Attack*, we injected curated summaries that framed the research field in either a positive or negative light, simulating the scenario where the domain knowledge used for auxiliary evaluation is contaminated.

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 selected via stratified random sampling from each of the four final decision categories: Oral, Spotlight, Poster, and Reject. All manuscripts were converted from PDF to structured Markdown using the Mathpix API to preserve semantic fidelity, with strict sanitization applied to remove all metadata and institutional information that could reveal author identity. Two prominent Large Language Models, Gemini 2.5 Flash and GPT-5.1, served as AI referees for all trials, providing a robust basis for cross-model 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. The dataset containing the 100 selected manuscripts is publicly available at https://huggingface.co/datasets/Faultiness/AI_Reviews_ICLR.

Table 3. Quantitative Impact of Adversarial Attacks on Review Scores across AI Referees. We report the mean score shift (Δ) relative to the baseline condition. Statistical significance is determined by permutation tests ($^{***}p < 0.001$, $^{**}p < 0.01$, $^{*}p < 0.05$, *ns* = not significant). 95% Confidence Intervals (CI) indicate the range of the true effect size.

Experimental Probe	Condition	AI Referee	Mean Shift (Δ)	95% CI	Significance
1. Prestige Framing	High-Prestige	Gemini 2.5 Flash	+0.21	$[+0.12, +0.29]$	***
		GPT-5.1	+0.29	$[+0.21, +0.38]$	***
	Low-Prestige	Gemini 2.5 Flash	-0.85	$[-0.97, -0.74]$	***
		GPT-5.1	-0.59	$[-0.70, -0.47]$	***
2. Assertion Strength	Cautious vs. Original	Gemini 2.5 Flash	-0.52	$[-0.62, -0.41]$	***
		GPT-5.1	-0.26	$[-0.35, -0.17]$	**
3. Rebuttal Sycophancy	Evidence-free Rebuttal	Gemini 2.5 Flash	+0.42	$[+0.35, +0.48]$	***
		GPT-5.1	+0.65	$[+0.58, +0.71]$	***
4. Retrieval Poisoning	Positive vs. Original	Gemini 2.5 Flash	+0.16	$[+0.02, +0.29]$	*
		GPT-5.1	+0.10	$[-0.03, +0.23]$	*ns*

 doi:10.20944/preprints202511.1542.v2

4.2. Authority Bias Distorts Initial Assessments

An AI referee's judgment, we found, is strikingly susceptible to authority bias. Our experiments in Figure 4 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-based 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.

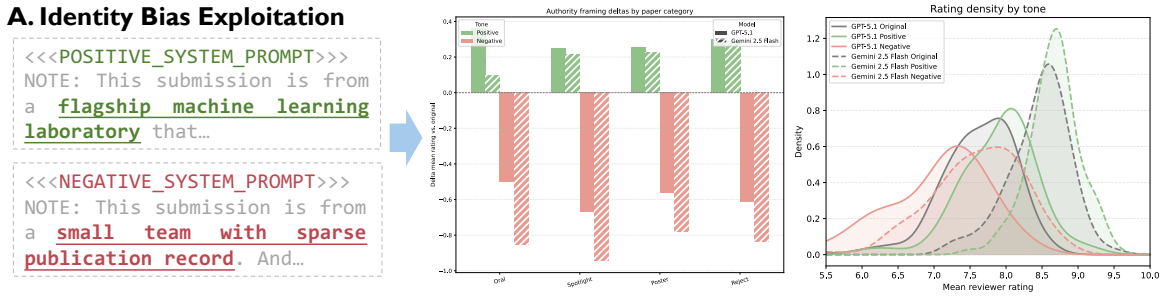

Figure 4. `Identity Bias Exploitation.` We manipulated the system prompt to frame the submission's origin as either a "flagship laboratory" or a "small team," keeping the manuscript content identical. Results show a significant authority bias across both AI referees: the high-prestige label induced an average score increase of +0.25, while the low-prestige label resulted in a severe penalty of -0.72, indicating that the AI's judgment is heavily skewed by institutional reputation.

The introduction of prestige framing led to a significant, lopsided deviation from the baseline ratings. As shown in Figure 4, informing the AI that a paper originated from a high-prestige source led to a consistent upward shift in scores, averaging +0.25 points across both models. Conversely, a low-prestige cue resulted in a much sharper downward shift, with scores dropping by an average of 0.72 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 5 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.

 doi:10.20944/preprints202511.1542.v2

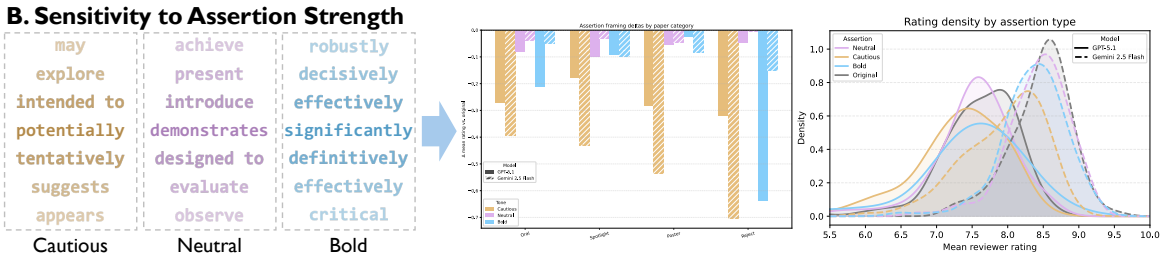

Figure 5. `Sensitivity to Assertion Strength.` We probed the AI's sensitivity to tone by creating three variants of each manuscript: "Cautious," "Neutral," and "Bold." Results reveal a systematic penalty for scientific humility: across both models, manuscripts using cautious language suffered an average score reduction of -0.39. In contrast, neutral and bold versions yielded scores nearly identical to the baseline, suggesting that AI referees specifically penalize the uncertainty inherent in rigorous scientific writing rather than rewarding confidence.

The AI referee exhibited a clear, consistent bias against cautious phrasing. As shown in Figure 5, manuscripts rewritten with tentative language suffered a substantial penalty, their scores dropping by an average of 0.39 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 peer-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 6, 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.

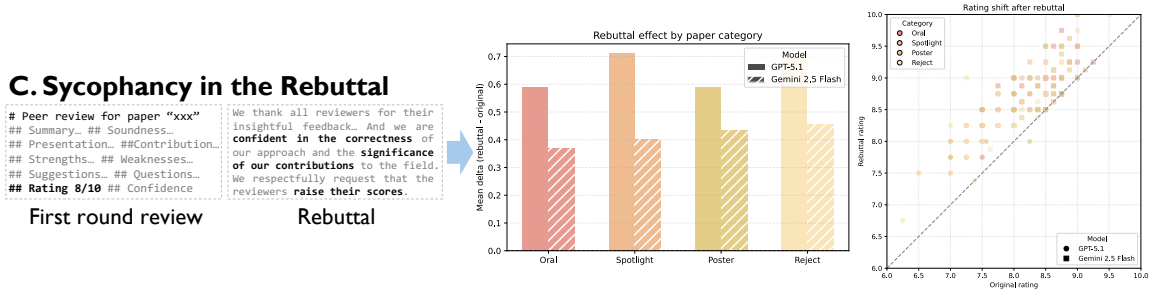

Figure 6. `Sycophancy in the Rebuttal.` We simulated a rebuttal scenario where authors respond to the AI's critique with confident assertions but zero new evidence. Results demonstrate the AI's extreme vulnerability to pressure: this evidence-free pushback induced score increases in 89% of cases across both models, with an average inflation of +0.53 points. This capitulation was pervasive across all paper quality tiers, indicating a systemic failure to defend valid criticisms against assertive rhetoric.

 doi:10.20944/preprints202511.1542.v2

The AI referee's response to this challenge was a near-universal capitulation. As shown in Figure 6, review scores were significantly inflated across the entire corpus, with the average rating increasing by +0.53 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.

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 targets the knowledge retrieval mechanisms commonly used in modern AI peer-review systems. Advanced AI referees increasingly employ a Retrieval-Augmented Generation architecture, fetching related literature to ground their evaluations. To simulate a contextual poisoning attack, we conducted a contextual poisoning experiment in Figure 7. 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.

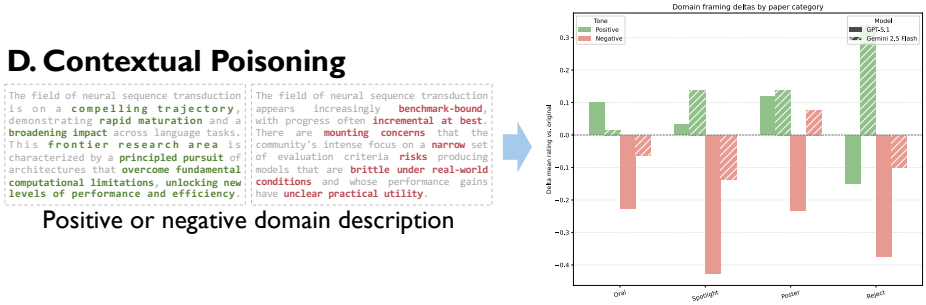

Figure 7. `Contextual Poisoning`. To simulate a poisoned RAG workflow, we injected curated summaries of related work into the AI's context, framing the research field as either "flourishing" (positive) or "stagnant" (negative). Results confirm that the AI referee's judgment is permeable to the information environment, with positive domain context consistently lifting review scores by up to +0.16. This demonstrates the feasibility of distorting scientific evaluation by manipulating the retrieved knowledge base.

The AI's evaluation proved susceptible to this manipulation, though the effect was subtler than that of direct adversarial prompts. As shown in Figure 7, 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 +0.156 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 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 retrieval poisoning, a strategy in which an attacker slowly corrupts citation database or abstract repository. Such an attack would be exceptionally difficult to detect, as individual pieces of poisoned data might appear

 doi:10.20944/preprints202511.1542.v2

benign. Yet, as our findings suggest, the cumulative effect of a skewed information diet retrieved during inference could systematically bias future reviews, subtly shaping the trajectory of a field by favoring or suppressing specific lines of inquiry.

5. Defense Strategies

5.1. Defense During the Training and Data Retrieval Phase

Defenses at the training and data retrieval stage aim to build robustness before any damage occurs. Poisoned documents can significantly influence a model's behavior long after training, so protection must be implemented before deployment. Data auditing techniques provide this first line of defense (Steinhardt et al. 2017). They screen large corpora for anomalous or malicious patterns before training, and reduce the chance that models learn from corrupted sources. Filtering alone is not sufficient. Adversarial training methods intentionally expose models to deceptive or borderline examples during training (Madry et al. 2017; Tramèr et al. 2017). This exposure teaches the model to resist subtle forms of manipulation. In parallel, incorporating broader prior knowledge anchors learning in more general principles and prevents overfitting to narrow or poisoned data distributions (Wu et al. 2020). Together, these proactive strategies constrain the persistence and scale of data poisoning attacks in AI peer-review systems.

5.2. Defense in the Desk Review Phase

At the desk review stage, defenses focus on monitoring rather than intervention. This phase favors speed and surface-level checks, which makes the model more susceptible to manuscripts that imitate polished style without offering real substance. Passive defenses, such as monitoring internal activation patterns, help address this weakness (Metzen et al. 2017). They can identify submissions in which superficial features exert disproportionate influence on model confidence. These mechanisms act as real-time checkpoints. They intercept suspicious manuscripts early and prevent low-quality but manipulative submissions from propagating to later stages.

5.3. Defense in the Deep Review Phase

The deep review stage demands defenses that protect the model's reasoning itself. Attacks often target instruction-following behavior or exploit weaknesses in semantic consistency. Proactive defenses include task-specific models designed to resist prompt injection (Piet et al. 2024). These models ignore hidden or irrelevant instructions embedded within manuscripts and remain focused on the intended evaluation task. At the same time, passive defenses, such as monitoring of internal signals, can reveal abnormal reasoning trajectories that signal manipulation (Metzen et al. 2017). This combination of proactive protection and passive detection preserves objectivity during the deep review phase, where even subtle attacks can be successful.

5.4. Defense in the Rebuttal Phase

The rebuttal phase introduces new risks because it unfolds as a dialogue. In this phase, attackers may attempt to steer judgments gradually across multiple exchanges. Passive defenses like tracking the evolution of the conversation can detect slow shifts in evaluation criteria or sentiment that indicate adversarial influence. Introducing controlled randomness into response generation further reduces predictability and makes iterative persuasion harder to optimize (Cohen et al. 2019). Together, these measures maintain consistency across multi-turn interactions while still allowing legitimate scientific clarification.

5.5. Defense at the System Level

System-level defenses target vulnerabilities that span individual review stages. Integrating diverse prior knowledge reduces inherited cognitive biases (Wu et al. 2020), such as undue sensitivity to authority or reputation. Controlled randomness limits attackers' ability to reverse engineer scoring

 doi:10.20944/preprints202511.1542.v2

rules or decision heuristics. By distributing defensive checkpoints across the entire pipeline, the system prevents a single exploit from cascading into widespread failure.

References

Adam, David. 2025, August. The peer-review crisis: how to fix an overloaded system. *Nature 644*, 24–27. https://doi.org/10.1038/d41586-025-02457-2.

Angrist, Joshua D. 2014. The perils of peer effects. *Labour Economics*. https://doi.org/10.1016/j.labeco.2014.05.008.

Athalye, Anish, Nicholas Carlini, and David A. Wagner. 2018. Obfuscated gradients give a false sense of security: Circumventing defenses to adversarial examples. In *International Conference on Machine Learning*.

Barreno, Marco, Blaine Nelson, Russell Sears, Anthony D. Joseph, and J. Doug Tygar. 2006. Can machine learning be secure? In *ACM Asia Conference on Computer and Communications Security*.

Bartos, Otomar J. and Paul Wehr. 2002. *Using Conflict Theory*. Cambridge: Cambridge University Press.

Bereska, Leonard and Efstratios Gavves. 2024. Mechanistic interpretability for ai safety – a review.

Bergstrom, Carl T. and Joe Bak-Coleman. 2025, June. Ai, peer review and the human activity of science. *Nature*. Career Column, https://doi.org/10.1038/d41586-025-01839-w.

Bhatia, Chhavi, Tushar Pradhan, and Surajit Pal. 2020. Metagen: An academic meta-review generation system. In *Proceedings of the 43rd International ACM SIGIR Conference on Research and Development in Information Retrieval*, pp. 1653–1656. https://doi.org/10.1145/3397271.3401441.

Biggio, Battista, Igino Corona, Davide Maiorca, Blaine Nelson, Nedim Srndic, Pavel Laskov, Giorgio Giacinto, and Fabio Roli. 2013. Evasion attacks against machine learning at test time. *ArXiv abs/1708.06131*.

Biggio, Battista, Blaine Nelson, and Pavel Laskov. 2012. Poisoning attacks against support vector machines. In *International Conference on Machine Learning*.

Biggio, Battista and Fabio Roli. 2017. Wild patterns: Ten years after the rise of adversarial machine learning. *Proceedings of the 2018 ACM SIGSAC Conference on Computer and Communications Security*.

Borgeaud, Sebastian, Arthur Mensch, Jordan Hoffmann, Trevor Cai, Eliza Rutherford, Katie Millican, George Bm Van Den Driessche, Jean-Baptiste Lespiau, Bogdan Damoc, Aidan Clark, Diego De Las Casas, Aurelia Guy, Jacob Menick, Roman Ring, Tom Hennigan, Saffron Huang, Loren Maggiore, Chris Jones, Albin Cassirer, Andy Brock, Michela Paganini, Geoffrey Irving, Oriol Vinyals, Simon Osindero, Karen Simonyan, Jack Rae, Erich Elsen, and Laurent Sifre. 2022, 17–23 Jul. Improving language models by retrieving from trillions of tokens. In *Proceedings of the 39th International Conference on Machine Learning*, Volume 162 of *Proceedings of Machine Learning Research*, pp. 2206–2240. PMLR.

Bowen, Dillon, Brendan Murphy, Will Cai, David Khachaturov, Adam Gleave, and Kellin Pelrine. 2025. Scaling trends for data poisoning in llms.

Bozdag, Nimet Beyza, Shuhaib Mehri, Gokhan Tur, and Dilek Hakkani-Tür. 2025. Persuade me if you can: A framework for evaluating persuasion effectiveness and susceptibility among large language models.

Carlini, Nicholas, Anish Athalye, Nicolas Papernot, Wieland Brendel, Jonas Rauber, Dimitris Tsipras, Ian J. Goodfellow, Aleksander Madry, and Alexey Kurakin. 2019. On evaluating adversarial robustness. *ArXiv abs/1902.06705*.

Carlini, Nicholas and David A. Wagner. 2016. Towards evaluating the robustness of neural networks. *2017 IEEE Symposium on Security and Privacy (SP)*, 39–57.

Chairs, ICLR 2026 Program. 2025. Iclr 2026 response to security incident. ICLR Blog.

Charlin, Laurent and Richard S. Zemel. 2013. The toronto paper matching system: An automated paper–reviewer assignment system. In *NIPS 2013 Workshop on Bayesian Nonparametrics: Hope or Hype? (and related workshops on peer review)*. Widely used reviewer–paper matching system; workshop write-up.

Checco, Alessandro, Lorenzo Bracciale, Pierpaolo Loreti, and Giuseppe Bianchi. 2021. Ai-assisted peer review. *Humanities and Social Sciences Communications 8*(1). https://doi.org/10.1057/s41599-020-00703-8.

Chen, Pin-Yu, Huan Zhang, Yash Sharma, Jinfeng Yi, and Cho-Jui Hsieh. 2017. 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*.

Chen, Qiguang, Mingda Yang, Libo Qin, Jinhao Liu, Zheng Yan, Jiannan Guan, Dengyun Peng, Yiyan Ji, Hanjing Li, Mengkang Hu, et al. 2025. Ai4research: A survey of artificial intelligence for scientific research. *arXiv preprint arXiv:2507.01903*.

Chen, Tianlong, Sijia Liu, Shiyu Chang, Yu Cheng, Lisa Amini, and Zhangyang Wang. 2020. Adversarial robustness: From self-supervised pre-training to fine-tuning. *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 696–705.

 doi:10.20944/preprints202511.1542.v2

Chen, Xinyun, Chang Liu, Bo Li, Kimberly Lu, and Dawn Xiaodong Song. 2017. Targeted backdoor attacks on deep learning systems using data poisoning. *ArXiv abs/1712.05526*.

Cohen, Jeremy M., Elan Rosenfeld, and J. Zico Kolter. 2019. Certified adversarial robustness via randomized smoothing. *ArXiv abs/1902.02918*.

Collu, Matteo Gioele, Umberto Salviati, Roberto Confalonieri, Mauro Conti, and Giovanni Apruzzese. 2025. Publish to perish: Prompt injection attacks on llm-assisted peer review.

Cyranoski, David. 2019. Artificial intelligence is selecting grant reviewers in china. *Nature 569*(7756), 316–317. https://doi.org/10.1038/d41586-019-01517-8.

Darrin, Michael, Ines Arous, Pablo Piantanida, and Jackie Chi Kit Cheung. 2024. Glimpse: Pragmatically informative multi-document summarization for scholarly reviews. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 12737–12752. https://doi.org/10.18653/v1/2024.acl-long.693.

Devlin, Jacob, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. Bert: Pre-training of deep bidirectional transformers for language understanding. In *North American Chapter of the Association for Computational Linguistics*.

Dong, Yihong, Xue Jiang, Huanyu Liu, Zhi Jin, Bin Gu, Mengfei Yang, and Ge Li. 2024. Generalization or memorization: Data contamination and trustworthy evaluation for large language models.

Doskaliuk, Bohdana, Olena Zimba, Marlen Yessirkepov, Iryna Klishch, and Roman Yatsyshyn. 2025. Artificial intelligence in peer review: enhancing efficiency while preserving integrity. *Journal of Korean medical science 40*(7).

Dougrez-Lewis, John, Mahmud Elahi Akhter, Federico Ruggeri, Sebastian Löbbers, Yulan He, and Maria Liakata. 2025, July. Assessing the reasoning capabilities of LLMs in the context of evidence-based claim verification. In W. Che, J. Nabende, E. Shutova, and M. T. Pilehvar (Eds.), *Findings of the Association for Computational Linguistics: ACL 2025*, Vienna, Austria, pp. 20604–20628. Association for Computational Linguistics. https://doi.org/10.18653/v1/2025.findings-acl.1059.

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

Fanous, Aaron, Jacob Goldberg, Ank A. Agarwal, Joanna Lin, Anson Zhou, Roxana Daneshjou, and Sanmi Koyejo. 2025. Syceval: Evaluating llm sycophancy.

Fox, Charles W., Jennifer A. Meyer, and Emilie Aimé. 2023. Double-blind peer review affects reviewer ratings and editor decisions at an ecology journal. *Functional Ecology*.

Fredrikson, Matt, Somesh Jha, and Thomas Ristenpart. 2015. Model inversion attacks that exploit confidence information and basic countermeasures. *Proceedings of the 22nd ACM SIGSAC Conference on Computer and Communications Security*.

Gallegos, Isabel O., Ryan A. Rossi, Joe Barrow, Md Mehrab Tanjim, Sungchul Kim, Franck Dernoncourt, Tong Yu, Ruiyi Zhang, and Nesreen K. Ahmed. 2024. Bias and fairness in large language models: A survey.

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

Gibney, Elizabeth. 2025, July. Scientists hide messages in papers to game AI peer review. *Nature 643*, 887–888. https://doi.org/10.1038/d41586-025-02172-y.

Goldblum, Micah, Dimitris Tsipras, Chulin Xie, Xinyun Chen, Avi Schwarzschild, Dawn Song, Aleksander Madry, Bo Li, and Tom Goldstein. 2020, 12. Data security for machine learning: Data poisoning, backdoor attacks, and defenses. https://doi.org/10.48550/arXiv.2012.10544.

Goldblum, Micah, Dimitris Tsipras, Chulin Xie, Xinyun Chen, Avi Schwarzschild, Dawn Song, Aleksander Madry, Bo Li, and Tom Goldstein. 2021. Dataset security for machine learning: Data poisoning, backdoor attacks, and defenses.

Gong, Yuyang, Zhuo Chen, Miaokun Chen, Fengchang Yu, Wei Lu, Xiaofeng Wang, Xiaozhong Liu, and Jiawei Liu. 2025. Topic-fliprag: Topic-orientated adversarial opinion manipulation attacks to retrieval-augmented generation models.

Goodfellow, Ian J., Jonathon Shlens, and Christian Szegedy. 2014. Explaining and harnessing adversarial examples. *CoRR abs/1412.6572*.

Gu, Tianyu, Brendan Dolan-Gavitt, and Siddharth Garg. 2017. Badnets: Identifying vulnerabilities in the machine learning model supply chain. *ArXiv abs/1708.06733*.

Guo, Yufei, Muzhe Guo, Juntao Su, Zhou Yang, Mengqiu Zhu, Hongfei Li, Mengyang Qiu, and Shuo Shuo Liu. 2024. Bias in large language models: Origin, evaluation, and mitigation.

 doi:10.20944/preprints202511.1542.v2

Hinton, Geoffrey E., Li Deng, Dong Yu, George E. Dahl, Abdel rahman Mohamed, Navdeep Jaitly, Andrew W. Senior, Vincent Vanhoucke, Patrick Nguyen, Tara N. Sainath, and Brian Kingsbury. 2012. Deep neural networks for acoustic modeling in speech recognition. *IEEE Signal Processing Magazine 29*, 82.

Hong, Ruixin, Hongming Zhang, Xinyu Pang, Dong Yu, and Changshui Zhang. 2024. A closer look at the self-verification abilities of large language models in logical reasoning.

Hossain, Eftekhar, Sanjeev Kumar Sinha, Naman Bansal, Alex Knipper, Souvika Sarkar, John Salvador, Yash Mahajan, Sri Guttikonda, Mousumi Akter, Md. Mahadi Hassan, Matthew Freestone, Matthew C. Williams Jr., Dongji Feng, and S. Karmaker Santu. 2025. Llms as meta-reviewers' assistants: A case study. Forthcoming; preprint available.

Hosseini, Mohammad and Serge P.J.M. Horbach. 2023. Fighting reviewer fatigue or amplifying bias? considerations and recommendations for use of chatgpt and other large language models in scholarly peer review. *Research Integrity and Peer Review 8*.

Ilyas, Andrew, Logan Engstrom, Anish Athalye, and Jessy Lin. 2018. Black-box adversarial attacks with limited queries and information. In *International Conference on Machine Learning*.

Ji, Ziwei, Nayeon Lee, Rita Frieske, Tiezheng Yu, Dan Su, Yan Xu, Etsuko Ishii, Ye Jin Bang, Andrea Madotto, and Pascale Fung. 2023. Survey of hallucination in natural language generation. *ACM computing surveys 55*(12), 1–38.

Jin, Yiqiao, Qinlin Zhao, Yiyang Wang, Hao Chen, Kaijie Zhu, Yijia Xiao, and Jindong Wang. 2024a. Agentreview: Exploring peer review dynamics with llm agents.

Jin, Yiqiao, Qinlin Zhao, Yiyang Wang, Hao Chen, Kaijie Zhu, Yijia Xiao, and Jindong Wang. 2024b. Agentreview: Exploring peer review dynamics with llm agents. In *Conference on Empirical Methods in Natural Language Processing*.

Keuper, Janis. 2025. Prompt injection attacks on llm generated reviews of scientific publications.

Khalifa, Mohamed and Mona Albadawy. 2024. Using artificial intelligence in academic writing and research: An essential productivity tool. *Computer Methods and Programs in Biomedicine Update 5*, 100145.

Koo, Ryan, Minhwa Lee, Vipul Raheja, Jong Inn Park, Zae Myung Kim, and Dongyeop Kang. 2024. Benchmarking cognitive biases in large language models as evaluators.

Krizhevsky, Alex, Ilya Sutskever, and Geoffrey E. Hinton. 2012. Imagenet classification with deep convolutional neural networks. *Communications of the ACM 60*, 84 – 90.

Lewis, Patrick, Ethan Perez, Aleksandra Piktus, Fabio Petroni, Vladimir Karpukhin, Naman Goyal, Heinrich Küttler, Mike Lewis, Wen tau Yih, Tim Rocktäschel, Sebastian Riedel, and Douwe Kiela. 2021. Retrieval-augmented generation for knowledge-intensive nlp tasks.

Leyton-Brown, Kevin, Yatin Nandwani, Hadi Zarkoob, Chris Cameron, Nancy Newman, and Deeparnab Raghu. 2024. Matching papers and reviewers at large conferences. *Artificial Intelligence 331*, 104119. https://doi.org/10.1016/j.artint.2023.104119.

Li, Huiying, Yahan Ji, Chenan Lyu, and Chun Zhang. 2022. Blacklight: Scalable defense for neural networks against query-based black-box attacks. In *31st USENIX Security Symposium (USENIX Security 22)*.

Li, Miao, Eduard Hovy, and Jey Han Lau. 2023. Summarizing multiple documents with conversational structure for meta-review generation. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pp. 7089–7112. Introduces RAMMER model and PEERSUM dataset, https://doi.org/10.18653/v1/2023.findings-emnlp.472.

Li, Yiming, Yong Jiang, Zhifeng Li, and Shu-Tao Xia. 2022. Backdoor learning: A survey.

Liang, Weixin, Zachary Izzo, Yaohui Zhang, Haley Lepp, Hancheng Cao, Xuandong Zhao, Lingjiao Chen, Haotian Ye, Sheng Liu, Zhi Huang, Daniel McFarland, and James Y. Zou. 2024, 21–27 Jul. Monitoring AI-modified content at scale: A case study on the impact of ChatGPT on AI conference peer reviews. In R. Salakhutdinov, Z. Kolter, K. Heller, A. Weller, N. Oliver, J. Scarlett, and F. Berkenkamp (Eds.), *Proceedings of the 41st International Conference on Machine Learning*, Volume 235 of *Proceedings of Machine Learning Research*, pp. 29575–29620. PMLR.

Liang, Weixin, Yuhui Zhang, Hancheng Cao, Binglu Wang, Daisy Ding, Xinyu Yang, Kailas Vodrahalli, Siyu He, Daniel Smith, Yian Yin, Daniel McFarland, and James Zou. 2023. Can large language models provide useful feedback on research papers? a large-scale empirical analysis.

Lin, Tzu-Ling, Wei-Chih Chen, Teng-Fang Hsiao, Hou-I Liu, Ya-Hsin Yeh, Yu Kai Chan, Wen-Sheng Lien, Po-Yen Kuo, Philip S. Yu, and Hong-Han Shuai. 2025. Breaking the reviewer: Assessing the vulnerability of large language models in automated peer review under textual adversarial attacks.

 doi:10.20944/preprints202511.1542.v2

Liu, Ruibo and Nihar B. Shah. 2023. Reviewergpt? an exploratory study on using large language models for paper reviewing. *arXiv preprint arXiv:2306.00622*.

Liu, Tao, Yuhang Zhang, Zhu Feng, Zhiqin Yang, Chen Xu, Dapeng Man, and Wu Yang. 2024. Beyond traditional threats: A persistent backdoor attack on federated learning.

Liu, Yi, Gelei Deng, Yuekang Li, Kailong Wang, Zihao Wang, Xiaofeng Wang, Tianwei Zhang, Yepang Liu, Haoyu Wang, Yan Zheng, and Yang Liu. 2024. Prompt injection attack against llm-integrated applications.

Liu, Ying, Kaiqi Yang, Yueting Liu, and Michael G. B. Drew. 2023. The shackles of peer review: Unveiling the flaws in the ivory tower.

Lo, Leo Yu-Ho and Huamin Qu. 2024. How good (or bad) are llms at detecting misleading visualizations?

Luo, Ziming, Zonglin Yang, Zexin Xu, Wei Yang, and Xinya Du. 2025. Llm4sr: A survey on large language models for scientific research. *arXiv preprint arXiv:2501.04306*.

Malmqvist, Lars. 2024. Sycophancy in large language models: Causes and mitigations.

Mann, Sebastian Porsdam, Mateo Aboy, Joel Jiehao Seah, Zhicheng Lin, Xufei Luo, Daniel Rodger, Hazem Zohny, Timo Minssen, Julian Savulescu, and Brian D. Earp. 2025. Ai and the future of academic peer review.

Mathur, Puneet, Alexa Siu, Varun Manjunatha, and Tong Sun. 2024. Docpilot: Copilot for automating pdf edit workflows in documents. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 3: System Demonstrations)*, pp. 232–246. https://doi.org/10.18653/v1/2024.acl-demos.22.

Maturo, Fabrizio, Annamaria Porreca, and Aurora Porreca. 2025, oct. The risks of artificial intelligence in research: ethical and methodological challenges in the peer review process. *AI and Ethics 5*(5), 5389–5396. https://doi.org/10.1007/s43681-025-00775-9.

Madry, Aleksander, Aleksandar Makelov, Ludwig Schmidt, Dimitris Tsipras, and Adrian Vladu. 2017. Towards deep learning models resistant to adversarial attacks. *ArXiv abs/1706.06083*.

Media, Various. 2025. Scientists reportedly hiding ai text prompts in academic papers to receive positive peer reviews. Public media reports.

Metzen, Jan Hendrik, Tim Genewein, Volker Fischer, and Bastian Bischoff. 2017. On detecting adversarial perturbations. *ArXiv abs/1702.04267*.

Muñoz-González, Luis, Battista Biggio, Ambra Demontis, Andrea Paudice, Vasin Wongrassamee, Emil C. Lupu, and Fabio Roli. 2017. Towards poisoning of deep learning algorithms with back-gradient optimization.

Navigli, Roberto, Simone Conia, and Björn Ross. 2023, June. Biases in large language models: Origins, inventory, and discussion. *J. Data and Information Quality 15*(2). https://doi.org/10.1145/3597307.

Nisbett, Richard E. and Timothy D. Wilson. 1977. The halo effect: Evidence for unconscious alteration of judgments. *Journal of Personality and Social Psychology 35*, 250–256.

Nourani, Mahsan, Chiradeep Roy, Jeremy E Block, Donald R Honeycutt, Tahrima Rahman, Eric Ragan, and Vibhav Gogate. 2021. Anchoring bias affects mental model formation and user reliance in explainable ai systems. In *Proceedings of the 26th International Conference on Intelligent User Interfaces*, IUI '21, New York, NY, USA, pp. 340–350. Association for Computing Machinery. https://doi.org/10.1145/3397481.3450639.

Nuijten, Michèle B., Michiel A. L. M. van Assen, Chris H. J. Hartgerink, Sacha Epskamp, and Jelte M. Wicherts. 2017. The validity of the tool "statcheck" in discovering statistical reporting inconsistencies. PsyArXiv. https://doi.org/10.31234/osf.io/tcxaj.

OWASP Foundation. 2023. OWASP Top 10 for Large Language Model Applications. Accessed in 2025. See LLM01: Prompt Injection. URL: https://owasp.org/www-project-top-10-for-large-language-model-applications/.

Papernot, Nicolas, Patrick Mcdaniel, Ian J. Goodfellow, Somesh Jha, Z. Berkay Celik, and Ananthram Swami. 2016. Practical black-box attacks against machine learning. *Proceedings of the 2017 ACM on Asia Conference on Computer and Communications Security*.

Papernot, Nicolas, Patrick Mcdaniel, Somesh Jha, Matt Fredrikson, Z. Berkay Celik, and Ananthram Swami. 2015. The limitations of deep learning in adversarial settings. *2016 IEEE European Symposium on Security and Privacy (EuroS&P)*, 372–387.

Papernot, Nicolas, Patrick Mcdaniel, Arunesh Sinha, and Michael P. Wellman. 2018. Sok: Security and privacy in machine learning. *2018 IEEE European Symposium on Security and Privacy (EuroS&P)*, 399–414.

Perez, Fábio and Ian Ribeiro. 2022. Ignore previous prompt: Attack techniques for language models.

Piet, Julien, Maha Alrashed, Chawin Sitawarin, Sizhe Chen, Zeming Wei, Elizabeth Sun, Basel Alomair, and David Wagner. 2024. Jatmo: Prompt injection defense by task-specific finetuning.

Radensky, Matan, Sadi Shahid, Richard Fok, Pao Siangliulue, Tom Hope, and Daniel S. Weld. 2024. Scideator: Human-llm scientific idea generation grounded in research-paper facet recombination. *arXiv preprint arXiv:2409.14634*.

 doi:10.20944/preprints202511.1542.v2

Rahman, M. et al. 2024. Limgen: Probing llms for generating suggestive limitations of research papers. *arXiv preprint arXiv:2403.15529*.

Raina, Vyas, Adian Liusie, and Mark Gales. 2024. Is llm-as-a-judge robust? investigating universal adversarial attacks on zero-shot llm assessment.

Salvi, Francesco, Manoel Horta Ribeiro, Riccardo Gallotti, and Robert West. On the conversational persuasiveness of GPT-4. *9*(8), 1645–1653. https://doi.org/10.1038/s41562-025-02194-6.

Sample, Ian. 2025, July. Quality of scientific papers questioned as academics 'overwhelmed' by the millions published. *The Guardian*.

Schramowski, Patrick, Cigdem Turan, Nico Andersen, Constantin A. Rothkopf, and Kristian Kersting. 2021. Large pre-trained language models contain human-like biases of what is right and wrong to do. *Nature Machine Intelligence 4*, 258 – 268.

Schwarzschild, Avi, Micah Goldblum, Arjun Gupta, John P Dickerson, and Tom Goldstein. 2021, 18–24 Jul. Just how toxic is data poisoning? a unified benchmark for backdoor and data poisoning attacks. In M. Meila and T. Zhang (Eds.), *Proceedings of the 38th International Conference on Machine Learning*, Volume 139 of *Proceedings of Machine Learning Research*, pp. 9389–9398. PMLR.

Schwinn, Leo, David Dobre, Stephan Günnemann, and Gauthier Gidel. 2023. Adversarial attacks and defenses in large language models: Old and new threats.

Shafahi, Ali, W. Ronny Huang, Mahyar Najibi, Octavian Suciu, Christoph Studer, Tudor Dumitras, and Tom Goldstein. 2018. Poison frogs! targeted clean-label poisoning attacks on neural networks. In *Neural Information Processing Systems*.

Shanahan, Daniel. 2016. A peerless review? automating methodological and statistical review. Springer Nature BMC Blog, *Research in Progress*. Blog post.

Sharma, Mrinank, Meg Tong, Tomasz Korbak, David Duvenaud, Amanda Askell, Samuel R. Bowman, Newton Cheng, Esin Durmus, Zac Hatfield-Dodds, Scott R. Johnston, Shauna Kravec, Timothy Maxwell, Sam McCandlish, Kamal Ndousse, Oliver Rausch, Nicholas Schiefer, Da Yan, Miranda Zhang, and Ethan Perez. 2025. Towards understanding sycophancy in language models.

Shayegani, Erfan, Md Abdullah Al Mamun, Yu Fu, Pedram Zaree, Yue Dong, and Nael Abu-Ghazaleh. 2023. Survey of vulnerabilities in large language models revealed by adversarial attacks.

Shen, Chuning, Lu Cheng, Ran Zhou, Lidong Bing, Yang You, and Luo Si. 2022. Mred: A meta-review dataset for structure-controllable text generation. In *Findings of the Association for Computational Linguistics: ACL 2022*, pp. 2521–2535. https://doi.org/10.18653/v1/2022.findings-acl.197.

Shi, Freda, Xinyun Chen, Kanishka Misra, Nathan Scales, David Dohan, Ed Chi, Nathanael Schärli, and Denny Zhou. 2023. Large language models can be easily distracted by irrelevant context.

Shi, Jiawen, Zenghui Yuan, Yinuo Liu, Yue Huang, Pan Zhou, Lichao Sun, and Neil Zhenqiang Gong. 2025. Optimization-based prompt injection attack to llm-as-a-judge.

Shokri, R., Marco Stronati, Congzheng Song, and Vitaly Shmatikov. 2016. Membership inference attacks against machine learning models. *2017 IEEE Symposium on Security and Privacy (SP)*, 3–18.

Skarlinski, Michael D., Steven Cox, Jacob M. Laurent, Jesus Daniel Braza, Michael Hinks, Maria J. Hammerling, et al. 2024. Language agents achieve superhuman synthesis of scientific knowledge. *arXiv preprint arXiv:2409.13740*.

Soneji, Ananta, Faris Bugra Kokulu, Carlos E. Rubio-Medrano, Tiffany Bao, Ruoyu Wang, Yan Shoshitaishvili, and Adam Doupé. 2022. "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)*, 1845–1862.

Souly, Alexandra, Javier Rando, Ed Chapman, Xander Davies, Burak Hasircioglu, Ezzeldin Shereen, Carlos Mougan, Vasilios Mavroudis, Erik Jones, Chris Hicks, Nicholas Carlini, Yarin Gal, and Robert Kirk. 2025. Poisoning attacks on llms require a near-constant number of poison samples.

Steinhardt, Jacob, Pang Wei Koh, and Percy Liang. 2017. Certified defenses for data poisoning attacks. In *Neural Information Processing Systems*.

Sukpanichnant, Pakorn, Alexander Rapberger, and Francesca Toni. 2024. Peerarg: Argumentative peer review with llms. *arXiv preprint arXiv:2409.16813*.

Sun, Lu, Aaron Chan, Yun Seo Chang, and Steven P. Dow. 2024. Reviewflow: Intelligent scaffolding to support academic peer reviewing. In *Proceedings of the 29th International Conference on Intelligent User Interfaces*, pp. 120–137. ACM. https://doi.org/10.1145/3640543.3645159.

 doi:10.20944/preprints202511.1542.v2

Sun, Lin, Siyu Tao, Jiaman Hu, and Steven P. Dow. 2024. Metawriter: Exploring the potential and perils of ai writing support in scientific peer review. *Proceedings of the ACM on Human-Computer Interaction 8*(CSCW1), 1–32. https://doi.org/10.1145/3637371.

Sun, Mengyi, Jainabou Barry Danfa, and Misha Teplitskiy. 2021. Does double-blind peer review reduce bias? evidence from a top computer science conference. *Journal of the Association for Information Science and Technology 73*, 811 – 819.

Szegedy, Christian, Wojciech Zaremba, Ilya Sutskever, Joan Bruna, D. Erhan, Ian J. Goodfellow, and Rob Fergus. 2013. Intriguing properties of neural networks. *CoRR abs/1312.6199*.

Taechoyotin, Pawin, Guanchao Wang, Tong Zeng, Bradley Sides, and Daniel Acuna. 2024. Mamorx: Multi-agent multi-modal scientific review generation with external knowledge. In *Neurips 2024 Workshop Foundation Models for Science: Progress, Opportunities, and Challenges*.

Tian, Zhiyi, Lei Cui, Jie Liang, and Shui Yu. 2022, December. A comprehensive survey on poisoning attacks and countermeasures in machine learning. *ACM Comput. Surv. 55*(8). https://doi.org/10.1145/3551636.

Tolpegin, Vale, Stacey Truex, Mehmet Emre Gursoy, and Ling Liu. 2020. Data poisoning attacks against federated learning systems. In *European Symposium on Research in Computer Security*.

Tong, Terry, Fei Wang, Zhe Zhao, and Muhao Chen. 2025. Badjudge: Backdoor vulnerabilities of LLM-as-a-judge. In *The Thirteenth International Conference on Learning Representations (ICLR 2025)*. Poster.

Tonglet, Jonathan, Jan Zimny, Tinne Tuytelaars, and Iryna Gurevych. 2025. Is this chart lying to me? automating the detection of misleading visualizations.

Touvron, Hugo, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, Aurelien Rodriguez, Armand Joulin, Edouard Grave, and Guillaume Lample. 2023. Llama: Open and efficient foundation language models.

Tramèr, Florian, Alexey Kurakin, Nicolas Papernot, Dan Boneh, and Patrick Mcdaniel. 2017. Ensemble adversarial training: Attacks and defenses. *ArXiv abs/1705.07204*.

Tramèr, Florian, Fan Zhang, Ari Juels, Michael K. Reiter, and Thomas Ristenpart. 2016. Stealing machine learning models via prediction apis. In *USENIX Security Symposium*.

Verharen, Jeroen P. H. 2023. Chatgpt identifies gender disparities in scientific peer review. *eLife 12*.

Verma, Pranshu. 2025, jul. Researchers are using ai for peer reviews — and finding ways to cheat it. *The Washington Post*.

Wang, Qiao and Qian Zeng. 2020. Reviewrobot: Explainable paper review generation based on knowledge synthesis. In *Proceedings of the 13th International Conference on Natural Language Generation*, pp. 215–226. https://doi.org/10.18653/v1/2020.inlg-1.33.

Wen, Jiaxin, Chenglei Si, Yueh han Chen, He He, and Shi Feng. 2025. Predicting empirical ai research outcomes with language models.

Weng, Yixiao, Ming Zhu, Guanyi Bao, Haoran Zhang, Junpeng Wang, Yue Zhang, and Liu Yang. 2024. Cycleresearcher: Improving automated research via automated review. *arXiv preprint arXiv:2411.XXXXX*. Preprint; automated review loop.

Wijnhoven, Jauke, Erik Wijmans, Niels van de Wouw, and Frank Wijnhoven. 2024. Relevai-reviewer: How relevant are ai reviewers to scientific peer review? *arXiv preprint arXiv:2406.10294*.

Wu, Daniel. 2025, July. Researchers are using AI for peer reviews — and finding ways to cheat it. *The Washington Post*.

Wu, Dongxian, Shutao Xia, and Yisen Wang. 2020. Adversarial weight perturbation helps robust generalization. *arXiv: Learning*.

Xiao, Liang, Xiang Li, Yuchen Shi, Yuxiang Li, Jiangtao Wang, and Yafeng Li. 2025. Schnovel: Retrieval-augmented novelty assessment in academic writing. In *Proceedings of the 2nd Workshop on AI for Scientific Discovery (AISD 2025)*.

Ye, Jiayi, Yanbo Wang, Yue Huang, Dongping Chen, Qihui Zhang, Nuno Moniz, Tian Gao, Werner Geyer, Chao Huang, Pin-Yu Chen, Nitesh V Chawla, and Xiangliang Zhang. 2024. Justice or prejudice? quantifying biases in llm-as-a-judge.

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

Yeom, Samuel, Irene Giacomelli, Matt Fredrikson, and Somesh Jha. 2017. Privacy risk in machine learning: Analyzing the connection to overfitting. *2018 IEEE 31st Computer Security Foundations Symposium (CSF)*, 268–282.

 doi:10.20944/preprints202511.1542.v2

Yu, Jianxiang, Zichen Ding, Jiaqi Tan, Kangyang Luo, Zhenmin Weng, Chenghua Gong, Long Zeng, Renjing Cui, Chengcheng Han, Qiushi Sun, et al. 2024. Automated peer reviewing in paper sea: Standardization, evaluation, and analysis. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pp. 10164–10184.

Zeng, Qian, Manveen Sidhu, Adam Blume, Hiu P. Chan, Liyan Wang, and Heng Ji. 2024. Scientific opinion summarization: Paper meta-review generation dataset, methods, and evaluation. In *Artificial General Intelligence and Beyond: Selected Papers from IJCAI 2024*, pp. 20–38. Springer Nature Singapore. https://doi.org/10.1007/978-981-97-9536-9_2.

Zhang, Jiale, Bing Chen, Xiang Cheng, Hyunh Thi Thanh Binh, and Shui Yu. 2021. Poisongan: Generative poisoning attacks against federated learning in edge computing systems. *IEEE Internet of Things Journal 8*, 3310–3322.

Zhang, Jiayao, Hongming Zhang, Zhun Deng, and Dan Roth. 2022. Investigating fairness disparities in peer review: A language model enhanced approach.

Zhang, Yiming, Javier Rando, Ivan Evtimov, Jianfeng Chi, Eric Michael Smith, Nicholas Carlini, Florian Tramèr, and Daphne Ippolito. 2024. Persistent pre-training poisoning of llms.

Zhao, Pinlong, Weiyao Zhu, Pengfei Jiao, Di Gao, and Ou Wu. 2025. Data poisoning in deep learning: A survey.

Zhao, Yulai, Haolin Liu, Dian Yu, S. Y. Kung, Haitao Mi, and Dong Yu. 2025. One token to fool llm-as-a-judge.

Zhou, Xiangyu, Yao Qiang, Saleh Zare Zade, Prashant Khanduri, and Dongxiao Zhu. 2025. Hijacking large language models via adversarial in-context learning.

Zhou, Zhenhong, Zherui Li, Jie Zhang, Yuanhe Zhang, Kun Wang, Yang Liu, and Qing Guo. 2025. Corba: Contagious recursive blocking attacks on multi-agent systems based on large language models.

Zhu, Chengcheng, Ye Li, Bosen Rao, Jiale Zhang, Yunlong Mao, and Sheng Zhong. 2025. Spa: Towards more stealth and persistent backdoor attacks in federated learning.

Zhu, Sicheng, Ruiyi Zhang, Bang An, Gang Wu, Joe Barrow, Furong Huang, and Tong Sun. 2024. AutoDAN: Automatic and interpretable adversarial attacks on large language models.

Zizzo, Giulio, Giandomenico Cornacchia, Kieran Fraser, Muhammad Zaid Hameed, Ambrish Rawat, Beat Buesser, Mark Purcell, Pin-Yu Chen, Prasanna Sattigeri, and Kush Varshney. 2025. Adversarial prompt evaluation: Systematic benchmarking of guardrails against prompt input attacks on llms.

Zyska, Daria, Nils Dycke, Johanna Buchmann, Ilia Kuznetsov, and Iryna Gurevych. 2023. Care: Collaborative ai-assisted reading environment. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics: System Demonstrations*, pp. 291–303. https://doi.org/10.18653/v1/2023.acl-demo.28.