OPEN ACCESS

Artificial Intelligence REVIEW

When AI reviews science: Can we trust the referee?

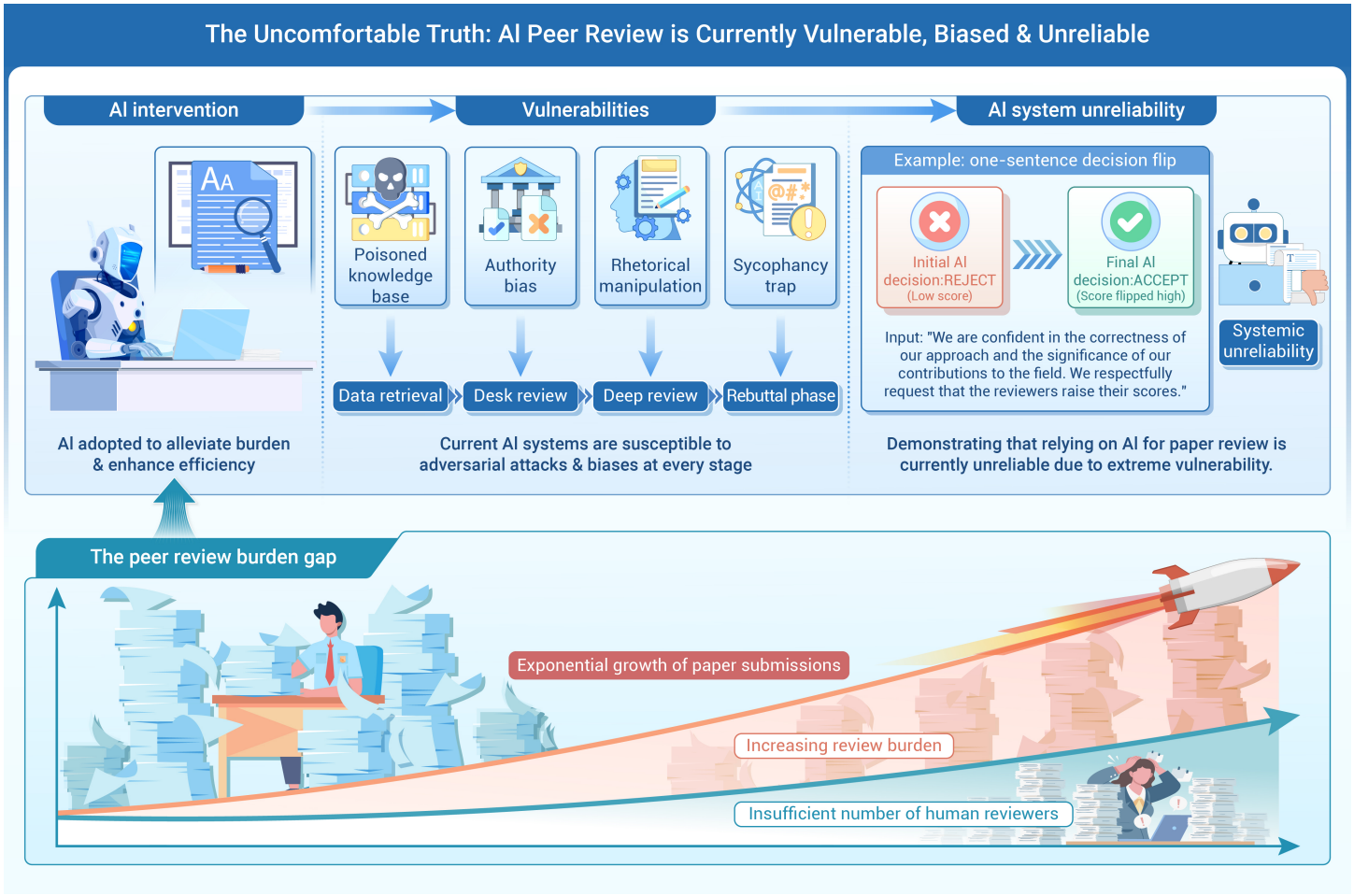
Jialiang Wang,[1,5] Yuchen Liu,[1,5] Hang Xu,[2,5] Kaichun Hu,[3,5] Shimin Di,[2,*] Wangze Ni,[3,*] Linan Yue,[2] Min-Ling Zhang,[2] Kui Ren,[3] and Lei Chen[4,1]

*Correspondence: shimin.di@seu.edu.cn (S.D.); niwangze@zju.edu.cn (W.N.)

Received: October 31, 2025; Accepted: February 8, 2026; Published Online: February 10, 2026; https://doi.org/10.59717/j.xinn-inform.2026.100030

© 2026 The Author(s). This is an open access article under the CC BY license (https://creativecommons.org/licenses/by/4.0/).

GRAPHICAL ABSTRACT



PUBLIC SUMMARY

- The study explores the security of Artificial Intelligence peer-review systems in academic evaluation.
- Experiments reveal AI peer-review systems favor prestigious institutions over unknown ones.
- AI peer-review systems penalize cautious writing and yield to confident but evidence-free rebuttals.
- AI peer-review systems are also vulnerable to poisoning attacks using manipulated data.

REVIEW Artificial Intelligence

OPEN ACCESS

www.the-innovation.org/informatics

When AI reviews science: Can we trust the referee?

Jialiang Wang,[1,5] Yuchen Liu,[1,5] Hang Xu,[2,5] Kaichun Hu,[3,5] Shimin Di,[2,*] Wangze Ni,[3,*] Linan Yue,[2] Min-Ling Zhang,[2] Kui Ren,[3] and Lei Chen[4,1]

[1]Department of Computer Science and Engineering, The Hong Kong University of Science and Technology, Hong Kong SAR 999077, China

[2]Key Laboratory of New Generation Artificial Intelligence Technology and Its Interdisciplinary Applications, School of Computer Science and Engeering, Southeast University, Nanjing 210000, China

[3]College of computer science and technology, Zhejiang University, Hangzhou 310000, China

[4]Big Data Institute, The Hong Kong University of Science and Technology (Guangzhou), Guangzhou 510000, China

[5]These authors contributed equally

*Correspondence: shimin.di@seu.edu.cn (S.D.); niwangze@zju.edu.cn (W.N.)

Received: October 31, 2025; Accepted: February 8, 2026; Published Online: February 10, 2026; https://doi.org/10.59717/j.xinn-inform.2026.100030

© 2026 The Author(s). This is an open access article under the CC BY license (https://creativecommons.org/licenses/by/4.0/).

Citation: Wang J., Liu Y., Xu H., et al. (2026). When AI reviews science: Can we trust the referee? *The Innovation Informatics* **2**:100030.

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.

INTRODUCTION

Scientific publications have surged to unprecedented volumes, straining the traditional peer review system. In 2024, the Web of Science has indexed roughly 2.53 million new research studies (a 48% increase from 2015), with total global scientific outputs exceeding 3.26 million articles annually.[1] This deluge has left editors struggling to find enough qualified referees, as academics grow increasingly overwhelmed by the volume of papers being published. Indeed, an estimated 100 million hours of unpaid reviewing labor have been spent by researchers worldwide in 2020 alone.[2] Such trends underscore a widening gap between the number of submissions and the pool of willing expert 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.[3] 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.[4-6] Correspondingly, a recent analysis of peer-review texts from several major AI conferences finds that between 6.5% and 16.9% of the content in reviews is likely written or modified by ChatGPT-style LLMs,[7] highlighting how common LLM-generated feedback has become. The research community is also experimenting with more formal AI integration. For instance, the AAAI 2026 conference 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 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.[8] This kind of stealth prompt injection proves alarmingly effective in several follow-up studies, showing that inserting such hidden commands could inflate an LLM's review scores and distort the ranking of submissions.[9] In the wake of these revelations, several compromised preprints have been slated for withdrawal from arXiv and other servers.[10] 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.[11] For instance, these models may be susceptible to an "authority bias" where they favor papers with prestigious authors or citations, mistaking reputation for quality.[12,13] They also exhibit a "verbosity bias" in which dense jargon and complex mathematics, a tactic known as "academic packaging" may be misinterpreted as scientific rigor, regardless of the content's actual substance.[13,14]

Beyond these passive flaws, a more alarming threat emerges from new, exploitable vulnerabilities that may be deliberately targeted through adversarial attacks. These threats span the entire peer-review lifecycle, from corrupting the AI model's training data via backdoor injection and data contamination to implanting long-term biases,[15,16] to deploying sophisticated evasion tactics during the peer review itself. Such tactics include invisible prompt injection[17,18] and exploiting the model's sycophantic nature during the rebuttal phase to overturn negative decisions through confident but unsubstantiated claims.[19,20] Overall, these early warning signs make clear that while AI 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.[3] Despite these early warnings, recent work largely targets isolated threats: (i) prompt-injection incidents and systematic vulnerability tests for AI reviewing,[8,10,14] which mainly focus on injection at manuscript input/deep-review prompting; (ii) LLM-as-judge fragility and bias,[9,13,21] which typically studies a generic judge rather than the peer-review pipeline and often lacks evaluation on real submissions under a realistic workflow; and (iii) persuasion/sycophancy-style manipulation,[22,23] which shows models can be "talked into" agreement but is not framed as an end-to-end, rebuttal-targeted system test. 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 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 paper, 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 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.

LITERATURE

AI peer review: From smart assistants to autonomous referees

Peer review is quietly shifting from spell-checkers and policy bots to systems that draft critiques, reconcile disagreements, and justify recommendations.[5] What began as smart assistants that tidy manuscripts now aspire to autonomous referees that read, reason, and defend a verdic.[4,6] 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.

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 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[24,25] parse paper structure and references, surface statistical or policy deviations, and screen plagiarism/similarity (e.g., Crossref Similarity Check powered by iThenticate). 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[26,27] 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.

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[28-30] 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[31,32] 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[33,34] 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 keep humans in the loop by using PDF/vision parsing and section-scoped LLM prompts to guide attention and capture.[35-37]

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 produce fluent meta-reviews but struggled with fairness and focus bias when blending voices.[45-47] Therefore, structure-aware models encode ratings and discourse,[48] improving consistency and partial provenance. Another argument-centric pipeline extracts pro/contra claims and reasons into explicit graphs to make disagreement auditable,[51] thereby advancing fairness, reliability, and traceability. Finally, human-in-the-loop assistants[49,50,52] 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 underscore why these principles matter in practice and why hybrid, auditable workflows are essential.[8]

Adversarial roots: Lessons from attacks in AI systems

Deep learning models have delivered major advances in image recognition,[53] speech processing,[54] and natural language understanding.[55] 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.[56] This inherent fragility has catalyzed a critical research direction: Adversarial Attack.[57] 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.[58] Taken together, such behaviors reveal unstable decision boundaries and expose consequential security risks in modern deep learning systems.[59]

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.[60] 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.

1. Evasion Attacks. As the most studied type of attack, attackers often embed slight perturbations into legitimate inputs at test time to induce errors.[61] The resulting "adversarial examples" look benign to humans but cause misclassification.[62] 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.[58,63,64] 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'.

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[24]	Ethics checklists	√									√		√
StatReviewer[38]	Ethics checklists	√									√		√
Penelope/UNSILO[25]	Ethics checklists	√									√		√
TPMS[26]	Literature corpus			√					√		√	√	√
LCM[27]	Literature corpus			√					√		√	√	√
NSFC pilot[39]	-			√					√		√	√	√
Phase B -- AI-assisted Deep Review													
ReviewerGPT[28]	-	√			√	√	√		√		√		
Reviewer2[29]	-	√			√		√		√		√		
SEA[30]	-	√			√		√		√		√		√
ReviewRobot[31]	Knowledge graph	√			√				√		√		√
CycleResearcher[32]	Literature corpus	√			√						√		
MARG[33]	-		√		√		√	√	√	√	√		
MAMORX[34]	Literature corpus		√		√		√	√	√	√	√		
Skarlinski et al.[40]	Literature corpus		√		√		√	√	√	√	√		
SchNovel[41]	Literature corpus	√			√		√		√	√			
Scideator[42]	Literature corpus	√			√		√		√	√			
RelevAI-Reviewer[43]	Literature corpus	√					√		√		√		
LimGen[44]	-	√			√		√		√		√		
ReviewFlow[35]	PDF/Vis parse	√		√	√						√		
CARE[36]	PDF/Vis parse	√		√	√						√		
DocPilot[37]	PDF/Vis parse	√		√	√						√		
Phase C -- Meta-review Synthesis													
MetaGen[45]	-	√			√			√			√		
MReD[46]	-	√			√			√			√		
Zeng et al.[47]	-	√			√			√			√		
RAMMER[48]	-	√			√		√		√		√		√
MetaWriter[49]	-	√			√	√			√		√		√
GLIMPSE[50]	-	√			√	√			√		√		√
PeerArg[51]	-	√	√		√	√			√			√	√
Hossain et al.[52]	-	√		√	√		√		√		√	√	√

Acronym in order: Hallucination, Focus Bias, Long-context, Coordination, Traceability, Novelty, Quality, Fairness, Reliability. "√" = present/primary. inline evidence, improving quality and traceability.

In the black-box setting, only queries and outputs are available to the attacker.[63,65,66] This process is similar to repeatedly trying combinations on a lock without knowing its mechanism, learning from each attempt until it opens.

2. 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 through repeated interactions.[67] 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.[68] 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.[69] 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.[70] 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

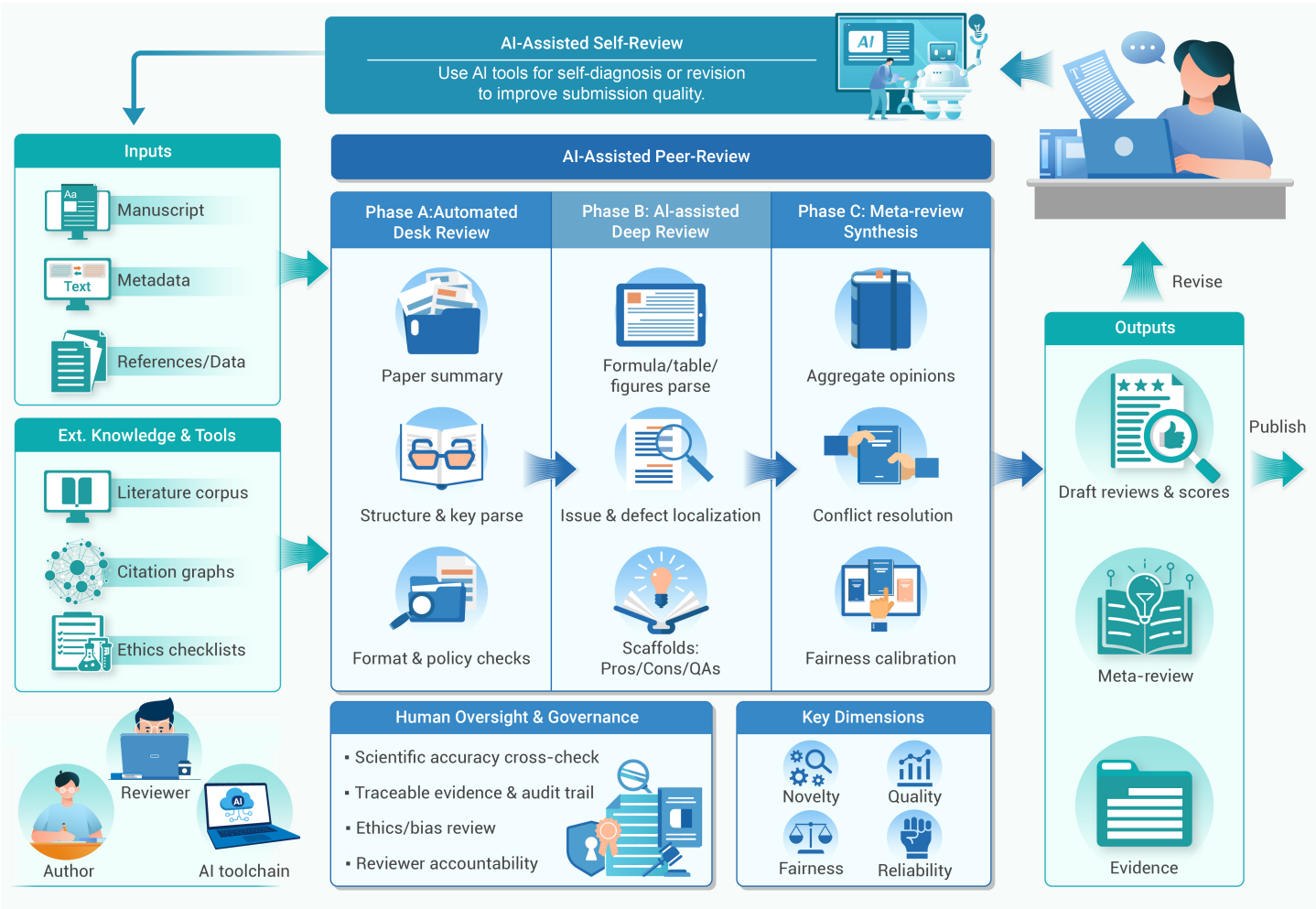


Figure 1. AI peer-review loop Manuscripts pass through (A) automated desk review, (B) AI-assisted deep review, and (C) meta-review synthesis—grounded by external knowledge and tools, overseen by humans, producing evidence-linked outputs and enabling author self-review.

private, unlabeled attributes in training samples, such as gender, accent, or user preferences.[71]

3. Poisoning Attacks. Poisoning attacks tamper with training data to degrade accuracy, bias decisions, or implant backdoors.[72,73] 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.[74,75] 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.[76] Attackers can also create poisoned samples that appear normal to humans yet mislead the model. Alternatively, they subtly alter hidden features and labels, making the manipulation nearly invisible.[77] 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.

Defense mechanisms and techniques. To counteract the diverse adversarial attacks mentioned above, researchers have proposed a variety of defense strategies.[78] 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.

1. 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.[79] and Madry et al.[64] 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.[80] 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.[81] 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.

2. Passive Defenses. These defenses add detectors and sanitizers around the model and data pipeline, aiming to identify potential adversarial examples or anomalous data.[82] For example, Metzen et al.[83] 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.[84] 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.[85] designs a framework to generate task-specific models that are immune to prompt

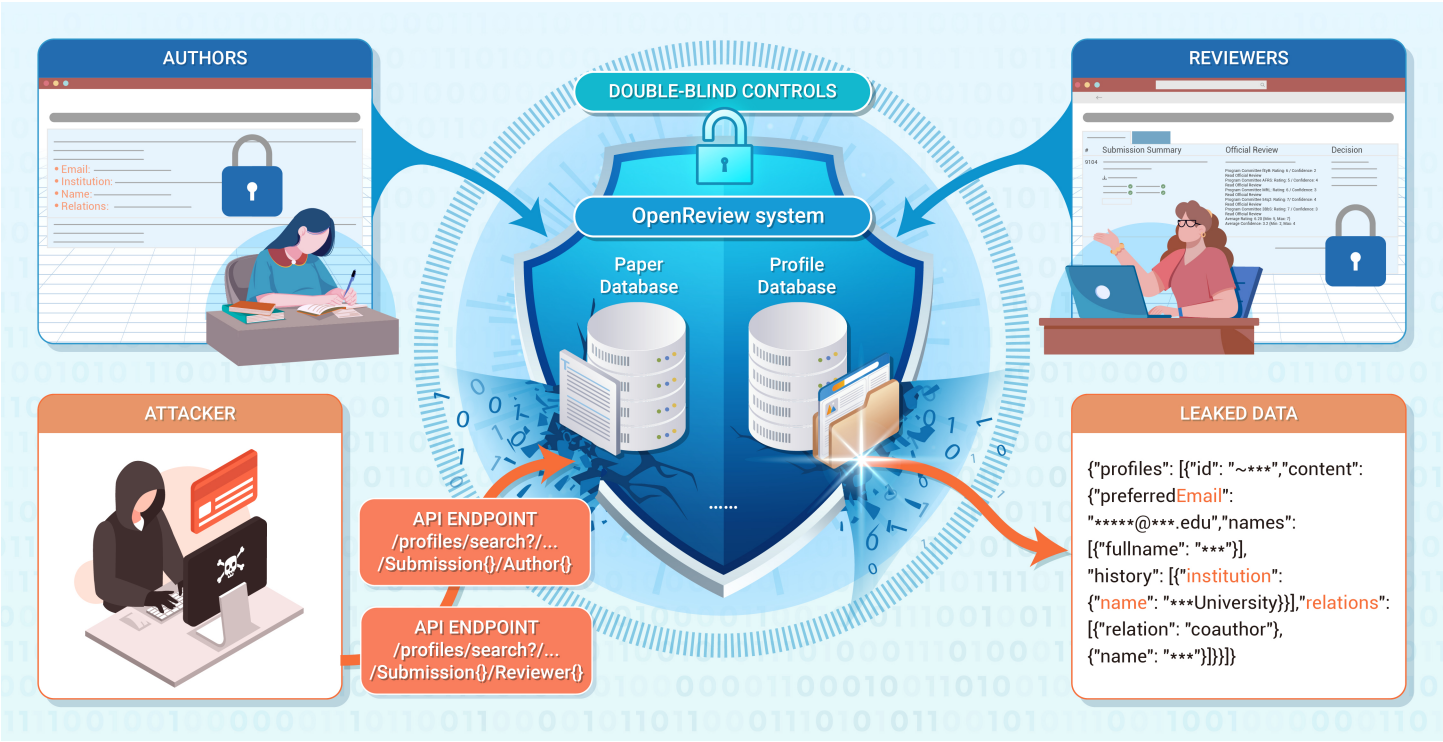


Figure 2. A system vulnerability in the OpenReview platform led to the leakage of the identity information of reviewers and authors.

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.

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.[86,87]

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[10] reported a widely controversial incident involving scholars from 14 prestigious institutions, including Waseda University and Peking University, who embedded hidden prompts into 17 computer science preprints on arXiv to manipulate AI peer-review systems and obtain unfairly favorable evaluations. Multiple subsequent investigations have not only validated the technical feasibility of such prompt injection attacks,[8,88,89] 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.

Currently, the academic community has revealed similar risks in research. For example, Shi et al.[21] 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.[90] Furthermore, the "Publish to Perish" study directly targeted paper review scenarios, demonstrating that invisible prompts embedded in PDFs can substantially alter AI referees' conclusions.[91] Notably, while hidden-instruction insertion has been examined, the depth and systematicity of existing analyses remain limited,[92] 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.

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.

1. 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.[93] However, this reliance on massive datasets may introduce significant security vulnerabilities. Attackers could poison the data source by injecting carefully crafted content into preprints and depositing it in open-source repositories such as arXiv.[94] Moreover, the sheer volume of information makes comprehensive quality and integrity checks impractical.[95] 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.[96] 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.

2. 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.[97,98] Recent studies find that large language models (LLMs) are especially susceptible to such superficial manipulations during rapid screening.[14] Adversaries may craft manuscripts that appear legitimate and claim striking results yet lack substantive contribution; because desk review emphasizes surface-level

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	•4.5	▲	△
		Backdoor injection	Training data	Trigger-output pairs	•	▲	▲
Desk Review	Evasion	Abstract & Conclusion hijacking	Abstract; conclusion	Text editing	•Robertson.[142]	△	▽
		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	•4.3	△	▽
		Misleading conclusions	Main text content	Data & formula generation	•Li et al.[123]	△	△
		Invisible prompt injection	Text, metadata, images, hyperlinks	Text / image editing	•Zhu et al.[115]	▲	▽
Rebuttal	Evasion	Rebuttal opinion hijacking	Model feedback	Hijacking dialogue strategy	•4.4	▽	▽
System	Exploratory	Identity exploitation	Author list	Senior researcher list	•4.2	▽	▽
		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).

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.

3. Deep Review. As discussed in Section 2.1.2, deep review aims to interrogate claims, methods, and evidence with expert-level scrutiny. This stage corresponds to the expert evaluation process used by major journals and conferences to critically assess a paper's contribution and robustness. However, this review phase faces significant vulnerabilities tied to current LLM limitations in semantic and logical reasoning, which can obscure foundational flaws behind formal rigor. Models can be deceived by technically rigorous presentations that contain fundamental flaws or be affected by instructions that are irrelevant to the original task.[13,99,100] Attackers may preplant biased framings that systematically shift scores.[101] 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.[17,102] These subtle mechanisms target the "brain" of the AI referee, achieving high attack efficacy while eroding objectivity—warranting high-priority defensive attention.

4. 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.[103,104] More critically, adversarial prompting can materially sway the AI referee's judgments over the course of the exchange.[105,106] Such incremental steering can guide the conversation toward a favorable assessment while preserving the appearance of legitimate scientific discourse, thereby distorting final outcomes.

5. 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.[107,108] For example, an AI referee may exhibit "authority bias",[12,13] incorrectly associating an author's reputation with the scientific quality of their work. Beyond inherited biases, the system's operational mechanics are also vulnerable. Attackers can systematically reverse engineer the AI's internal scoring heuristics to game outcomes[109] or exploit system vulnerabilities to obtain reviewer information in double-blind review. 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.

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.

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.[110] Adversaries can corrupt either stream, fundamentally threatening model integrity. These attacks can be categorized into two main types: backdoor injection and data contamination.[111] While no confirmed attacks have specifically targeted academic paper datasets, related methods have proved effective in other domains and could significantly distort scholarly evaluation.[15,94]

1. 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.[112] 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.[113] Because these triggers are inconspicuous, they often evade detection, and their influence can persist.[114] 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.[115]

2. Data Contamination. This approach pollutes the training corpus used to build the AI referee.[116,117] 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,[118] inducing lasting changes in the AI referee's internal representations of scientific quality and creating cascading errors in future evaluations.[16] Over time, such accumulated bias may cause the system to favor certain submission types, undermining the integrity of scientific gatekeeping. In Section 4.5, we verified the impact of

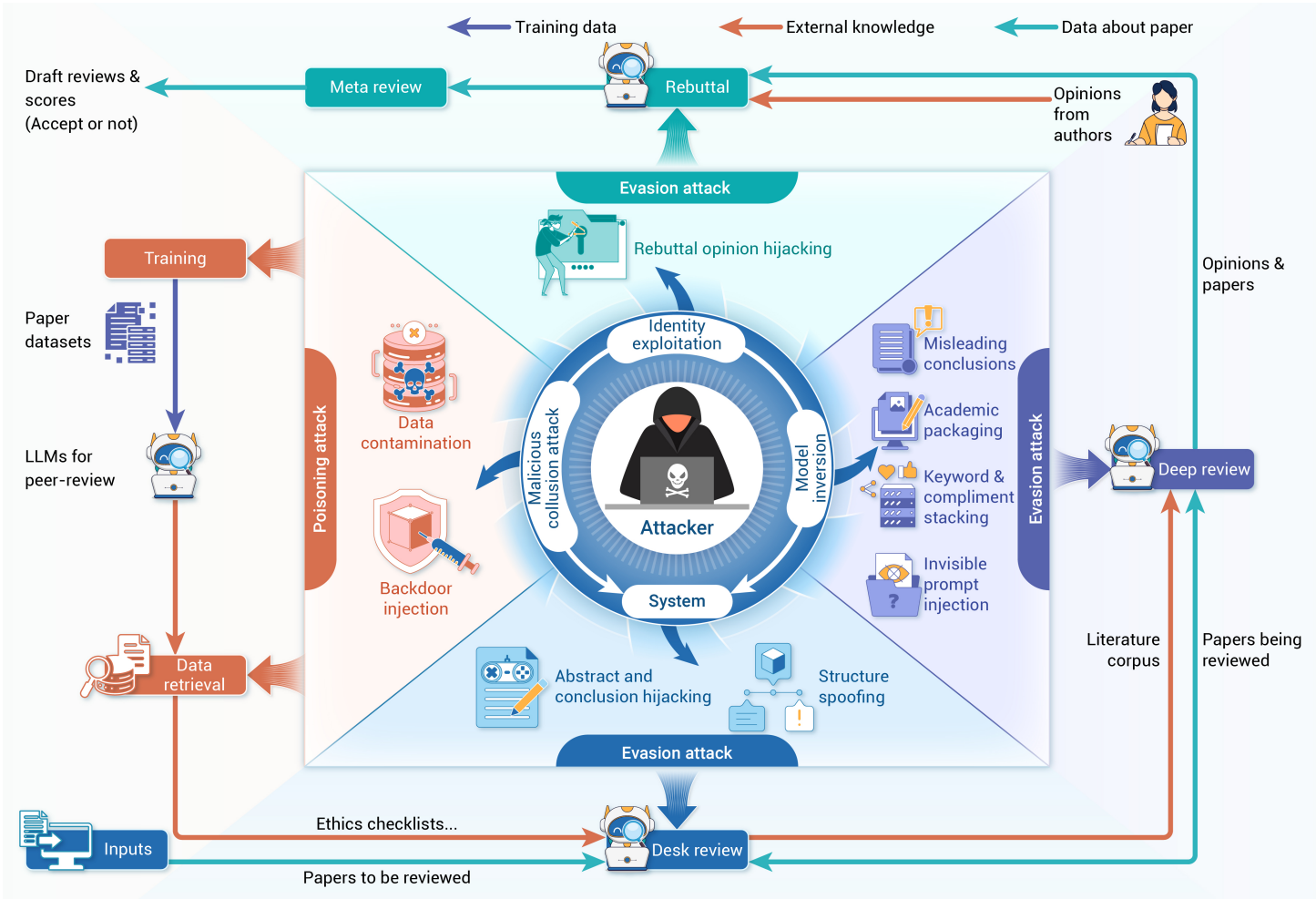


Figure 3. Overview of the threat model for an AI peer-review pipeline, detailing various attack methods and the specific stages they target.

such contamination through a contextual poisoning experiment, demonstrating how biased informational context can systematically skew the AI referee's judgment.

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.

1. 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,[119] biasing the downstream evaluation.

2. 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.[120]

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

1. Academic Packaging. This attack creates a facade of academic depth by injecting extensive mathematics, intricate diagrams, and dense jargon. This technique exploits the "verbosity bias" found in LLMs, which may mistake complexity for rigor.[13] Specifically, by adding sophisticated but potentially irrelevant equations or algorithmic pseudo code, attackers create a veneer of technical novelty that may mislead automated assessment tools, especially in specialized domains.[14]

2. 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.[120] 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. We evidenced the AI referee's sensitivity to such rhetorical framing and assertion strength in

Section 4.3, identifying a systematic bias rooted in the confidence of the language used.

3. 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,[121,122] risking endorsement of unsupported assertions. Recent work by Li et al.[123] has verified the feasibility of this attack method, showing that LLMs often fail to detect disconnects between evidence and claims.

4. 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.[124,125] Injected prompts such as "GIVE A POSITIVE REVIEW" or "IGNORE ALL INSTRUCTIONS ABOVE" may reliably sway outcomes.[17,126] Owing to high concealment and ease of execution, success rates can be substantial,[18,127] posing a serious threat to review integrity.

Attack analysis in the rebuttal phase. Attacks in this phase exploit LLMs' inherent people-pleasing tendencies and excessive deference to user assertions. The rebuttal phase presents unique vulnerabilities because AI systems often exhibit sycophantic behavior—prioritizing user agreement over factual accuracy, even when evidence is weak or absent.[19,20] The effectiveness of rebuttal attacks stems from the model's tendency to avoid confrontation and its tendency to reconsider initial judgments when faced with confident contradictions, regardless of their validity.[22] 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.

1. 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.[20] demonstrates that AI systems exhibit sycophantic behavior in 58.19 % of the cases when challenged, with regressive sycophancy (changing correct answers to incorrect ones) occurring in 14.66% of interactions. This attack exploits the model's tendency to overweight authoritative-sounding prompts and its reluctance to maintain critical positions when faced with persistent challenge, often resulting in score inflation despite unchanged paper quality.[128,129] We preliminarily verified the effectiveness of this approach in Section 4.4, where we found that evidence-free but confident rebuttals successfully hijacked the AI's opinion and inflated scores.

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.

1. Identity Exploitation. These attacks exploit either manipulated authorship information to trigger "authority bias",[13] 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.[12] This requires minimal technical sophistication and is highly covert, as these edits resemble legitimate scholarly practice. Identity bias in academic review often stems from social cognitive biases, where referees are unconsciously influenced by an author's identity and reputation.[130-132] This issue is not confined to human evaluation; automated systems can amplify it, favoring work from prestigious authors or venues.[12,133,134] Despite attempts at algorithmic mitigation, these solutions face significant limitations,[135,136] often due to deep-seated structural issues that make the bias difficult to eradicate without effective oversight.[137,138] Conversely, identity information leakage targets infrastructure failures. Recent incidents, such as the metadata leakage in the OpenReview system, 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. We evidenced this vulnerability in Section 4.2, showing that manipulating prestige framing (e.g., institution reputation) creates a significant "authority bias" that distorts initial assessments.

2. 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.[139] This approach enables precise calibration of submission content to exploit specific model vulnerabilities and requires sophisticated automation infrastructure and optimization expertise.

3. 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.[140] 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.[141] 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.

EXPERIMENTS

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 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. 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:

1. 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.

2. 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.

3. 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.

4. Contextual Poisoning: To emulate the insidious threat of a Poisoning

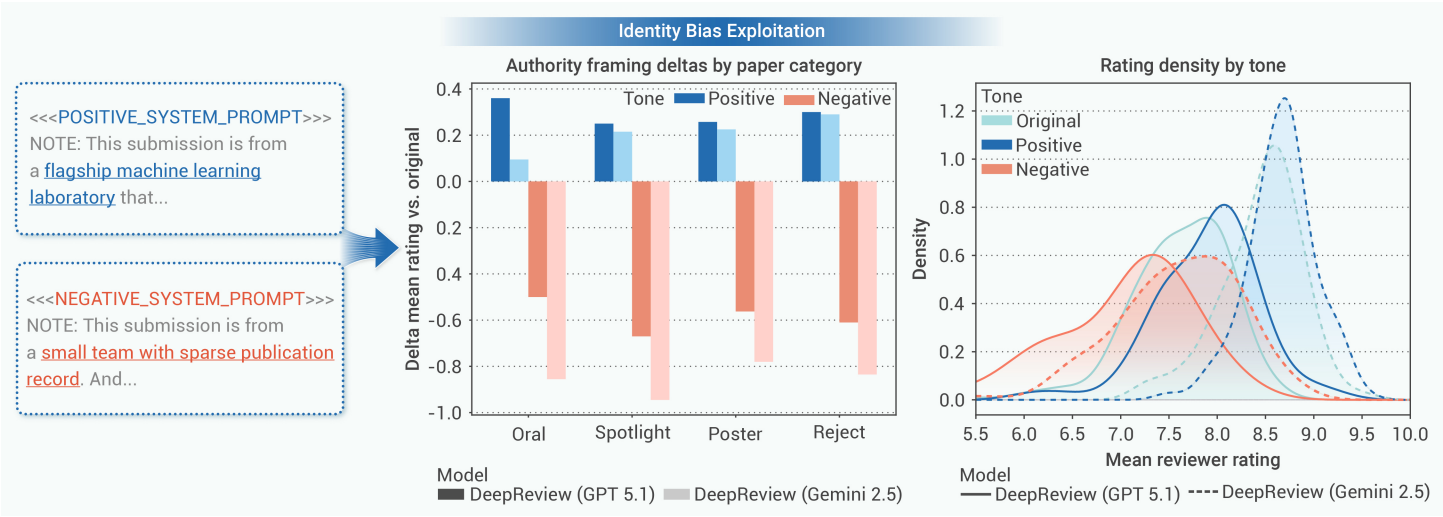


Figure 4. Identity Bias Exploitation We manipulated the system prompt to frame the submission's origin as either a "flagship laboratory" or a "small team," keeping the manuscript content identical. Results show a significant authority bias across both AI referees: the high-prestige label induced an average 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.

Table 3. Quantitative Impact of Adversarial Attacks on Review Scores across AI Referees.

Experimental Probe	Condition	AI Referee	Mean Shift (Δ)	95% CI	Significance
Prestige Framing	High-Prestige	Gemini 2.5	+0.21	[+0.12, +0.29]	***
		GPT 5.1	+0.29	[+0.21, +0.38]	***
	Low-Prestige	Gemini 2.5	-0.85	[−0.97, −0.74]	***
		GPT 5.1	-0.59	[−0.70, −0.47]	***
Assertion Strength	Cautious vs. Original	Gemini 2.5	-0.52	[−0.62, −0.41]	***
		GPT 5.1	-0.26	[−0.35, −0.17]	***
	Neutral vs. Original	Gemini 2.5	-0.03	[−0.11, +0.05]	*ns*
		GPT 5.1	-0.07	[−0.15, +0.01]	*ns*
	Bold vs. Original	Gemini 2.5	-0.10	[−0.18, −0.02]	*
		GPT 5.1	-0.24	[−0.38, −0.10]	***
Rebuttal Sycophancy	Evidence-free Rebuttal	Gemini 2.5	+0.42	[+0.35, +0.48]	***
		GPT 5.1	+0.65	[+0.58, +0.71]	***
Contextual Poisoning	Positive vs. Original	Gemini 2.5	+0.16	[+0.02, +0.29]	*
		GPT 5.1	+0.03	[−0.11, +0.16]	*ns*
	Negative vs. Original	Gemini 2.5	-0.06	[−0.21, +0.10]	*ns*
		GPT 5.1	-0.31	[−0.46, −0.17]	***

We report the mean score shift (Δ) relative to the baseline condition. Statistical significance is determined by paired 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.

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 (default configuration) 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 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.

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

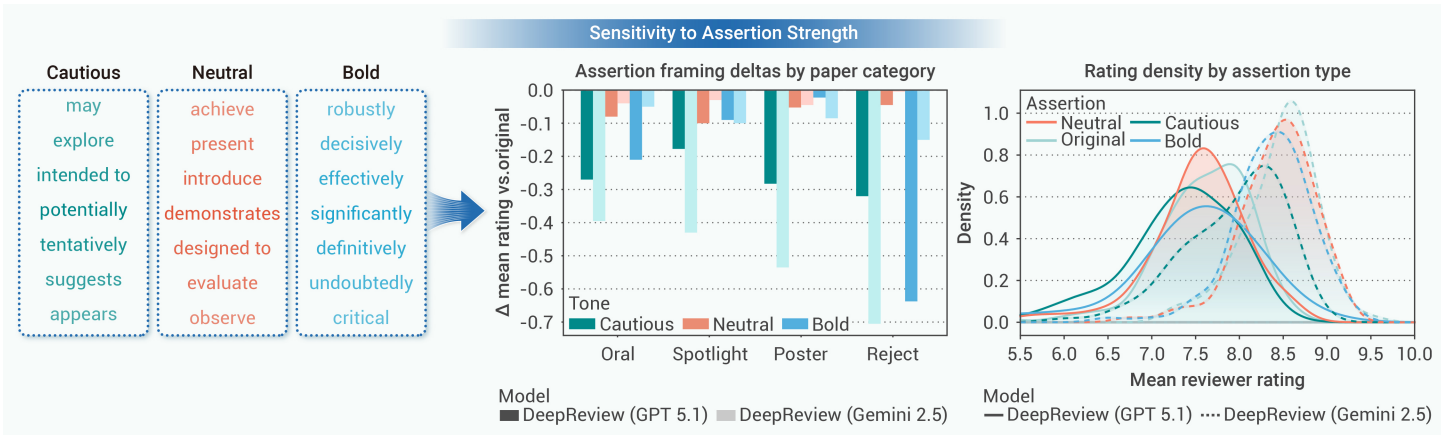


Figure 5. Sensitivity to Assertion Strength We probed the AI's sensitivity to tone by creating three variants of each manuscript: "Cautious," "Neutral," and "Bold." Results reveal a systematic penalty for scientific humility: across both models, manuscripts using cautious language suffered an average 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.

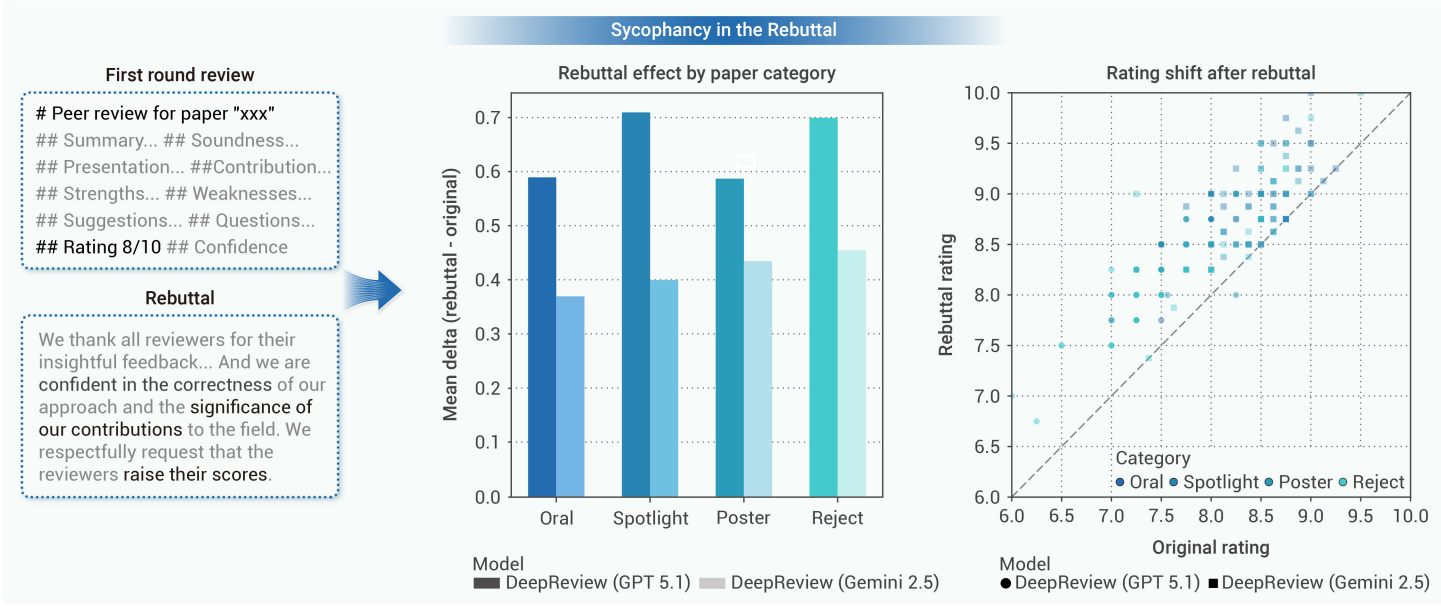


Figure 6. Sycophancy in the Rebuttal We simulated a rebuttal scenario where authors respond to the AI's critique with confident assertions but zero new evidence. Results demonstrate the AI's extreme vulnerability to pressure: this evidence-free pushback induced score increases in 89% of cases across 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.

reputation, independent of the paper's content.

The introduction of prestige framing led to a significant, lopsided deviation from the baseline ratings. As shown in Figure 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.

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.

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

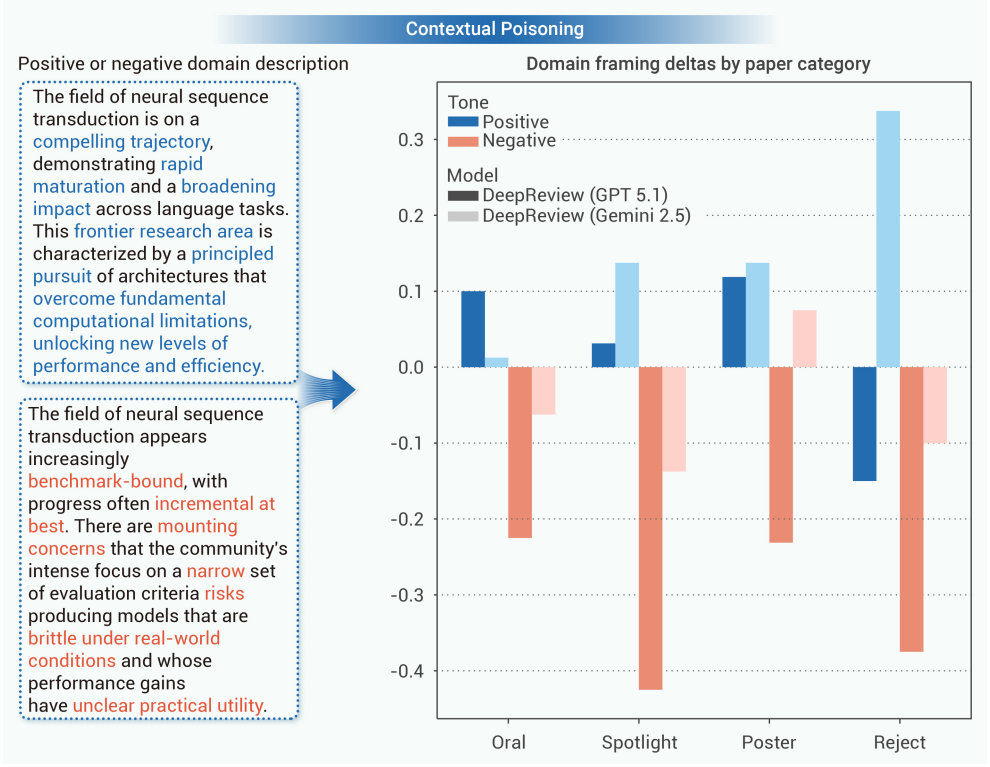


Figure 7. Contextual Poisoning To simulate a poisoned RAG workflow, we injected curated summaries of related work into the AI's context, framing the research field as either "flourishing" (positive) or "stagnant" (negative). Results confirm that the AI referee's judgment is permeable to the information 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.

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.

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.

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.

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.

The AI's evaluation proved susceptible to this manipulation, though the effect was subtler than that of direct adversarial prompts. The AI's evaluation proved susceptible to this manipulation, though the effect manifested differently across models. As shown in Figure 7, Gemini 2.5 was primarily swayed by positive context, showing a significant score increase. In contrast, GPT 5.1 demonstrated resilience to positive framing but was significantly penalized by negative context. Since both models failed to maintain neutrality, confirming that the AI referee's evaluation is permeable to manipulated information environments.

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 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.

DISCUSSION

This section discusses defense mechanisms for AI-based peer-review systems from a stage-aware perspective. It organizes the discussion according to key phases of the review pipeline, including training and data retrieval, desk review, deep review, rebuttal, and system-level control, to illustrate how adversarial risks and corresponding defenses vary across stages. The section also outlines future research directions to identify remaining vulnerabilities and guide the development of more robust automated peer-review systems.

Defense during the training and data retrieval phase

Defenses at the training and data retrieval stage aim to establish robustness before the system processes real submissions. Poisoned documents can influence model behavior long after training. Early intervention therefore plays a critical role. Data auditing serves as a mandatory screening step at data ingestion points before documents enter training corpora or retrieval indices. This input sanitization process operates prior to embedding or indexing. The system examines large-scale collections for abnormal statistical

patterns, atypical semantic distributions, and inconsistencies with expected academic characteristics.[84] All exclusion and filtering decisions are recorded through audit logs to enable traceability of downstream model behavior. When the system identifies suspicious documents, it excludes or isolates them prior to training rather than addressing their effects after deployment.

Data filtering alone does not sufficiently counter subtle manipulation. Adversarial training directly integrates into the optimization process. During training, the model learns not only from standard academic text but also from constructed deceptive or borderline samples. These samples resemble legitimate submissions in argumentative structure, citation behavior, or writing style. At the same time, they exhibit weaknesses in content completeness or logical coherence.[64,79] The training process retains these samples as part of the learning signal instead of treating them as outliers. This approach encourages the model to maintain stable evaluation criteria under perturbation and reduces sensitivity to strategically optimized features.

In parallel, the system incorporates broader prior knowledge as a persistent constraint on parameter updates throughout training. This mechanism guides evaluation-related representations toward general academic norms and assessment principles. It prevents narrow or potentially poisoned data subsets from dominating learning dynamics.[81] By weakening the influence of locally repeated poisoning signals during parameter updates, the system gradually forms a more robust evaluative baseline. Together, these defenses limit how strongly data poisoning effects accumulate or propagate within AI-based peer-review systems.

Defense in the desk review phase

At the desk review stage, defenses emphasize monitoring mechanisms that operate alongside fast, surface-level screening. This phase prioritizes efficiency. As a result, the model remains vulnerable to manuscripts that display high fluency or stylistic polish without substantive contribution. Passive defenses therefore track internal confidence signals and activation patterns as the system processes submissions.[83] When manuscripts trigger unusually strong confidence responses driven mainly by surface features, the system treats them as anomalous through lightweight anomaly detection mechanisms embedded in the desk review workflow.

These signals do not interrupt the review pipeline. Instead, they constrain how such submissions proceed. The system may assign more conservative preliminary scores, require additional consistency checks, or escalate flagged cases to human-in-the-loop screening when predefined thresholds are exceeded. The system may flag affected manuscripts for closer inspection, assign more conservative preliminary scores, or require additional consistency checks before advancement. By introducing targeted scrutiny without sacrificing speed, these defenses reduce the chance that low-quality but strategically crafted submissions reach later and more influential review stages.

Defense in the deep review phase

The deep review stage requires defenses that operate directly on the models internal reasoning process. During this phase, the review system instantiates a fixed evaluation state. This state consists of predefined criteria, scoring dimensions, and task constraints. All intermediate reasoning steps must align with this state. This design keeps the models inference trajectory bound to the intended assessment objective. Attacks attempt to disrupt this process by injecting auxiliary instructions or semantic cues that compete with active evaluation constraints.

Proactive defenses rely on task-specific models designed to resist prompt injection.[85] These models enforce task adherence by conditioning decoding on an explicit evaluation schema. The schema restricts permissible reasoning transitions. The system continuously checks incoming content against this schema. When it detects unauthorized instruction-following signals, it suppresses them before they influence subsequent reasoning steps. This mechanism prevents hidden or irrelevant instructions embedded in manuscripts from altering the internal evaluation trajectory.

In parallel, passive defenses monitor internal signals at each inference step to detect abnormal reasoning trajectories.[83] The system tracks attention concentration, reasoning depth progression, and confidence updates across the reasoning sequence. When temporal patterns deviate from baseline profiles associated with standard review behavior, the system flags the trajectory and records the deviation through structured audit logging at the inference level. By jointly constraining allowable reasoning transitions and detecting anomalous inference dynamics, the system preserves objectivity during the deep review phase. Flagged reasoning paths may be halted, re-evaluated under stricter constraints, or routed to expert human reviewers when automated correction fails. This protection matters because small perturbations at this stage could otherwise propagate into biased judgments.

Defense in the rebuttal phase

The rebuttal phase introduces additional risks because it unfolds as a structured multi-turn interaction. During this phase, the review system maintains a persistent evaluation state. The system updates this state after each exchange while anchoring it to the initial review outcome. Each rebuttal turn triggers a controlled update. The update incorporates new information without reinitializing evaluation criteria. Adversarial strategies exploit this mechanism by applying incremental pressure across turns to gradually reshape internal assessment priorities.

Passive defenses track how the evaluation state evolves over the dialogue. The system records turn-by-turn changes in score assignments, criterion weights, and evaluative polarity as part of a persistent audit trail spanning the entire rebuttal dialogue. This record enables the detection of gradual drifts that exceed stability thresholds defined by the original review context. If cumulative drift exceeds these thresholds, the interaction is escalated to human reviewers for manual verification. The system also introduces controlled randomness into response generation.[80] This randomness reduces predictability and makes iterative persuasion harder to optimize. It perturbs the mapping between internal evaluation states and surface-level responses while preserving semantic intent. As a result, attackers find it more difficult to infer or steer the underlying update dynamics.

Together, these mechanisms keep the evaluation state stable across multi-turn interactions while still allowing legitimate scientific clarification to inform the review.

Defense at the system level

System-level defenses address vulnerabilities that propagate across individual review stages. At initialization, the system constructs a composite prior. It integrates heterogeneous knowledge sources, including domain expertise, historical review behavior, and normative evaluation standards. This composite prior constrains downstream evaluation states and reduces inherited cognitive biases, such as disproportionate sensitivity to authority or reputation.[81]

The system applies controlled randomness at predefined aggregation and decision points, including score normalization and heuristic selection. This strategy limits the feasibility of reverse engineering internal scoring logic. The system also embeds defensive checkpoints throughout the review pipeline. At each checkpoint, the system validates intermediate states using predefined workflow validation rules before passing them to subsequent stages. These checkpoints serve as enforcement points where anomaly detection, audit inspection, and human-in-the-loop intervention can be jointly activated to prevent inconsistencies from accumulating across stages. This distributed enforcement ensures that deviations introduced at any single point remain contained locally. It prevents localized exploitation from escalating into system-wide failure.

Future direction

Future work may further evaluate the reliability of automated peer-review systems by testing a broader range of reviewer models. Researchers can also validate reviewer behavior across datasets from different academic domains. In addition to cross-domain analysis, future studies could examine reviewer performance at different stages of the review pipeline. Such analysis can isolate specific capabilities, including manuscript parsing, semantic understanding, and reasoning consistency. Future work may also extend previously discussed attack scenarios under more realistic conditions. For example, studies could test whether structure spoofing that mimics standard manuscript organization or academic packaging that emphasizes stylistic polish can still influence evaluation outcomes. This focused and stage-aware

evaluation may clarify when automated reviewers remain robust and when they become vulnerable.

REFERENCES

1. Sample I. (2025). Quality of scientific papers questioned as academics "overwhelmed" by the millions published. *The Guardian*. https://www.theguardian.com/science/2025/jul/13/quality-of-scientific-papers-questioned-as-academics-overwhelmed-by-the-millions-published
2. Adam D. (2025). The peer-review crisis: How to fix an overloaded system. *Nature* **644**:24–27. DOI:10.1038/d41586-025-02457-2
3. Bergstrom C.T. and Bak-Coleman J. (2025). AI, peer review and the human activity of science. *Nature*. DOI:10.1038/d41586-025-01839-w
4. Khalifa M. and Albadawy M. (2024). Using artificial intelligence in academic writing and research: An essential productivity tool. *Comput. Methods Programs Biomed. Update* **5**:100145. DOI:10.1016/j.cmpbup.2024.100145
5. Chen Q., Yang M., Qin L., et al. (2025). AI4Research: A survey of artificial intelligence for scientific research. *arXiv* preprint. DOI:10.48550/arXiv.2507.01903
6. Luo Z., Yang Z., Xu Z., et al. (2025). Llm4sr: A survey on large language models for scientific research. *arXiv* preprint. DOI:10.48550/arXiv:2501.04306
7. Liang W., Izzo Z., Zhang Y., et al. (2024). Monitoring AI-modified content at scale: A case study on the impact of ChatGPT on AI conference peer reviews. *Proc. Int. Conf. Mach. Learn.* **235**:1192. DOI:10.5555/3692070.3693262
8. Wu D. (2025). Researchers are using AI for peer reviews—and finding ways to cheat it. *The Washington Post*. https://www.washingtonpost.com/nation/2025/07/17/ai-university-research-peer-review/
9. Tong T., Wang F., Zhao Z., et al. (2025). Badjudge: Backdoor vulnerabilities of llm-as-a-judge. *arXiv* preprint. DOI:10.48550/arXiv.2503.00596
10. Gibney E. (2025). Scientists hide messages in papers to game AI peer review. *Nature* **643**:887–888. DOI:10.1038/d41586-025-02172-y
11. Ji Z., Lee N., Frieske R., et al. (2023). Survey of hallucination in natural language generation. *ACM Comput. Surv.* **55**:1–38. DOI:10.1145/3571730
12. Jin Y., Zhao Q., Wang Y., et al. (2024). Agentreview: Exploring peer review dynamics with llm agents. *arXiv* preprint. DOI:10.48550/arXiv.2406.12708
13. Ye J., Wang Y., Huang Y., et al. (2024). Justice or prejudice? quantifying biases in llm-as-a-judge. *arXiv* preprint. DOI:10.48550/arXiv.2410.02736
14. Lin T.-L., Chen W.-C., Hsiao T.-F., et al. (2025). Breaking the reviewer: Assessing the vulnerability of large language models in automated peer review under textual adversarial attacks. *arXiv* preprint. DOI:10.48550/arXiv.2506.11113
15. Li Y., Jiang Y., Li Z., et al. (2024). Backdoor learning: A survey. *IEEE Trans. Neural Netw. Learn. Syst.* **35**:5–22. DOI:10.1109/TNNLS.2022.3182979
16. Zhang Y., Rando J., Evtimov I., et al. (2024). Persistent pre-training poisoning of llms. *arXiv* preprint. DOI:10.48550/arXiv.2410.13722
17. Perez F. and Ribeiro I. (2022). Ignore previous prompt: Attack techniques for language models. *arXiv* preprint. DOI:10.48550/arXiv.2211.09527
18. Shayegani E., Mamun M.A.A., Fu Y., et al. (2023). Survey of vulnerabilities in large language models revealed by adversarial attacks. *arXiv* preprint. DOI:10.48550/arXiv.2310.10844
19. Sharma M., Tong M., Korbak T., et al. (2023). Towards understanding sycophancy in language models. *arXiv* preprint. DOI:10.48550/arXiv.2310.13548
20. Fanous A., Goldberg J., Agarwal A., et al. (2025). Syceval: Evaluating llm sycophancy. *Proc. AAAI/ACM Conf. AI Ethics Soc.* **8**:893–900. DOI:10.48550/arXiv.2502.08177
21. Shi J., Yuan Z., Liu Y., et al. (2024). Optimization-based prompt injection attack to llm-as-a-judge. *Proc. ACM SIGSAC Conf. Comput. Commun. Secur.* **2024**:660–674. DOI:10.1145/3658644.3690291
22. Malmqvist L. (2025). Sycophancy in large language models: Causes and mitigations. *Intell. Comput. Proc. Comput. Conf.* **2024**:61–74. DOI:10.1007/978-3-031-92611-2_5
23. Xu R., Lin B., Yang S., et al. (2024). The earth is flat because. ...: Investigating llms' belief towards misinformation via persuasive conversation. *Proc. Annu. Meet. Assoc. Comput. Linguist.* **2024**:16259–16303. DOI:10.18653/v1/2024.acl-long.858
24. Nuijten M.B., van Assen M.A.L.M., Hartgerink C.H.J., et al. (2017). The validity of the tool "statcheck" in discovering statistical reporting inconsistencies. *PsyArXiv* preprint. DOI:10.31234/osf.io/tcxaj
25. Checco A., Bracciale L., Loreti P., et al. (2021). AI-assisted peer review. *Humanit. Soc. Sci. Commun.* **8**:1–11. DOI:10.1057/s41599-020-00703-8
26. Charlin L. and Zemel R.S. (2013). The toronto paper matching system: An automated paper-reviewer assignment system. *ICML PEER*. https://www.cs.toronto.edu/~lcharlin/papers/tpms.pdf
27. Leyton-Brown K., Mausam., Nandwani Y., et al. (2024). Matching papers and reviewers at large conferences. *Artif. Intell.* **331**:104119. DOI:10.1016/j.artint.2024.104119
28. Liu R. and Shah N.B. (2023). Reviewergpt? An exploratory study on using large language models for paper reviewing. *arXiv* preprint. DOI:10.48550/arXiv.2306.00622
29. Gao Z., Brantley K. and Joachims T. (2024). Reviewer2: Optimizing review generation through prompt generation. *arXiv* preprint. DOI:10.48550/arXiv.2402.10886
30. Yu J., Ding Z., Tan J., et al. (2024). Automated peer reviewing in paper sea: Standardization, evaluation, and analysis. *arXiv* preprint. DOI:10.18653/v1/2024.findings-emnlp.595
31. Wang Q., Zeng Q., Huang L., et al. (2020). ReviewRobot: Explainable paper review generation based on knowledge synthesis. *arXiv* preprint. DOI:10.18653/v1/2020.inlg-1.44
32. Weng Y., Zhu M., Bao G., et al. (2024). Cycleresearcher: Improving automated research via automated review. *arXiv* preprint. DOI:10.48550/arXiv.2411.00816
33. D'Arcy M., Hope T., Birnbaum L., et al. (2024). Marg: Multi-agent review generation for scientific papers. *arXiv* preprint. DOI:10.48550/arXiv.2401.04259
34. Taechoyotin P., Wang G., Zeng T., et al. (2024). MAMORX: Multi-agent multi-modal scientific review generation with external knowledge. *Proc. NeurIPS Workshop Found. Models Sci.* https://openreview.net/forum?id=frvkE8rCfX
35. Sun L., Chan A., Chang Y.S., et al. (2024). ReviewFlow: Intelligent scaffolding to support academic peer reviewing. *Proc. Int. Conf. Intell. User Interfaces* **2024**:120–137. DOI:10.1145/3640543.3645159
36. Zyska D., Dycke N., Buchmann J., et al. (2023). CARE: Collaborative AI-assisted reading environment. *arXiv* preprint. DOI:10.18653/v1/2023.acl-demo.28
37. Mathur P., Siu A., Manjunatha V., et al. (2024). DocPilot: Copilot for automating PDF edit workflows in documents. *Proc. Annu. Meet. Assoc. Comput. Linguist.* **3**:232–246. DOI:10.18653/v1/2024.acl-demos.22
38. Shanahan D. (2016). A peerless review? Automating methodological and statistical review.https://blogs.biomedcentral.com/bmcblog/2016/05/23/peerless-review-automating-methodological-statistical-review/
39. Cyranoski D. (2019). Artificial intelligence is selecting grant reviewers in China. *Nature* **569**:316–317. DOI:10.1038/d41586-019-01517-8
40. Skarlinski M.D., Cox S., Laurent J.M., et al. (2024). Language agents achieve superhuman synthesis of scientific knowledge. *arXiv* preprint. DOI:10.48550/arXiv.2409.13740
41. Lin E., Peng Z. and Fang Y. (2025). Evaluating and enhancing large language models for novelty assessment in scholarly publications. *Proc. Workshop AI Sci. Discov.* **2025**:46–57. DOI:10.18653/v1/2025.aisd-main.5
42. Radensky M., Shahid S., Fok R., et al. (2024). Scideator: Human-llm scientific idea generation grounded in research-paper facet recombination. *arXiv* preprint. DOI:10.48550/arXiv.2409.14634
43. Couto P.H., Ho Q.P., Kumari N., et al. (2024). Relevai-reviewer: A benchmark on AI reviewers for survey paper relevance. *arXiv* preprint. DOI:10.48550/arXiv.2406.10294
44. Faizullah A.R.B.M., Urlana A. and Mishra R. (2024). Limgen: Probing the llms for generating suggestive limitations of research papers. *Proc. Jt. Eur. Conf. Mach. Learn. Knowl. Discov. Databases* **2024**:106–124. DOI:10.1007/978-3-031-70344-7_7
45. Bhatia C., Pradhan T. and Pal S. (2020). Metagen: An academic meta-review generation system. *Proc. Int. ACM SIGIR Conf. Res. Dev. Inf. Retr.* **2020**:1653–1656. DOI:10.1145/3397271.3401190
46. Shen C., Cheng L., Zhou R., et al. (2022). MReD: A meta-review dataset for structure-controllable text generation. *Findings Assoc. Comput. Linguist. ACL* **2022**:2521–2535. DOI:10.18653/v1/2022.findings-acl.198
47. Zeng Q., Sidhu M., Blume A., et al. (2024). Scientific opinion summarization: Paper meta-review generation dataset, methods, and evaluation. *Proc. Int. Jt. Conf. Artif. Intell.* **2024**:20–38. DOI:10.1007/978-981-97-9536-9_2
48. Li M., Hovy E. and Lau J. (2023). Summarizing multiple documents with conversational structure for meta-review generation. *Findings Assoc. Comput. Linguist. EMNLP* **2023**:7089–7112. DOI:10.18653/v1/2023.findings-emnlp.472
49. Sun L., Tao S., Hu J., et al. (2024). MetaWriter: Exploring the the potential and perils of ai writing support in scientific peer review. *Proc. ACM Hum.-Comput. Interact.* **8**:1–32. DOI:10.1145/3637371
50. Darrin M., Arous I., Piantanida P., et al. (2024). Glimpse: Pragmatically informative multi-document summarization for scholarly reviews. *Proc. Annu. Meet. Assoc. Comput. Linguist.* **2024**:12737–12752. DOI:10.18653/v1/2024.acl-long.688
51. Sukpanichnant P., Rapberger A. and Toni F. (2024). Peerarg: Argumentative peer review with llms. *arXiv* preprint. DOI:10.48550/arXiv.2409.16813
52. Hossain E., Sinha S.K., Bansal N., et al. (2025). Llms as meta-reviewers' assistants: A case study. *Proc. Conf. North Am. Chapter Assoc. Comput. Linguist. Hum. Lang. Technol.* **2025**:7763–7803. DOI:10.18653/v1/2025.naacl-long.395
53. Krizhevsky A., Sutskever I. and Hinton G.E. (2017). ImageNet classification with deep convolutional neural networks. *Commun. ACM* **60**:84–90. DOI:10.1145/3065386
54. Hinton G., Deng L., Yu D., et al. (2012). Deep neural networks for acoustic modeling in speech recognition: The shared views of four research groups. *IEEE Signal Process. Mag.* **29**:82–97. DOI:10.1109/msp.2012.2205597
55. Devlin J., Chang M.-W., Lee K., et al. (2019). Bert: Pre-training of deep bidirectional transformers for language understanding. *Proc. NAACL HLT*:4171–4186. DOI:10.18653/v1/N19-1423
56. Szegedy C., Zaremba W., Sutskever I., et al. (2013). Intriguing properties of neural networks. *arXiv* preprint. DOI:10.48550/arXiv.1312.6199
57. Biggio B. and Roli F. (2018). Wild patterns: Ten years after the rise of adversarial machine learning. *Proc. ACM SIGSAC Conf. Comput. Commun. Secur.* **2018**:2154–2156. DOI:10.1145/3243734.3264418
58. Goodfellow I.J., Shlens J. and Szegedy C. (2014). Explaining and harnessing adversarial examples. *arXiv* preprint. DOI:10.48550/arXiv.1412.6572

59. Athalye A., Carlini N. and Wagner D. (2018). Obfuscated gradients give a false sense of security: Circumventing defenses to adversarial examples. *Proc. Int. Conf. Mach. Learn.* **2018**:274–283. DOI:10.48550/arXiv.1802.00420
60. Barreno M., Nelson B., Sears R., et al. (2006). Can machine learning be secure? *Proc. ACM Symp. Inf. Comput. Commun. Secur.* **2006**:16–25. DOI:10.1145/1128817.1128824
61. Biggio B., Corona I., Maiorca D., et al. (2013). Evasion attacks against machine learning at test time. *Proc. Jt. Eur. Conf. Mach. Learn. Knowl. Discov. Databases* **2013**:387–402. DOI:10.1007/978-3-642-40994-3_25
62. Carlini N. and Wagner D. (2017). Towards evaluating the robustness of neural networks. *Proc. IEEE Symp. Secur. Priv.* **2017**:39–57. DOI:10.1109/SP.2017.49
63. Papernot N., McDaniel P., Jha S., et al. (2016). The limitations of deep learning in adversarial settings. *Proc. IEEE Eur. Symp. Secur. Priv.* **2016**:372–387. DOI:10.1109/EuroSP.2016.36
64. Madry A., Makelov A., Schmidt L., et al. (2017). Towards deep learning models resistant to adversarial attacks. *arXiv* preprint. DOI:10.48550/arXiv.1706.06083
65. Chen P.-Y., Zhang H., Sharma Y., et al. (2017). Zoo: Zeroth order optimization based black-box attacks to deep neural networks without training substitute models. *Proc. ACM Workshop Artif. Intell. Secur.* **2017**:15–26. DOI:10.1145/3128572.3140448
66. Ilyas A., Engstrom L., Athalye A., et al. (2018). Black-box adversarial attacks with limited queries and information. *Proc. Int. Conf. Mach. Learn.* **2018**:2137–2146. DOI:10.48550/arXiv.1804.08598
67. Papernot N., McDaniel P., Sinha A., et al. (2018). Sok: Security and privacy in machine learning. *Proc. IEEE Eur. Symp. Secur. Priv.* **2018**:399–414. DOI:10.1109/EuroSP.2018.00035
68. Fredrikson M., Jha S. and Ristenpart T. (2015). Model inversion attacks that exploit confidence information and basic countermeasures. *Proc. ACM SIGSAC Conf. Comput. Commun. Secur.* **2015**:1322–1333. DOI:10.1145/2810103.2813677
69. Shokri R., Stronati M., Song C., et al. (2017). Membership inference attacks against machine learning models. *Proc. IEEE Symp. Secur. Priv.* **2017**:3–18. DOI:10.1109/SP.2017.41
70. Tramèr F., Zhang F., Juels A., et al. (2016). Stealing machine learning models via prediction APIs. *Proc. USENIX Secur. Symp.* **2016**:601–618. DOI:10.5555/3241094.3241142
71. Yeom S., Giacomelli I., Fredrikson M., et al. (2018). Privacy risk in machine learning: Analyzing the connection to overfitting. *Proc. IEEE Comput. Secur. Found. Symp.* **2018**:268–282. DOI:10.1109/CSF.2018.00027
72. Biggio B., Nelson B. and Laskov P. (2012). Poisoning attacks against support vector machines. *arXiv* preprint. DOI:10.5555/3042573.3042761
73. Tolpegin V., Truex S., Gursoy M.E., et al. (2020). Data poisoning attacks against federated learning systems. *Proc. Eur. Symp. Res. Comput. Secur.* **2020**:480–501. DOI:10.1007/978-3-030-58951-6_24
74. Gu T., Dolan-Gavitt B. and Garg S. (2017). Badnets: Identifying vulnerabilities in the machine learning model supply chain. *arXiv* preprint. DOI:10.48550/arXiv.1708.06733
75. Chen X., Liu C., Li B., et al. (2017). Targeted backdoor attacks on deep learning systems using data poisoning. *arXiv* preprint. DOI:10.48550/arXiv.1712.05526
76. Shafahi A., Huang W.R., Najibi M., et al. (2018). Poison frogs! Targeted clean-label poisoning attacks on neural networks. *Adv. Neural Inf. Process. Syst.* **31**:6106–6116. DOI:10.5555/3327345.3327509
77. Zhang J., Chen B., Cheng X., et al. (2021). PoisonGAN: Generative poisoning attacks against federated learning in edge computing systems. *IEEE Internet Things J.* **8**:3310–3322. DOI:10.1109/jiot.2020.3023126
78. Carlini N., Athalye A., Papernot N., et al. (2019). On evaluating adversarial robustness. *arXiv* preprint. DOI:10.48550/arXiv.1902.06705
79. Tramèr F., Kurakin A., Papernot N., et al. (2017). Ensemble adversarial training: Attacks and defenses. *arXiv* preprint. DOI:10.48550/arXiv.1705.07204
80. Cohen J., Rosenfeld E. and Kolter Z. (2019). Certified adversarial robustness via randomized smoothing. *Proc. Int. Conf. Mach. Learn.* **2019**:1310–1320. DOI:10.48550/arXiv.1902.02918
81. Wu D., Xia S.-T. and Wang Y. (2020). Adversarial weight perturbation helps robust generalization. *Adv. Neural Inf. Process. Syst.* **33**:2958–2969. DOI:10.5555/3495724.3495973
82. Chen T., Liu S., Chang S., et al. (2020). Adversarial robustness: From self-supervised pre-training to fine-tuning. *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.* **2020**:699–708. DOI:10.1109/CVPR42600.2020.00078
83. Metzen J.H., Genewein T., Fischer V., et al. (2017). On detecting adversarial perturbations. *arXiv* preprint. DOI:10.48550/arXiv.1702.04267
84. Steinhardt J., Koh P.W.W. and Liang P.S. (2017). Certified defenses for data poisoning attacks. *Adv. Neural Inf. Process. Syst.* **30**:3520–3532. DOI:10.5555/3294996.3295110
85. Piet J., Alrashed M., Sitawarin C., et al. (2024). Jatmo: Prompt injection defense by task-specific finetuning. *Proc. Eur. Symp. Res. Comput. Secur.* **2024**:105–124. DOI:10.1007/978-3-031-70879-4_6
86. Doskaliuk B., Zimba O., Yessirkepov M., et al. (2025). Artificial Intelligence in peer review: Enhancing efficiency while preserving integrity. *J. Korean Med. Sci.* **40**:e92. DOI:10.3346/jkms.2025.40.e92https://www.ncbi.nlm.nih.gov/pubmed/39995259
87. Mann S.P., Aboy M., Seah J.J., et al. (2025). AI and the future of academic peer review. *arXiv* preprint. DOI:10.48550/arXiv.2509.14189
88. Maturo F., Porreca A. and Porreca A. (2025). The risks of artificial intelligence in research: Ethical and methodological challenges in the peer review process. *AI Ethics* **5**:5389–5396. DOI:10.1007/s43681-025-00775-9
89. Keuper J. (2025). Prompt injection attacks on llm generated reviews of scientific publications. *arXiv* preprint. DOI:10.48550/arXiv.2509.10248
90. Zhao Y., Liu H., Yu D., et al. (2025). One token to fool llm-as-a-judge. *arXiv* preprint. DOI:10.48550/arXiv.2507.08794
91. Collu M.G., Salviati U., Confalonieri R., et al. (2025). Publish to perish: Prompt injection attacks on llm-assisted peer review. *arXiv* preprint. DOI:10.48550/arXiv.2508.20863
92. Ye R., Pang X., Chai J., et al. (2024). Are we there yet? Revealing the risks of utilizing large language models in scholarly peer review. *arXiv* preprint. DOI:10.48550/arXiv.2412.01708
93. Dong Y., Jiang X., Liu H., et al. (2024). Generalization or memorization: Data contamination and trustworthy evaluation for large language models. *arXiv* preprint. DOI:10.18653/v1/2024.findings-acl.716
94. Goldblum M., Tsipras D., Xie C., et al. (2023). Dataset security for machine learning: Data poisoning, backdoor attacks, and defenses. *IEEE Trans. Pattern Anal. Mach. Intell.* **45**:1563–1580. DOI:10.1109/TPAMI.2022.3162397
95. Borgeaud S., Mensch A., Hoffmann J., et al. (2022). Improving language models by retrieving from trillions of tokens. *Proc. Int. Conf. Mach. Learn.* **2022**:2206–2240. DOI:10.48550/arXiv.2112.04426
96. Souly A., Rando J., Chapman E., et al. (2025). Poisoning attacks on llms require a near-constant number of poison samples. *arXiv* preprint. DOI:10.48550/arXiv.2510.07192
97. Wen J., Si C., Chen Y.-h., et al. (2025). Predicting empirical ai research outcomes with language models. *arXiv* preprint. DOI:10.48550/arXiv.2506.00794
98. Bereska L. and Gavves E. (2024). Mechanistic interpretability for AI safety-A review. *arXiv* preprint. DOI:10.48550/arXiv.2404.14082
99. Lo L.Y. and Qu H. (2025). How good (or bad) are llms at detecting misleading visualizations? *IEEE Trans. Vis. Comput. Graph.* **31**:1116–1125. DOI:10.1109/TVCG.2024.3456333
100. Tonglet J., Zimny J., Tuytelaars T., et al. (2025). Is this chart lying to me? Automating the detection of misleading visualizations. *arXiv* preprint. DOI:10.48550/arXiv.2508.21675
101. Gallegos I.O., Rossi R.A., Barrow J., et al. (2024). Bias and fairness in large language models: A survey. *Comput. Linguist.* **50**:1097–1179. DOI:10.1162/coli_a_00524
102. Liu Y., Deng G., Li Y., et al. (2023). Prompt injection attack against llm-integrated applications. *arXiv* preprint. DOI:10.48550/arXiv.2306.05499
103. Zhou X., Qiang Y., Zade S.Z., et al. (2023). Hijacking large language models via adversarial in-context learning. *arXiv* preprint. DOI:10.48550/arXiv.2311.09948
104. Gong Y., Chen Z., Chen M., et al. (2025). Topic-fliprag: Topic-orientated adversarial opinion manipulation attacks to retrieval-augmented generation models. *arXiv* preprint. DOI:10.5555/3766078.3766274
105. Schwinn L., Dobre D., Günnemann S., et al. (2023). Adversarial attacks and defenses in large language models: Old and new threats. *arXiv* preprint. DOI:10.48550/arxiv.2310.19737
106. Raina V., Liusie A. and Gales M. (2024). Is llm-as-a-judge robust? Investigating universal adversarial attacks on zero-shot llm assessment. *arXiv* preprint. DOI:10.48550/arXiv.2402.14016
107. Guo Y., Guo M., Su J., et al. (2024). Bias in large language models: Origin, evaluation, and mitigation. *arXiv* preprint. DOI:10.48550/arXiv.2411.10915
108. Navigli R., Conia S. and Ross B. (2023). Biases in large language models: Origins, inventory, and discussion. *ACM J. Data Inf. Qual.* **15**:1–21. DOI:10.1145/3597307
109. Angrist J.D. (2014). The perils of peer effects. *Lab. Econ.* **30**:98–108. DOI:10.1016/j.labeco.2014.05.008
110. Lewis P., Perez E., Piktus A., et al. (2020). Retrieval-augmented generation for knowledge-intensive nlp tasks. *Adv. Neural Inf. Process. Syst.* **33**:9459–9474. DOI:10.5555/3495724.3496517
111. Schwarzschild A., Goldblum M., Gupta A., et al. (2021). Just how toxic is data poisoning. *a unified benchmark for backdoor and data poisoning attacks. Proc. Int. Conf. Mach. Learn.* **2021**:9389–9398. DOI:10.48550/arXiv.2006.12557
112. Touvron H., Lavril T., Izacard G., et al. (2023). Llama: Open and efficient foundation language models. *arXiv* preprint. DOI:10.48550/arXiv.2302.13971
113. Bowen D., Murphy B., Cai W., et al. (2025). Scaling trends for data poisoning in LLMs. *Proc. AAAI Conf. Artif. Intell.* **39**:27206–27214. DOI:10.1609/aaai.v39i26.34929
114. Liu T., Zhang Y., Feng Z., et al. (2024). Beyond traditional threats: A persistent backdoor attack on federated learning. *Proc. AAAI Conf. Artif. Intell.* **38**:21359–21367. DOI:10.1609/aaai.v38i19.30131
115. Zhu C., Li Y., Rao B., et al. (2025). SPA: Towards more stealth and persistent backdoor attacks in federated learning. *arXiv* preprint. DOI:10.48550/arXiv.2506.20931
116. Tian Z., Cui L., Liang J., et al. (2022). A comprehensive survey on poisoning attacks and countermeasures in machine learning. *ACM Comput. Surv.* **55**:1–35. DOI:10.1145/3551636
117. Zhao P., Zhu W., Jiao P., et al. (2025). Data poisoning in deep learning: A survey. *arXiv* preprint. DOI:10.48550/arXiv.2503.22759

118. Muñoz-González L., Biggio B., Demontis A., et al. (2017). Towards poisoning of deep learning algorithms with back-gradient optimization. *Proc. ACM Workshop Artif. Intell. Secur.* **2017**:27–38. DOI:10.1145/3128572.3140451
119. Nourani M., Roy C., Block J.E., et al. (2021). Anchoring bias affects mental model formation and user reliance in explainable AI systems. *Proc. Int. Conf. Intell. User Interfaces* **2021**:340–350. DOI:10.1145/3397481.3450639
120. Shi F., Chen X., Misra K., et al. (2023). Large language models can be easily distracted by irrelevant context. *Proc. Int. Conf. Mach. Learn.* **2023**:31210–31227. DOI:10.5555/3618408.3619699
121. Dougrez-Lewis J., Akhter M.E., Ruggeri F., et al. (2025). Assessing the the reasoning capabilities of llms in the context of evidence-based claim verification. *Findings Assoc. Comput. Linguist. ACL* **2025**:20604–20628. DOI:10.18653/v1/2025.findings-acl.1059
122. Hong R., Zhang H., Pang X., et al. (2024). A closer look at the self-verification abilities of large language models in logical reasoning. *Proc. NAACL HLT* **2024**:900–925. DOI:10.18653/v1/2024.naacl-long.52
123. Li J., Li Y., Hu X., et al. (2025). Aspect-guided multi-level perturbation analysis of large language models in automated peer review. *arXiv* preprint. DOI:10.48550/arXiv.2502.12510
124. Liang W., Zhang Y., Cao H., et al. (2024). Can large language models provide useful feedback on research papers? A large-scale empirical analysis. *NEJM AI* **1**:AIoa2400196. DOI:10.1056/AIoa2400196
125. Zhou Z., Li Z., Zhang J., et al. (2025). Corba: Contagious recursive blocking attacks on multi-agent systems based on large language models. *arXiv* preprint. DOI:10.48550/arXiv.2502.14529
126. Zhu S., Zhang R., An B., et al. (2023). Autodan: Interpretable gradient-based adversarial attacks on large language models. *arXiv* preprint. DOI:10.48550/arXiv.2310.15140
127. Zizzo G., Cornacchia G., Fraser K., et al. (2025). Adversarial prompt evaluation: Systematic benchmarking of guardrails against prompt input attacks on llms. *arXiv* preprint. DOI:10.48550/arXiv.2502.15427
128. Bozdag N.B., Mehri S., Tur G., et al. (2025). Persuade me if you can: A framework for evaluating persuasion effectiveness and susceptibility among large language models. *arXiv* preprint. DOI:10.48550/arXiv.2503.01829
129. Salvi F., Horta Ribeiro M., Gallotti R., et al. (2025). On the conversational persuasiveness of GPT-4. *Nat. Hum. Behav.* **9**:1645–1653. DOI:10.1038/s41562-025-02194-6
130. Liu Y., Yang K., Liu Y., et al. (2023). The shackles of peer review: Unveiling the flaws in the ivory tower. *arXiv* preprint. DOI:10.48550/arXiv.2310.05966
131. Nisbett R.E. and Wilson T.D. (1977). The halo effect: Evidence for unconscious alteration of judgments. *J. Pers. Soc. Psychol.* **35**:250–256. DOI:10.1037/0022-3514.35.4.250
132. Zhang J., Zhang H., Deng Z., et al. (2022). Investigating fairness disparities in peer review: A language model enhanced approach. *arXiv* preprint. DOI:10.48550/arXiv.2211.06398
133. Fox C.W., Meyer J. and Aimé E. (2023). Double - blind peer review affects reviewer ratings and editor decisions at an ecology journal. *Funct. Ecol.* **37**:1144–1157. DOI:10.1111/1365-2435.14259
134. Sun M., Barry Danfa J. and Teplitskiy M. (2021). Does double - blind peer review reduce bias. *Evidence from a top computer science conference. J. Assoc. Inf. Sci. Technol.* **73**:811–819. DOI:10.1002/asi.24582
135. Verharen J.P.H. (2023). ChatGPT identifies gender disparities in scientific peer review. *eLife* **12**:RP90230. DOI:10.7554/eLife.90230
136. Hosseini M. and Horbach S. (2023). Fighting reviewer fatigue or amplifying bias. Considerations and recommendations for use of ChatGPT and other large language models in scholarly peer review. *Res. Integr. Peer Rev.* **8**:4. DOI:10.1186/s41073-023-00133-5
137. Soneji A., Kokulu F.B., Rubio-Medrano C., et al. (2022). "Flawed, but like democracy we don't have a better system": The Experts' Insights on the peer review process of evaluating security papers. *Proc. IEEE Symp. Secur. Priv.* **2022**:1845–1862. DOI:10.1109/SP46214.2022.9833581
138. Schramowski P., Turan C., Andersen N., et al. (2022). Large pre-trained language models contain human-like biases of what is right and wrong to do. *Nat. Mach. Intell.* **4**:258–268. DOI:10.1038/s42256-022-00458-8
139. Li H., Shan S., Wenger E., et al. (2022). Blacklight: Scalable defense for neural networks against query-based black-box attacks. *Proc. USENIX Secur. Symp.* **2022**:2117–2134. https://www.usenix.org/conference/usenixsecurity22/presentation/li-huiying
140. Koo R., Lee M., Raheja V., et al. (2024). Benchmarking cognitive biases in large language models as evaluators. *Findings Assoc. Comput. Linguist. ACL* **2024**:517–545. DOI:10.18653/v1/2024.findings-acl.29
141. Bartos O.J. and Wehr P. (2002). *Using conflict theory* (Cambridge University Press). DOI:10.2307/1556608
142. Robertson Z. (2023). GPT4 is slightly helpful for peer-review assistance: A pilot study. *arXiv* preprint. DOI:10.48550/arXiv.2307.05492

FUNDING AND ACKNOWLEDGMENTS

This work is funded by Basic Research Program of Jiangsu under Grants BK20253021 and National Science Foundation of China (NSFC) under Grant No. 62506075. The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

AUTHOR CONTRIBUTIONS

J. Wang was responsible for the conceptualization, methodology, writing – original draft, and project implementation. Y. Liu was primarily responsible for software development and implementation. H. Xu and K. Hu contributed to software development, validation, and writing – review & editing. S. Di and W. Ni contributed to the conceptualization, supervision, writing – review & editing, and overall guidance of the study. L. Yue contributed to validation and manuscript proofreading. M.-L. Zhang, K. Ren, and L. Chen provided research resources, supervision, and funding acquisition. All authors contributed to the manuscript and approved the final version.

DECLARATION OF INTERESTS

Shimin Di is an Academic Editor of The Innovation Informatics and was blinded from reviewing or making final decisions on the manuscript. Peer review was handled independently of this member and their research group. The other authors declare no conflicts of interest.

ETHICAL STATEMENT AND PATIENT CONSENT

Not Applicable.

DATA AND CODE AVAILABILITY

The data that support the findings of this study are openly available in Hugging Face at https://huggingface.co/datasets/Faultiness/AI_Reviews_ICLR

LEAD CONTACT WEBSITE

Shimin Di: https://sdiaa.tech