

The Emerging AI Paper-Review Arms Race: Adversarial Co-Evolution in Scholarly Publishing

Chenguang Wang¹ Ming Li² Adebayo Braimah³ Chenrui Fan²
Tuo Wang¹ Weijie Guan¹ Ruiyi Zhang Tianyi Zhou⁴ Dawei Zhou¹

¹Virginia Tech ²University of Maryland ³Stony Brook University ⁴MBZUAI
{cswang, zhoud}@vt.edu, minglii@umd.edu

Project website: https://github.com/MingLiiii/Awesome_AI_Arm_Race

Abstract

Generative and agentic AI are reshaping both the production and evaluation of scientific research. These developments are often studied separately, as questions of how AI can produce research and how AI can review it. We argue that this separation misses an increasingly important feature of scholarly publishing: changes on one side alter the incentives, constraints, and behavior of the other. We synthesize 230 scholarly publications and institutional records using a taxonomy of six connected dynamics: production scaling, evaluation automation, evaluation manipulation, defense mechanisms and policy responses, evasion and side effects, and long-horizon ecosystem feedback. The literature shows an emerging progression in which cheaper and faster research production increases pressure on evaluation, AI-mediated evaluation becomes more scalable and repeatable, participants can exploit evaluator regularities, and institutions respond with technical safeguards and policy controls. These responses can in turn induce evasion, redistribute errors and workload, and shape the scholarly records reused by future research and evaluation systems. Evidence is strongest for production and evaluation at scale, reproducible manipulation, and institutional response, while post-policy adaptation and artifact-level long-horizon feedback remain less directly observed. This systems view shifts attention from isolated AI capabilities toward how scholarly actors and AI systems adapt to one another over time.

1 Introduction

Generative and agentic AI are increasingly entering both sides of scientific research and evaluation. On the production side, researchers use LLM-based tools throughout the research cycle, from exploring ideas and reviewing relevant literature to conducting experiments and preparing manuscripts (Baek et al., 2025; Liang et al., 2025; Lu et al., 2024; Yamada et al., 2025; Tang et al., 2026b; Kong et al., 2026). On the evaluation side, generative AI is increasingly used across peer review, extending from manuscript summarization and critique generation to evaluation and publication decision-making (Liang et al., 2024b; Nguyen and Ahmadi, 2026; Thakkar et al., 2026; Biswas et al., 2026). These developments are commonly studied through two corresponding questions: how AI can support or automate scientific research, and how AI can perform or assist scholarly evaluation. While this separation is useful for studying the capabilities of autonomous research systems and AI reviewers in isolation, it is incomplete once the two sides operate within the same publishing process. AI-mediated production changes the volume, speed, and form of what enters evaluation, while AI-mediated evaluation creates new signals and regularities to which authors can respond. The relevant challenge therefore lies not only in the capabilities of either side, but also in how changes on one side reshape the conditions, costs, and incentives faced by the other.

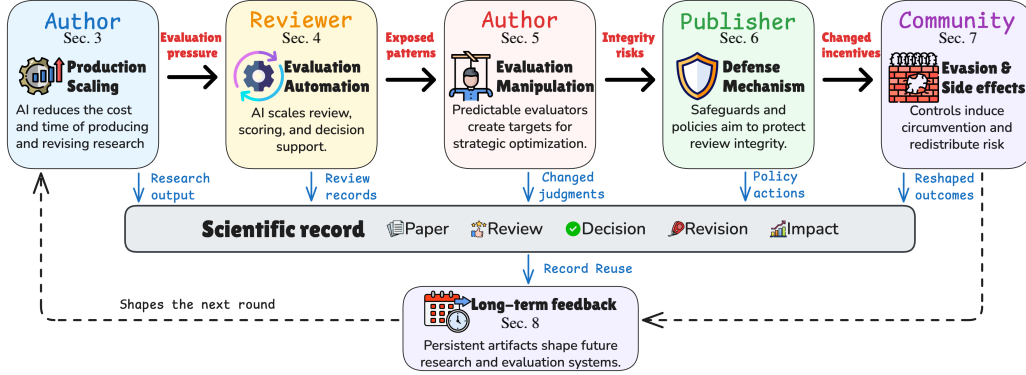


Figure 1: Six connected dynamics in the emerging AI paper-review arms race. Solid arrows show the principal response order, while the dotted path represents slower feedback from accumulated scholarly records into future production, evaluation, and institutional response.

The most immediate form of this coupling is a capacity imbalance. As AI reduces the time and marginal cost required to conduct, write, and revise research, scientific output and submission volume can grow faster than available human reviewing capacity, increasing pressure to scale evaluation as well. Evidence for this dynamic is already emerging from empirical studies and real-world deployments. Observational studies associate LLM adoption with shorter research and revision cycles, increased scientific output, rising submission volume, and growing pressure on limited reviewing capacity (Kusumegi et al., 2025; Qi et al., 2025; Filimonovic et al., 2025; Gartenberg et al., 2026). Increasingly capable agentic AI systems extend this scaling from assistance with individual tasks toward integrated idea-to-paper workflows, further reducing the time and marginal cost required to produce complete research artifacts (Kong et al., 2026; Tang et al., 2026b; Song et al., 2026; Yuan et al., 2026). Evaluation is beginning to scale alongside this growth: AI is now used both to assist individual reviewers (Thakkar et al., 2026) and within official conference workflows, as illustrated by the large-scale AAAI-26 AI-review pilot (Biswas et al., 2026). At the same time, machine-mediated evaluation makes parts of the evaluation process more repeatable and therefore easier for authors to adapt to. Hidden instructions targeting AI-assisted reviewers have been identified in public manuscripts (Lin, 2026), while controlled studies show that reviewer judgments can be shifted without corresponding improvements in the underlying scientific evidence and that repeated evaluator feedback can be used to optimize manuscripts toward more favorable assessments (Yang et al., 2026c; Zhou et al., 2025; Li et al., 2026c). Venues have, in turn, introduced AI-specific review policies, submission controls, detection mechanisms, and integrity safeguards (ICLR 2026 Program Chairs, 2025a; Agarwal et al., 2026; NeurIPS 2026 Communication Chairs, 2026; Li et al., 2024b; Fichtl et al., 2026). Taken together, these developments form an emerging sequence of scaling, adaptation, and institutional response rather than two independent trends in AI adoption.

Existing surveys provide valuable accounts of AI-assisted scientific production and peer review, typically organizing the literature around tasks, system components, or stages of the research and review workflow (Kuznetsov et al., 2024; Zhuang et al., 2025; Wu et al., 2026c; Nguyen and Ahmadi, 2026; Kong et al., 2026). Recent surveys have also highlighted the verification gap in autonomous research agents and discussed governance and human oversight in AI-assisted peer review (Ding et al., 2026b; Mann et al., 2025; Wei et al., 2025). Other recent work has also begun to connect review with rebuttal, revision, and further experimentation (Weng et al., 2025; Han et al., 2026; Ma et al., 2026b; Wu, 2026). These perspectives explain what individual systems do and how outputs move through a workflow, but they do not fully capture how one actor’s use of AI changes another actor’s signals, costs, and feasible responses. Production scaling can alter evaluation demand, machine-mediated evaluation can create new targets for optimization, and institutional interventions can change the incentives for subsequent adaptation and evasion. What is therefore missing is not another inventory of AI tools, but a relational account of how these developments interact across actors in the scientific ecosystem. We synthesize these relations as a coupled process that we term the *AI paper-review arms race*. The term serves as a descriptive lens for documented or testable patterns of adaptation and counter-adaptation across actors in the scientific ecosystem.

Figure 1 summarizes this coupled sequence and the scholarly actors through which it operates.

To better characterize this adaptive process and provide a coherent structure for the survey, we develop a descriptive taxonomy of six linked dynamics: (1) *production scaling*, (2) *evaluation automation*, (3) *evaluation manipulation*, (4) *defense mechanisms and policy responses*, (5) *evasion and side effects*, and (6) *long-horizon ecosystem feedback*. The taxonomy is organized around the principal response relations among actors in the scientific ecosystem rather than around technologies alone. Production scaling changes the demand placed on evaluation; evaluation automation makes parts of the evaluation process increasingly repeatable and therefore more susceptible to manipulation and optimization; venues respond through technical and institutional defenses; and those interventions can induce evasion or shift costs and risks to other actors. The resulting papers, reviews, and decisions can also persist and influence future research and evaluation systems. These dynamics are therefore relation-oriented and non-exclusive: each reported finding, deployment, or institutional action is assigned a primary dynamic according to the scholarly function it directly observes, tests, or implements. We then examine whether individual studies directly connect adjacent dynamics or whether a proposed connection is synthesized across separate bodies of literature. Throughout the survey, we avoid treating a capability demonstration as population prevalence, a policy announcement as an effective intervention, or a plausible downstream pathway as an observed causal cycle.

Tracing these six dynamics as a connected process yields three recurring conclusions about the emerging system. (1) Trustworthy evaluation is the binding constraint. Generation, revision, and review generation can scale rapidly, whereas validating evidence, assessing scientific soundness and novelty, and making accountable decisions remain costly. (2) Adaptation weakens static evaluation. Static evaluations of AI reviewers can overestimate their real-world reliability once authors are able to observe, query, and adapt to machine-mediated evaluation. (3) Safeguards redistribute risk and burden. No single detector or policy can govern the resulting interaction; effective safeguards must account for how labor, error, confidentiality risk, and enforcement burden shift across actors and rounds. Support for this account is nevertheless uneven across the loop. Production scaling and evaluation automation have been observed in large-scale deployments, evaluation manipulation has been both observed and experimentally reproduced, and defense mechanisms and policy responses are documented, whereas post-policy counter-adaptation and long-horizon ecosystem feedback remain substantially less established.

In summary, this survey shifts the unit of analysis from individual AI tools to the response relations through which scholarly actors adapt to one another, while identifying which connections have and have not yet been directly observed. The survey is organized around this connected progression. Section 2 defines the system boundary, institutional roles, taxonomy, classification rules, and literature scope. Sections 3-8 trace the six dynamics in response order, and Section 9 synthesizes their implications, limitations, falsifiable boundaries, and priority studies before Section 11 concludes.

2 Taxonomy

This section formalizes the taxonomy used throughout the survey. We first define the system boundary and the institutional roles through which AI operates within the scientific ecosystem, then describe the six dynamics and their principal response order, and finally specify the literature-mapping and source-selection procedures. The taxonomy is descriptive and relation-oriented, organizing the literature around scholarly functions and the response relations examined across studies.

2.1 Survey scope and actors

This survey examines the coupled and recursive effects of emerging AI tools across the scientific ecosystem. The scope begins with research production, including idea development, experimentation, manuscript preparation, revision, and rebuttal, and extends to peer review, publication decisions, and the policies, tools, and safeguards introduced by conferences and publishers. We also consider how these stages interact, as changes in research production can alter evaluation practices, while evaluation and institutional responses can in turn shape subsequent author behavior. Beyond these immediate interactions, we examine their longer-term consequences, including how papers, reviews, decisions, corrections, and citations can be reused by retrieval systems, scientific agents, training pipelines, and AI-based evaluators, thereby influencing future rounds of research and evaluation. The survey is organized around the progression of these interactions and responses. Section 3 examines

Table 1: Coverage in recent surveys and broad syntheses of AI-assisted research and scholarly peer review. ✓ indicates that a dimension is explicitly identified and incorporated into the discussion; × indicates that it is absent or mentioned only in passing.

Survey	Organizing frame	Production	Evaluation	Strategic interaction	Adaptive coevolution	Governance	Long-term feedback
Kuznetsov et al. (2024)	Peer-review workflow	✓	✓	✓	×	×	×
Zhuang et al. (2025)	Automated-review pipeline	×	✓	×	×	✓	×
Eger et al. (2025)	AI-assisted science tasks	✓	✓	×	×	×	×
Kong et al. (2026)	End-to-end research lifecycle	✓	✓	✓	×	✓	×
Ding et al. (2026b)	Autonomous research agents and verification gap	✓	✓	×	×	✓	×
Nguyen and Ahmadi (2026)	Review methods and reliability	×	✓	✓	×	✓	×
Mann et al. (2025)	AI-assisted peer-review design and governance	×	✓	✓	×	✓	×
Wu et al. (2026c)	Review and post-review tasks	✓	✓	✓	×	×	×
This survey	Coupled dynamics and responses	✓	✓	✓	✓	✓	✓

AI-enabled production scaling and the resulting pressure on evaluation, while Section 4 covers the automation of scholarly evaluation. Sections 5-7 focus most directly on strategic interaction and adaptive response, tracing the progression from manipulation of AI-mediated evaluation through institutional defense to evasion and unintended side effects. Section 8 then considers the longer-term recursive effects of these dynamics on future research and evaluation.

Under this scope, we distinguish several functional roles rather than a fixed set of mutually exclusive actors. Authors and AI-enabled research systems produce and revise scientific work and respond to evaluation, while reviewers, editors, and committees assess that work and contribute to publication decisions. At the institutional level, venues, publishers, and platform operators shape the rules and technical conditions under which research and evaluation take place, including review policies, evaluation rubrics, and safeguards. AI-enabled systems may operate within these constraints and may also reuse papers, reviews, decisions, and other scholarly records in subsequent research or evaluation. These roles can overlap, and the same actor or system may occupy multiple roles. The taxonomy therefore distinguishes scholarly functions rather than fixed actor identities, allowing us to trace how changes in one part of the ecosystem alter the conditions and responses of others.

We use the paper-review arms race as an analytic lens for sequences of adaptive response, not as a synonym for any use of AI in research or review. Operationally, an arms-race episode involves an actor taking an action that changes a target or signal relevant to another actor, a counter-response by that actor or an institution, and a resulting shift in incentives, costs, or capabilities that can motivate further adaptation. A single instance of AI-assisted production, evaluation, or defense can therefore belong to the taxonomy without constituting an arms-race episode unless it participates in such a response relation. Because complete sequences are rarely observed within a single study, connections synthesized across separate bodies of literature are presented as research questions rather than as directly observed causal sequences.

2.2 Six categories in the taxonomy

Our taxonomy comprises six process categories that describe how AI changes scientific research and evaluation and how different parts of the ecosystem respond to those changes. We refer to these categories as dynamics because they capture processes of change and response rather than static system components. Table 1 situates this relation-oriented organization relative to recent surveys organized around research and review tasks or workflows. Figure 2 provides a detailed hierarchy of the six dynamics, their recurring mechanisms, and representative studies.

For Table 1, coverage is coded by substantive organizing scope rather than by isolated keyword mention. *Production* requires discussion of AI-supported research creation or revision; *evaluation*, of review, assessment, or decision support; *strategic interaction*, of attempts to influence an evaluator or evaluative outcome; *adaptive coevolution*, of linked response and counter-response across actors or institutions; *governance*, of technical safeguards, institutional rules, or accountability; and *long-term feedback*, of persistent scholarly artifacts shaping later research or evaluation systems.

Table 2: Operational interpretation and boundaries of the six taxonomy dynamics.

Dynamic	Organizing question	Included scope	Main interpretive boundary
Production scaling	Does AI increase the rate or reduce the marginal cost of producing or revising research artifacts?	Idea development, literature work, experimentation, writing, rebuttal and revision, and end-to-end research workflows.	A capability demonstration or lower production cost does not establish a causal increase in submissions, open-ended scientific progress, or scientific validity.
Evaluation automation	Does AI generate, improve, aggregate, or support judgments about scientific work?	Manuscript assessment, review and feedback generation, scoring, meta-review, reviewer assistance, and decision support.	Helpful or human-like reviews do not establish that evaluation is reliable, diverse, autonomous, or accountable.
Evaluation manipulation	Is an evaluative outcome shifted without a commensurate improvement in scientific evidence?	Hidden instructions, prompt injection, evaluator-aware framing, presentation optimization, and iterative feedback gaming.	A score change with the scientific content held fixed shows evaluator sensitivity. It does not by itself distinguish clearer presentation from manipulation or show an effect on venue decisions.
Defense mechanisms and policy responses	Does an intervention protect evaluation capacity, integrity, or accountability?	Detection and verification, reviewer safeguards, submission and review controls, disclosure rules, audits, and appeals.	Adoption or benchmark performance does not establish deterrence, net benefit, or equitable burden.
Evasion and side effects	How do actors adapt to controls, and where do errors, risks, and costs move?	Trace removal, paraphrasing, routing or strategy changes, false positives, verification labor, confidentiality risk, and compliance burden.	Evasion must respond to a control; side effects may arise without strategic intent and should not be treated as evasion.
Long-horizon ecosystem feedback	Do persistent scholarly artifacts shape later research or evaluation systems?	Reuse of papers, reviews, decisions, corrections, and citations in retrieval, training, scientific agents, and AI evaluators.	Reuse or exposure is not a demonstrated closed loop; causal feedback requires a traceable artifact-to-successor path.

(1) Production scaling. This category captures uses of AI that reduce the time, cost, or effort required to produce and revise research outputs. It includes assistance with individual activities such as literature search, experimentation, writing, and rebuttal, as well as increasingly integrated systems that combine these activities into end-to-end research workflows.

(2) Evaluation automation. This category covers the use of AI to support or automate the evaluation of scientific work. It includes manuscript assessment, review generation and revision, scoring, meta-review, and decision support, ranging from assistance to individual reviewers to AI components incorporated into official evaluation workflows.

(3) Evaluation manipulation. This category refers to attempts to influence AI-mediated evaluation without a corresponding improvement in the underlying scientific evidence. Such attempts can directly target an AI reviewer through hidden instructions or other attacks, but can also exploit presentation, framing, or other evaluator-visible signals that systematically change review outcomes.

(4) Defense mechanisms and policy responses. This category includes technical, procedural, and institutional responses intended to protect the reliability, integrity, or accountability of research and evaluation. These responses include detection and verification methods, safeguards for AI reviewers, submission and review controls, disclosure requirements, and policies governing the use of AI by different participants in the scientific process.

(5) Evasion and side effects. This category captures the responses and consequences that arise after defenses or restrictions are introduced. Participants may adapt their tools, traces, or strategies to avoid detection or other controls. At the same time, defensive measures can create unintended effects by shifting verification effort, false-positive risk, confidentiality concerns, or other costs to different participants in the ecosystem.

(6) Long-horizon ecosystem feedback. This category extends the analysis beyond an individual research or review cycle. Papers, reviews, decisions, corrections, citations, and other scholarly records can later be reused by retrieval systems, scientific agents, training pipelines, or AI-based evaluators. Through this reuse, outputs produced in one round of research and evaluation can influence the information, objectives, and behavior of future scientific systems.

These six categories are connected by a principal response order rather than treated as independent topics. Production scaling can increase the amount and pace of research entering evaluation, encouraging evaluation to scale through greater automation. As AI-mediated evaluation becomes more repeatable and observable, it can create new opportunities for manipulation or optimization. Such behavior

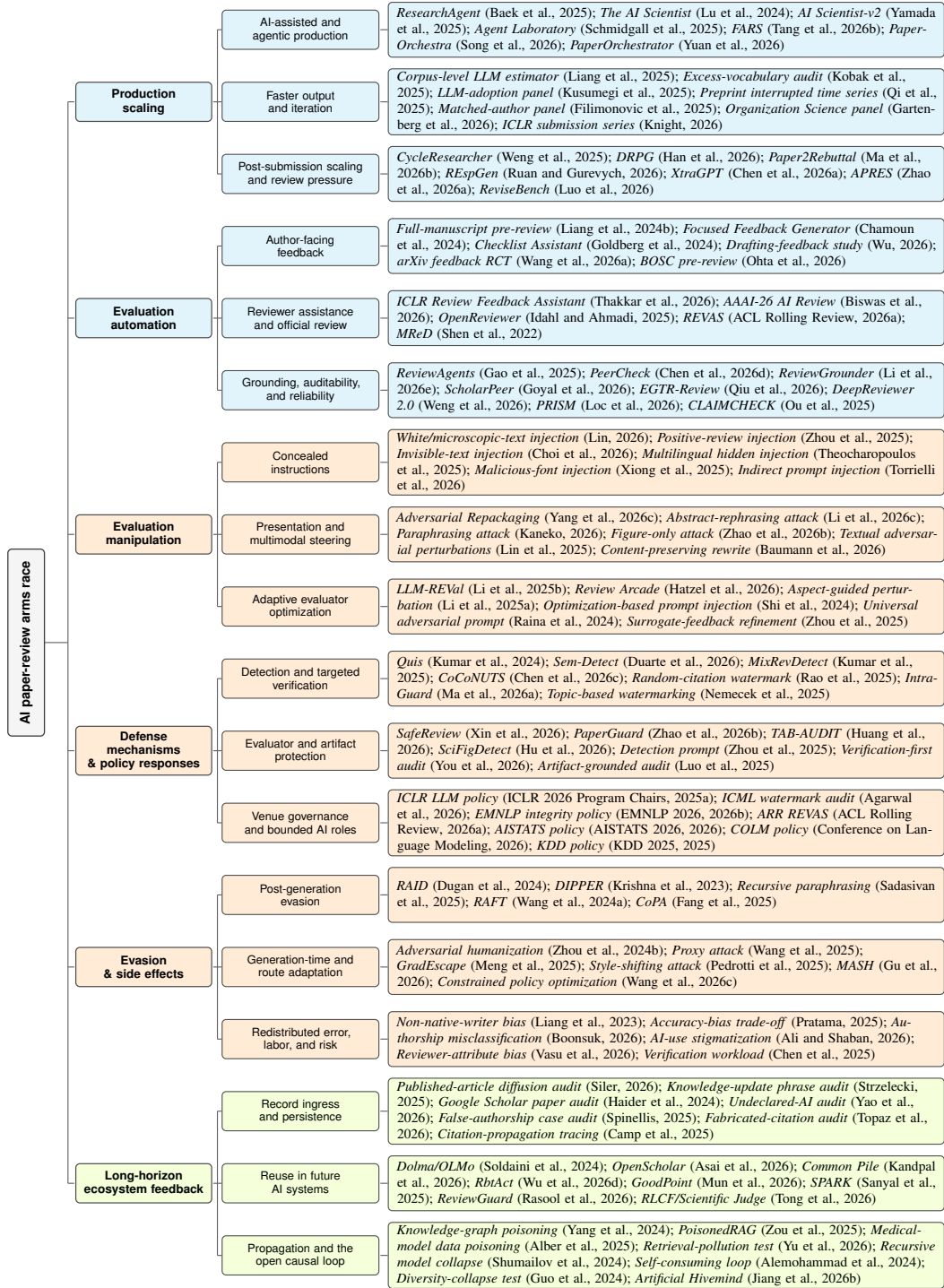


Figure 2: Detailed taxonomy of the emerging AI paper-review arms race. The middle and right columns summarize recurring mechanisms and representative methods, systems, benchmarks, datasets, and policies, with colors distinguishing scaling and automation, strategic response and governance, and long-horizon feedback.

can motivate technical and institutional defenses, which may in turn induce evasion or redistribute costs and risks. Finally, the papers, reviews, decisions, and other records produced through these interactions can persist and influence later rounds of research and evaluation. The taxonomy therefore

captures both the individual processes represented in the literature and the relations through which one process can reshape the next.

2.3 Survey framework

We use a consistent procedure to map the retained literature to the taxonomy. The framework addresses two questions: how each study contributes to the six categories and whether it directly examines a response relation between them.

2.3.1 Taxonomy mapping

To organize the literature consistently, we map distinct findings rather than forcing each paper into a single category. Each finding is assigned a primary dynamic according to the scholarly function it directly observes, tests, or implements. A paper may therefore contribute to multiple dynamics when it reports findings that address different parts of the process.

We separately identify response relations when a study directly examines how one dynamic changes the conditions, behavior, or outcomes of another. For example, a study may evaluate an automated reviewer while also testing how authors adapt their manuscripts to its feedback; the former finding contributes to evaluation automation, the latter to evaluation manipulation, and the study directly examines the relation between the two dynamics. Connections synthesized across separate bodies of literature, rather than directly examined within a study, are presented as open research questions.

This mapping separates the function represented by each finding from the relation it may support, allowing individual studies to contribute to multiple parts of the taxonomy without treating the categories as mutually exclusive at the paper level.

2.3.2 Literature collection and scope

To assemble the evidence base, we conducted a structured literature search for work available through July 1, 2026. Query families aligned with the six taxonomy categories were applied across arXiv, OpenReview, ACL Anthology, PMLR, JMLR, publisher and proceedings pages, DOI records, and official conference websites. We complemented this search by tracing original studies from existing surveys and retrieving related work based on titles, abstracts, and topical similarity. We later conducted targeted updates for newly released work; as discussed in Section 10, these updates did not rerun every search query.

The search and retrieval process yielded approximately 500 candidate papers and institutional records. After deduplication and manual screening for relevance to AI-mediated scientific production, evaluation, and the interactions represented in our taxonomy, we retained 230 sources. Recent preprints were included when they provided directly relevant evidence, with publication status tracked separately. Work with only weak connections to the AI paper-review process was excluded unless it established a mechanism that clearly transferred to one of the six categories. For retained sources, claims used in the survey were checked against the corresponding original paper, report, policy, or official record whenever available.

Study settings and designs are described when they materially constrain interpretation. In particular, controlled demonstrations are not used to infer real-world prevalence, announced policies are not treated as evidence of effectiveness, and plausible downstream pathways are not presented as observed causal cycles.

Appendix Tables A.1 and A.2 complement this mapping by summarizing reusable scholarly records and representative empirical evidence for the six dynamics and their response relations.

3 Scaling scholarly production

Production scaling refers to how AI can enable research outputs to be produced in greater volume or in less time. By reducing the time and effort required for research production, revision, and resubmission, AI can enable researchers or automated systems to produce more outputs, complete them faster, or both. This is important because production capacity and evaluation capacity do not necessarily scale together.

3.1 From industrialized production to AI-enabled scaling

Large-scale production of scholarly artifacts is not entirely new. Before generative AI, its most visible problematic form was industrialized production through paper mills and related authorship-for-sale practices (Porter and McIntosh, 2024; Abalkina, 2023). Such operations showed that research-like artifacts could be produced systematically when production was organized around volume rather than individual scholarly workflows. Generative AI changes the accessibility and scope of this capability. Tasks that previously required substantial individual time, including literature search, drafting, experimentation, and formatting, can increasingly be assisted or automated within ordinary research workflows.

The first visible change has been the widespread adoption of AI-assisted writing. Corpus studies find distributional changes in key sections of scientific papers that are consistent with substantial LLM-associated modification in computer-science preprints and biomedical publishing (Liang et al., 2025; Kobak et al., 2025). These studies confirm the diffusion of AI-assisted writing and indicate that AI-enabled writing tools have become widely used in scientific communication.

More capable research systems extend this shift beyond writing assistance. ResearchAgent supports iterative idea development over the literature; the AI Scientist and its successor integrate ideation, experimentation, manuscript production, and evaluation; and CycleResearcher uses automated review as a revision signal (Baek et al., 2025; Lu et al., 2024; Yamada et al., 2025; Weng et al., 2025). Agent Laboratory, PaperOrchestra, and PaperOrchestrator further connect literature, code, experimental logs, and manuscript generation (Schmidgall et al., 2025; Song et al., 2026; Yuan et al., 2026). Related systems extend these workflows to coding, bioinformatics, and multimodal medical research (Jansen et al., 2025; Tang et al., 2026a; Ifargan et al., 2025; Kamran et al., 2026; Wu et al., 2026b). Other systems emphasize particular stages of the research process rather than full paper production. The AI co-scientist, for example, uses multiple agents to generate, debate, and refine hypotheses that are subsequently tested by human scientists in biomedical settings (Gottweis et al., 2025). Together, these systems illustrate different levels of automation across the research workflow.

More importantly, the key change is not simply that individual systems contain more components, but that multiple stages of research can increasingly be coordinated within the same automated workflow. This makes it cheaper to explore multiple research directions, discard unsuccessful branches, and turn successful ones into complete artifacts. AgentRxiv, for example, allows agent laboratories to share intermediate reports so that research trajectories can be compared and reused rather than executed independently (Schmidgall and Moor, 2025). EvoScientist and AutoResearchClaw similarly carry information across experiments or runs (Lyu et al., 2026; Liu et al., 2026), while Jr. AI Scientist, PaperClaw, and Claw AI Lab introduce mechanisms such as provenance, experimental logs, rollback, and human intervention (Miyai et al., 2025; Ye et al., 2026; Wu et al., 2026a).

Deployment-scale systems further show that this increase in throughput is no longer limited to small demonstrations. Fully Automated Research System (FARS) reports a public deployment that produced 166 complete papers across 67 AI/ML topics while preserving proposals, code, logs, results, and manuscripts (Tang et al., 2026b). This demonstrates that complete research artifacts can be generated at substantial scale, but producing a complete artifact is not the same as producing scientifically strong or successful research. In 282 structured reviews of 140 FARS outputs, the mean overall score was 3.17 on an ICLR-style scale, and only 11.4% reached a mean score of six. Artifact-grounded audits also identified integrity problems, with insufficient experimental evidence emerging as a major weakness. A complementary result comes from two shadow evaluations in which frontier agents completed substantial engineering work and produced full manuscripts but did not solve the main open-ended research problems (Kirgis et al., 2026). Human-verifiable and laboratory-coupled systems provide alternative designs (Ifargan et al., 2025; Ghareeb et al., 2026), but the broader pattern remains the same: the capacity to produce research artifacts can scale faster than the capacity to establish that their scientific claims and conclusions are sound.

3.2 Empirical evidence of faster and greater production

Evidence from real scholarly corpora increasingly connects AI adoption with changes in research output. Across 2.1 million preprints from arXiv, bioRxiv, and SSRN, Kusumegi et al. (2025) estimate that inferred AI adoption is associated with output increases of 36.2%, 52.9%, and 59.8%, respectively. Estimated gains vary substantially across author backgrounds and fields, and are particularly large for

some researchers facing higher English-writing costs. The same study suggests that lower production barriers may also broaden access to different forms of literature and participation. However, adoption is inferred from text, and the design does not provide causal identification.

Renault et al. (2026) question the size of this estimate. Because adoption is defined by an author’s first detector-flagged paper, authors whose output is already rising are more likely to be classified as adopters. Their simulations and placebo tests recover similar post-adoption increases even when LLM use has no effect. The analysis does not show that LLMs have no productivity benefit; it shows that this design does not identify that benefit.

Other studies provide complementary evidence for changes in scholarly production. Using matched author panels in the social and behavioral sciences, Filimonovic et al. (2025) associate adoption with 15% higher productivity in 2023 and 36% in 2024, although adoption is again inferred from textual markers and residual selection remains possible. At the aggregate level, Qi et al. (2025) analyze more than 2.1 million records from four preprint repositories and find faster submission and revision cycles after the release of ChatGPT, particularly in computational fields. Their interrupted time-series design cannot fully separate the effect of LLM availability from other contemporaneous changes.

Taken together, these studies show that AI-assisted writing is widespread and that inferred adoption is associated with higher output or faster production cycles. They do not yet establish how much of this relationship is causal or how broadly it generalizes across fields and researcher groups.

3.3 Scaling continues after submission

Production does not end when a manuscript is submitted. Review, rebuttal, revision, and resubmission create additional production cycles, and AI can reduce the time and effort required at these stages as well. Response systems now decompose reviews into specific concerns, retrieve relevant manuscript evidence, plan responses, and learn from earlier response failures (Li and Zhao, 2025; Han et al., 2026). Paper2Rebuttal and REspGen introduce structured plans, human checkpoints, author-specific information, and response controls, while XtraGPT applies context-aware revisions directly to the manuscript (Ma et al., 2026b; Ruan and Gurevych, 2026; Chen et al., 2026a).

These systems can be beneficial when they help authors identify missing evidence, clarify existing results, or respond more efficiently to reviewers. Their significance for production scaling is that the reduction in time and effort extends beyond the initial submission. Reviewer feedback can be processed and converted into a revised manuscript or rebuttal more quickly, allowing the same submission to move through evaluation and response cycles at a faster pace.

The difficulty of post-submission revision depends on what the reviewer asks the author to change. Some revisions mainly reorganize or improve the presentation of existing work, while others require new experiments, code, or scientific reasoning. APRES focuses on the former setting, keeping the core scientific content fixed while optimizing presentation; experts often preferred its revised versions (Zhao et al., 2026a). ReviseBench evaluates the harder setting, where agents must respond to reviews through writing, coding, and experimentation. Across 12 ICLR papers, model-generated revisions remained substantially behind human camera-ready versions and sometimes introduced incomplete experiments, formatting errors, or fabricated results (Luo et al., 2026). Together, these studies suggest that AI is currently better at revising how existing work is presented than at carrying out the new scientific work required to address substantive reviewer concerns.

Production scaling after submission is therefore uneven: AI can substantially reduce the cost of responding to evaluation, while revisions that require new evidence, experiments, or scientific judgment remain harder to automate.

3.4 From production scaling to evaluation pressure

The downstream consequence appears when this greater volume and faster rate of iteration reach an evaluation process whose capacity remains limited. The clearest evidence comes from settings where submission activity and reviewing workload can be observed directly. At Organization Science, Gartenberg et al. (2026) analyze 6,957 initial submissions and 10,389 textual reviews from 2021–2026. Submission volume increased by 42% after ChatGPT’s release, with much of the increase concentrated in manuscripts receiving high aggregate AI-assistance scores, while the journal experienced additional editorial and review load. Because this is an observational study of one journal

using a commercial detector, it does not establish a general causal effect. It nevertheless provides direct evidence that AI-associated production growth can translate into additional evaluation demand.

Evidence from ICLR points in the same direction. Using an extended submission series, Knight (2026) estimate that annual submission growth accelerated from 25% before ChatGPT to 59% afterward. In a sample of 18,787 submissions, they also estimate that the relationship between acceptance and later citation weakened. The authors further propose two related mechanisms: increased volume spreads evaluator attention more thinly, while cheaper linguistic polish reduces the usefulness of presentation quality as a signal for triage.

Beyond increasing evaluation volume, cheaper production can also make redundant or overlapping work harder to identify. In a biomedical corpus, Maupin et al. (2026) identify 411 effectively redundant papers among 3,465 studies of selected NHANES exposure-outcome pairs. In a separate demonstration, three syntactically different manuscripts could be generated in roughly two hours each while remaining below individual-source similarity warning thresholds. This illustrates a related problem: when producing alternative versions becomes cheap, evaluators may need to distinguish semantically redundant work even when surface-level similarity checks fail. Reports of reworded dual submissions provide a related conference-level example (Wagstaff et al., 2026).

Synthesis. In conclusion, the evidence shows that generative AI can make scholarly production both faster and more scalable, from initial research and manuscript preparation to later revision and rebuttal. Evaluation capacity, however, remains constrained by the time and expertise required to read, verify, and judge scientific work. Production scaling therefore shifts pressure downstream and creates growing demand to scale evaluation as well. The next section examines how AI is being used to support and automate scholarly evaluation.

4 Automating scholarly evaluation

As scholarly production scales, evaluation faces the same pressure to process more work with limited expert attention. AI offers one way to expand this capacity by supporting or automating parts of peer review. However, evaluation automation covers systems with very different levels of authority. An AI system may provide private feedback to authors, assist reviewers in preparing reports, contribute an official review, or support scoring, meta-review, and publication decisions. As AI moves closer to consequential decisions, the central question shifts from whether it can generate useful review text to whether its judgments are reliable enough for the authority it receives.

4.1 From review tasks to evaluative roles

The use of AI to scale evaluation builds on earlier computational work on peer review. PeerRead released 14.7K paper drafts, decisions, and 10.7K expert reviews, enabling tasks such as acceptance and aspect-score prediction (Kang et al., 2018). Later datasets supported review aggregation and meta-review generation (Shen et al., 2022; Li et al., 2023, 2024a). With the emergence of LLMs, the range of automated tasks has expanded further to submission checking, reviewer-paper matching, review drafting, claim verification, rebuttal support, meta-review, and decision assistance (Kuznetsov et al., 2024; Zhuang et al., 2025; Wu et al., 2026c; Du et al., 2024; Hossain et al., 2025; Yu et al., 2024a).

As this range expands, the more important distinction is not simply which task is automated, but how much evaluative authority the system receives. Author-facing feedback has no direct influence on publication decisions. Reviewer assistance may change an official report while leaving the reviewer responsible for the judgment. An additional AI review introduces a new evaluative voice into formal deliberation, while scoring, triage, and meta-review place the model closer to publication decisions. Capability at one level therefore does not automatically justify use at the next. Producing helpful feedback is different from identifying a fatal scientific flaw, and agreement with existing reviews is different from making a reliable publication decision.

4.2 From assistance to official evaluation

This progression in evaluative authority is already visible in current systems and deployments. The lowest-authority form of evaluation automation is feedback provided directly to authors. Liang

et al. (2024b) generated GPT-4 feedback for full manuscripts and compared it with reviews of 3,096 accepted papers across 15 Nature-family journals and 1,709 ICLR submissions. Model feedback showed overlap with human reviews, and 57.4% of 308 researchers in a prospective sample rated it helpful or very helpful. The system nevertheless struggled with deeper methodological criticism, suggesting that automated pre-review can expand access to feedback without reproducing the full role of expert review.

Other author-facing systems combine planning, literature search, and critique to suggest revisions or identify concerns later raised by reviewers (Chamoun et al., 2024; Wu, 2026). A NeurIPS 2024 deployment similarly used an LLM-based checklist assistant on 234 voluntary submissions from 184 papers. More than 70% of post-use respondents found the tool useful, and a similar share intended to revise their paper or checklist (Goldberg et al., 2024). Importantly, these systems remained outside the formal decision process.

Although author-facing feedback remains outside the formal decision process, it can still influence subsequent author behavior at scale. In a randomized experiment with 31,020 arXiv manuscripts, customized AI feedback increased the one-month revision rate by 12.55% relative to the control group (Wang et al., 2026a). The effect was larger for authors with less access to established feedback networks. This shows that low-authority AI evaluation can affect how authors respond to their work even without directly influencing publication decisions.

Reviewer-facing assistance moves one step closer to official evaluation. At ICLR 2025, Thakkar et al. (2026) conducted a large-scale randomized intervention in which an AI system provided optional suggestions for improving review reports. Among 44,831 reports, 18,946 received feedback, and 5,031 were revised, incorporating 12,222 suggestions. Revised reports were judged clearer and more informative, and the intervention increased engagement during rebuttal. The AI system, however, did not write the official review, change scores, or enter the acceptance decision. The experiment therefore shows that AI can change formal review practice at scale while leaving the final evaluative authority with human reviewers.

AAAI-26 moved further along this authority ladder. Every one of the 22,977 main-track papers entering full review received a clearly labeled AI review generated in less than 24 hours (Biswas et al., 2026). The system evaluated multiple aspects of the paper, used code and search tools, critiqued its own draft, and retained logs for oversight. Across 5,834 survey responses, participants often preferred the AI reports on technical accuracy, focus, and research suggestions.

The AI review still did not provide a score or acceptance recommendation and did not replace a human reviewer. It nevertheless became part of the official evaluation record and could be consulted by senior committee members. This deployment therefore marks an important transition: AI-generated evaluation has moved from private assistance into live conference workflows, even while venues continue to limit its formal decision authority.

4.3 Toward grounded and auditable evaluation

As AI receives greater evaluative responsibility, simply generating a plausible review becomes insufficient. Recent systems increasingly aim to make review judgments more structured, grounded, and auditable.

One direction is workflow decomposition. Instead of generating a review in a single step, systems divide evaluation into stages such as summarization, retrieval, critique, and decision. ReviewAgents follows a process of summarization, retrieval, critique, and decision, while Generative Agent Reviewers combines personas, claim graphs, simulated reports, and meta-review (Gao et al., 2025; Bougie and Watanabe, 2025). PeerCheck similarly shows that review quality depends on how reasoning and retrieval are incorporated, with effects varying across models (Chen et al., 2026d). Reviewer guidelines also influence model judgments: official conference guidelines can move model scores closer to human averages, while reviewer-imitating or rigid additive guidelines may perform less consistently (Li et al., 2026b). The evaluation workflow is therefore part of the reviewer itself rather than a neutral wrapper around the model.

A second direction is grounding judgments in evidence. Novelty assessment requires comparison with external literature (Afzal et al., 2026), while scientific support requires checking whether claims are justified by evidence inside the manuscript (Ballakuraya et al., 2026). Review grounding further

asks whether a criticism is supported, contradicted, or left unresolved by the paper (Ghorbanpour et al., 2026). ReviewGrounder and ScholarPeer combine paper-specific criteria with retrieval and specialist agents to support these functions (Li et al., 2026e; Goyal et al., 2026). The broader goal is to make criticism depend less on surface plausibility and more on traceable scientific evidence.

A third direction is auditability. EGTR-Review retains retrieved evidence, sources, and uncertainty while generating reviews (Qiu et al., 2026). ProReviewer represents review as a sequential investigation whose log records claims, questions, and observations that guide later inspection (Fang et al., 2026). DeepReviewer 2.0 similarly produces a claim-evidence-risk record with anchored criticism and repair actions (Weng et al., 2026). These systems move automated review away from reviewer imitation and toward evaluations whose supporting evidence and intermediate judgments can be inspected.

4.4 Automation already affects selection

Formal deployments capture only part of current evaluation automation because reviewers can also use general-purpose LLMs independently. A distributional analysis of ICLR, NeurIPS, CoRL, and EMNLP reviews estimates that 6.5–16.9% of post-ChatGPT review text was substantially AI-modified (Liang et al., 2024a). This suggests that AI-assisted evaluation entered review practice even before many venues established official systems.

More importantly, AI-associated reviewing is already linked to selection outcomes. Russo Latona et al. (2025) estimate that at least 15.8% of ICLR 2024 reviews were AI-assisted. Within-paper comparisons find these reports slightly more lenient, while matched analyses associate receiving an AI-assisted review with a 3.1 percentage-point increase in acceptance and a 4.9-point increase for borderline papers. Although the study relies on inferred AI use rather than direct observation, it shows that AI-associated reviewing can coincide with consequential differences in evaluation outcomes.

A larger analysis of more than 125,000 paper-review pairs further suggests that automation can change calibration and information aggregation (Sharma et al., 2026). Reviews classified as LLM-associated show greater leniency toward weaker papers, while fully generated reviews in synthetic experiments exhibit stronger rating compression. LLM-associated meta-reviews are also more likely to recommend acceptance at equivalent reviewer scores. These findings indicate that evaluation automation can affect not only how reviews are written, but also how judgments are distributed and combined.

4.5 The reliability bottleneck

Despite rapid progress in system design, reliable scientific judgment remains the central limitation. Specialized models, multi-agent workflows, retrieval, question decomposition, and reinforcement learning continue to improve review generation (Idahl and Ahmadi, 2025; D’Arcy et al., 2024; Zhu et al., 2025b; Chang et al., 2025; Zeng et al., 2025). However, conventional evaluation of AI reviewers often measures fluency, overlap with human reviews, or general helpfulness rather than whether the system correctly identifies and prioritizes scientifically important problems (Zhou et al., 2024a).

PRISM evaluates automated reviewers across analytical depth, novelty verification, flaw identification and prioritization, and constructiveness (Loc et al., 2026). Individual systems perform strongly on selected dimensions, but no single system reproduces the same balance across all of them. ReviewEval similarly evaluates multiple dimensions of generated reports (Garg et al., 2025), while CLAIMCHECK asks whether a criticism is actually grounded in the scientific claim it challenges (Ou et al., 2025). Together, these benchmarks shift attention from whether a review sounds plausible to whether its judgments are scientifically supported.

Results on criticism-level tasks are more mixed. In a study of 82 Nature-family papers with 45 expert scientists, GPT-5.2 outperformed the top-rated human review on a composite measure of correct, important, and well-supported criticisms (60.0% versus 48.2%) (Kim et al., 2026). However, its individual criticisms were less often correct, and AI reviews overlapped much more with one another than human reviews. These results show that strong performance on criticism generation can coexist with lower per-criticism reliability and less diversity across reviews.

Strong criticism generation, however, does not necessarily imply sensitivity to scientific soundness. Counterfactual evaluation reveals a sharper limitation. Across 133 accepted AI and NLP papers,

Dyck and Gurevych (2026) construct 931 variants, including 391 edits that break critical scientific support relations and 540 soundness-neutral controls. Across the tested automated reviewers, soundness-critical changes did not produce statistically significant differences in review aspects, sentiment, or scores compared with surface-level control edits. The models were also sensitive to irrelevant wording changes.

Direct error-detection benchmarks point to the same verification gap. On SPOT, which contains 91 author-validated errors from 83 published papers, no tested model exceeded 21.1% recall or 6.1% precision (Son et al., 2025). The models also rarely found the same error across repeated runs. In MLReplicate, automated review accepted 10 of 37 valid agent-generated papers, and 59% of the accepted papers contained fabricated or unsupported claims (Gaddipati et al., 2026). The benchmark covers only eight machine-learning tasks, but it shows that weak generation and weak evaluation can compound each other.

These findings narrow the evaluation bottleneck. Generating criticism is becoming cheap, and AI can perform well on some bounded tasks. The harder problem is to verify claims, identify the flaws that matter most, preserve different perspectives, and make accountable decisions.

Synthesis Evaluation automation can expand review capacity, and its role has already progressed from private feedback to assistance within formal review and official AI-generated reports. At the same time, greater authority places greater demands on the reliability of scientific judgment, which remains difficult to automate. Machine-mediated evaluation also makes parts of the review process more repeatable and easier to query. Once authors can observe and respond to these regularities, the evaluator itself becomes a target for optimization. Section 5 examines this next dynamic.

5 Evaluation manipulation

As AI-mediated evaluation becomes more repeatable and queryable, participants can begin to exploit its regularities to influence how their work is evaluated. This creates the third dynamic in our taxonomy: evaluation manipulation. Such manipulation can directly target an AI reviewer through hidden instructions, but it can also exploit presentation, framing, or other evaluator-visible signals that systematically affect review outcomes. Repeated access to evaluator feedback can further turn these signals into targets for iterative optimization.

5.1 Defining evaluation manipulation

Not every revision that improves an AI review is manipulation. Clearer writing, corrected analysis, additional evidence, or a better explanation of limitations can improve both the paper and its evaluation. The relevant distinction is whether a change primarily improves the scientific work or instead targets how the evaluator responds to it. Importantly, evaluation manipulation differs from evasion in both its target and its position in the response sequence. Manipulation targets the evaluator in order to shift its judgment without a corresponding improvement in the underlying scientific evidence. Evasion instead occurs after a defense or restriction is introduced and adapts behavior to avoid or circumvent that control. The same technique may therefore count as manipulation when it directly targets an evaluator, and as evasion when it is adopted in response to a safeguard.

This distinction is clearest for direct instructions to the reviewer, but becomes less obvious for presentation changes. Repositioning related work may genuinely clarify novelty or simply make the same contribution appear stronger. A revised discussion may explain a limitation more clearly or reduce the weight that an evaluator assigns to an unchanged weakness. Controlled comparisons that hold methods, data, figures, equations, and numerical results fixed while changing presentation provide one way to study this boundary (Yang et al., 2026c; Li et al., 2026c). The key question is therefore not simply whether a score changes, but what was changed and why the evaluator responded differently.

Controlled revision studies further illustrate this distinction. LLM-REVal finds that iterative model feedback can produce genuine corrections while also exposing preferences that authors can exploit, and Review Arcade shows that repeated revision against an LLM reviewer does not consistently improve evaluation across cases (Li et al., 2025b; Hatzel et al., 2026). A higher score alone does not show that a revision improved the science or manipulated the reviewer. The relevant distinction is

whether a revision improves the work or its assessability, or instead changes the evaluator’s judgment without a corresponding change in what the scientific evidence supports.

5.2 From explicit to subtle manipulation

The clearest form of evaluation manipulation directly targets the AI reviewer. A July 2025 audit identified 18 arXiv preprints containing concealed instructions, often presented as white or microscopic text, asking an AI reviewer to provide favorable assessments or particular review content (Lin, 2026). These observations show that attempts to manipulate AI-mediated review have already appeared in public manuscripts. Related work shows that such instructions can also be concealed through font encodings, multilingual text, or external resources (Collu et al., 2026; Choi et al., 2026; Theocharopoulos et al., 2025; Xiong et al., 2025).

Controlled studies show that these attacks can affect AI-review outcomes. Hidden instructions have increased ratings and reduced reported weaknesses across multiple review systems (Ye et al., 2024; Zhu et al., 2025a). Larger evaluations also reveal substantial variation across models: some tested reviewers remain highly vulnerable, while others are considerably more resistant (Li et al., 2026d). That is to say, a reviewer that is robust in one setting may still be vulnerable under a different model, prompt, or evaluation pipeline.

Manipulation also extends beyond a literal hidden sentence. Placement, layout, and domain-specific instructions can change attack effectiveness (Torrielli et al., 2026; Sahoo et al., 2025), while character-, word-, and sentence-level perturbations can influence reviewer behavior without remaining recognizable as explicit instructions (Lin et al., 2025). Scientific papers are also multimodal documents. PaperGuard shows that attacks can operate through figures as well as text, with figure-only perturbations changing aggregate reviewer scores in its benchmark (Zhao et al., 2026b). The relevant attack surface therefore includes the broader document presented to the evaluator, not only visible prose.

More importantly, evaluation manipulation does not require an explicit instruction. Meaning-preserving rewrites can also change AI-review outcomes. Abstract-level paraphrasing has produced score gains across different papers and reviewers while preserving the intended meaning of the text (Li et al., 2026c; Kaneko, 2026). Full-paper experiments make this effect more direct. Yang et al. (2026c) hold scientific evidence fixed while iteratively changing presentation based on AI-reviewer feedback. Across three reviewer models, the attack achieved a 75.1% success rate and increased scores by an average of 1.21 points on a ten-point scale. Changes to related-work positioning and analytical discussion were particularly effective, and criticisms of unchanged limitations sometimes weakened or disappeared.

Baumann et al. (2026) report a similar effect: rewriting papers without changing their scientific content increased AI-review scores by an average of 0.45 points. They also find that AI reviews are more similar to one another than human reviews, suggesting that susceptibility to presentation changes may coexist with reduced diversity in evaluation.

These results show a progression from explicit instructions to ordinary-looking changes in presentation. An evaluator may still recognize the same scientific weakness while changing how much that weakness affects its overall judgment. Related work on human review and LLM judges similarly identifies sensitivity to citation, position, rubrics, style, confidence, and framing (Stelmakh et al., 2023; Wang et al., 2024b; Ding et al., 2026a; Yang et al., 2026b; Wang et al., 2026b). Such sensitivities are not necessarily manipulation by themselves, but they create predictable signals that can be exploited.

A more severe adjacent failure occurs when convincing presentation is combined with unreliable or fabricated scientific content. BadScientist shows that agent-generated manuscripts without real experiments can nevertheless receive favorable assessments from LLM reviewers under controlled conditions (Jiang et al., 2026a). This differs from presentation-based manipulation, but it reinforces the broader concern that an evaluator can assign favorable judgments when plausible presentation is not matched by sufficient scientific verification.

5.3 From one-shot to adaptive manipulation

The more consequential shift occurs when manipulation becomes iterative. A one-time hidden instruction demonstrates that an evaluator can be manipulated. A repeatable evaluator, however, also

reveals whether a particular change worked. The manuscript can then be revised, evaluated again, and further adjusted based on the new response.

Controlled prompt-injection studies already demonstrate this process. Iterative attacks use reviewer scores or surrogate-model feedback to refine hidden instructions, producing stronger attacks that can transfer across evaluators (Zhou et al., 2025). The same principle applies to presentation. Repeated reviewer feedback can guide which sections to rewrite, which framing to emphasize, and which changes are most likely to alter the eventual assessment (Li et al., 2026c; Yang et al., 2026c).

This process does not require exact knowledge of the deployed evaluator. Aspect-level perturbation studies reveal systematic sensitivities in review models, while work on generic LLM judges shows that scalar evaluator feedback can be used to optimize transferable prompts without corresponding improvements in human judgments (Li et al., 2025a; Raina et al., 2024; Shi et al., 2024; Alazraki et al., 2026). Keeping the deployed model or prompt secret can make optimization more difficult, but shared preferences and cross-model transfer can still provide useful signals.

The central mechanism is therefore straightforward: repeated evaluation turns reviewer regularities into an optimization signal. Once this happens, manipulation is no longer limited to a fixed hidden prompt or a single rewrite. Authors or automated systems can search over instructions, wording, framing, layout, and other evaluator-visible features and retain changes that produce more favorable evaluations. Evaluation manipulation can therefore become adaptive rather than one-shot.

Synthesis. The evidence shows that AI-mediated evaluation can be manipulated through both explicit attacks and more subtle changes to evaluator-facing signals. Repeated access to evaluator feedback can further turn these sensitivities into an adaptive optimization process. Evidence that such manipulation changes official venue decisions remains limited, but the underlying mechanisms are already observed and experimentally reproducible. These risks create pressure for venues to protect the reliability of evaluation, motivating the technical and institutional defenses examined next.

6 Defense mechanisms and policy responses

As evaluation manipulation becomes observable and reproducible, defenses have expanded from detecting suspicious AI-generated content or hidden instructions to more targeted verification and more robust evaluation pipelines. These approaches aim not only to identify manipulation after it occurs, but also to reduce the vulnerability of AI-mediated evaluation itself. At the same time, conferences, journals, and publishers are increasingly incorporating such safeguards into review policies and procedures that define permitted AI use, accountability, verification, and enforcement.

6.1 From AI detection to targeted verification

A natural first response to AI-mediated risks is to detect whether AI was used. General-purpose AI-text detectors and review-specific classifiers attempt to distinguish human-written from AI-generated or AI-assisted text. In peer review, paper-conditioned similarity and review-specific methods can improve detection under low false-positive requirements (Yu et al., 2024b; Duarte et al., 2026). However, these systems infer an unobserved writing process from the final text, and their performance changes across models, editing styles, and operating thresholds.

Real venue use illustrates this limitation. In a NeurIPS 2026 position-paper audit, a default document-level detector marked 28.2% of 969 submissions with the maximum AI score, while a 100-word-window analysis reduced the high-confidence share to 12.7% (NeurIPS 2026 Communication Chairs, 2026). The large difference shows that detector outputs depend strongly on how the artifact is segmented and scored. Human judgment is also not a reliable substitute: experienced reviewers can struggle to distinguish human from AI-augmented writing (Hadan et al., 2024).

Detection becomes particularly difficult when policies allow some forms of AI assistance but prohibit others. On a dataset of 50,156 reviews spanning human and simulated human-AI collaboration, commercial detectors identified most fully generated reviews but also classified some compliant AI-polished reviews as fully AI-generated (Saha et al., 2026). This makes post-hoc attribution useful for screening, but much weaker as direct evidence of how a particular review was produced.

These limitations motivate a shift from passive detection toward targeted verification. Instead of asking whether a text merely looks AI-generated, a venue can introduce a known signal and test whether a prohibited process actually occurred. Rao et al. (2025), for example, randomly assign covert manuscript markers that cause an LLM generating a review to emit a rare phrase or citation. Random-citation markers appeared in 98.6% of tested LLM reviews on average and often survived one-shot paraphrasing. Unlike ordinary stylometry, this approach creates a controlled probe of a specific behavior rather than inferring provenance from writing style.

This idea has already moved into practice. An ICML 2026 report describes the use of randomized manuscript watermarks followed by manual verification and sanctions for reviewers who had agreed to a no-LLM policy (Agarwal et al., 2026; Saha et al., 2026). The mechanism does not detect every form of AI assistance, but it illustrates a broader change in defensive design: when attribution from text is unreliable, venues can verify narrower behaviors using controlled and auditable signals.

Related approaches extend targeted verification beyond randomized watermarks. IntraGuard embeds refusal instructions or markers in manuscripts to expose end-to-end LLM review outsourcing (Ma et al., 2026a), while other work studies watermark, refusal, and monitored-redirection mechanisms under different attack settings (Nemecsek et al., 2025; Torrielli et al., 2026). These methods can be effective against the behaviors they target, but rephrasing, modality changes, or jailbreaks can weaken them. Together, they reinforce the shift from inferring AI use from writing style toward testing specific behaviors through controlled signals.

6.2 Protecting AI-mediated evaluation

Detection alone does not protect an AI reviewer from manipulation. A second line of defense therefore focuses directly on the path from manuscript to evaluation.

The first step is to control the input presented to the evaluator. Hidden or microscopic text can be identified by comparing rendered and extracted content, suspicious metadata and external resources can be isolated, and the manuscript can be treated as untrusted data rather than as an extension of the reviewing prompt. These measures can block common prompt-injection channels, although they cannot address manipulation that operates through legitimate-looking presentation alone.

More targeted systems explicitly test the evaluator against adversarial inputs. A detection prompt in Zhou et al. (2025) identifies many non-adaptive hidden instructions and moves reviewer scores back toward their baseline, but its effectiveness drops when the attacker is allowed to adapt to the defense. SafeReview addresses this problem by jointly exposing the reviewer to progressively stronger attacks during training (Xin et al., 2026). Under the evaluated attacker configurations, it preserves paper ranking better than static adversarial training and avoids some of the excessive blocking produced by generic prompt detectors. The broader lesson is that defenses against adaptive manipulation must themselves be evaluated under adaptive attacks.

Long and multimodal papers create additional difficulties. PaperGuard searches manuscript passages and figures for instruction-like content and then verifies the retrieved candidates (Zhao et al., 2026b). Artifact-specific methods similarly target fabricated scientific tables or figures (Huang et al., 2026; Hu et al., 2026). These tools broaden the protected surface beyond plain text, but they also illustrate the limit of attack-specific detection: presentation-level manipulation may contain no discrete malicious string or artifact to identify.

Defenses may also need to examine the research process, not just the final paper. Verification-first proposals argue that AI should produce evidence and checks that can be audited, rather than imitate review scores or final decisions (You et al., 2026). A controlled study supports this direction: adding code, logs, and execution traces raised audit accuracy from 55% to 82% and exposed errors that were hard to see in the paper alone (Luo et al., 2025). These results extend defensive verification beyond the submitted manuscript to the evidence and intermediate artifacts that support it.

6.3 Institutional and policy responses

Technical safeguards operate within venue rules that determine which uses of AI are allowed and how violations are handled. Current policies vary widely. Some venues prohibit reviewers from uploading confidential manuscripts to external models, some permit limited language assistance, others require disclosure, and some provide bounded venue-controlled tools while keeping human

reviewers responsible for the final report (ICLR 2026 Program Chairs, 2025a; ACL Rolling Review, 2026b; Association for Computational Linguistics, 2026; EMNLP 2026, 2026a). Similar variation appears across journals: among the top 100 medical journals, 78 provided AI guidance in 2024, with policies ranging from prohibition to limited permitted use (Li et al., 2024b).

Policy studies show large differences in whether guidance exists, where it appears, and what authors must disclose. In 2023, 24% of the 100 largest publishers and 87% of 100 highly ranked journals had generative-AI guidance (Ganjavi et al., 2024). Another study found public AI policies at 50 of 59 disaster-research journals; all allowed some AI use but imposed different conditions (Armitage, 2026). Guides and expert-consensus work generally call for clear disclosure, human verification, and author responsibility (Cleland et al., 2026; Fettiplace et al., 2026). These studies describe policy design, not policy effectiveness.

Conference policies make these differences concrete. AISTATS 2026 permits author-side AI use under human responsibility but prohibits reviewer-side LLM use and treats favorable hidden instructions as misconduct. COLM requires disclosure of substantive AI assistance by authors and reviewers, while KDD combines review disclosure with restrictions on sending manuscript text to external services (AISTATS 2026, 2026; Conference on Language Modeling, 2026; KDD 2025, 2025). Other venues combine automated screening with human follow-up: EMNLP routes integrity flags to senior verification, while ICLR uses automated triage together with multi-person checking and appeals (ECCV 2026, 2026; EMNLP 2026, 2026b; ICLR 2026 Program Chairs, 2025b, 2026).

Venues are also beginning to constrain the role AI plays rather than treating automation as all-or-nothing. Author-facing preflight systems have checked formal requirements and presentation before review, while reviewer-facing tools such as the ICLR feedback system and ARR’s self-hosted assistant operate within deliberately limited roles (Jayaram et al., 2026; Ohta et al., 2026; Chen et al., 2026b; ACL Rolling Review, 2026a). These deployments keep particular inputs and outputs under venue control and preserve an identifiable human decision point.

Whether these policies materially change author behavior remains less clear. Comparative evidence finds similar growth in detected AI-assisted writing across journals with and without AI policies, but the causal effect of policy adoption is difficult to establish without adoption dates and pre-policy trends (He and Bu, 2026a; Wang, 2026; He and Bu, 2026b). Current evidence therefore documents the spread of AI policies more clearly than their effectiveness.

Institutional responses also address the pressure created by scaling production. Submission limits, reciprocal-review requirements, desk screening, and workload-aware assignment can reduce evaluation demand or increase review supply without requiring a venue to determine whether a submission was AI-generated (Computer Vision Foundation, 2025; AISTATS 2026, 2026; Conference on Language Modeling, 2026; KDD 2025, 2025). These mechanisms can relieve capacity pressure, but they can also introduce trade-offs in access, workload, and fairness (Cao et al., 2025).

Synthesis. Defense mechanisms have evolved from passive AI detection toward targeted verification, more robust evaluation pipelines, and formal policy responses. Technical safeguards can identify suspicious behavior or reduce evaluator vulnerability, but they increasingly operate within venue-controlled processes in which human reviewers, chairs, or editors retain responsibility for consequential decisions. Conferences and publishers therefore combine technical safeguards with bounded AI roles, disclosure requirements, verification, and enforcement procedures. These interventions also change the costs and incentives faced by participants, creating opportunities for evasion and new side effects examined in the next section.

7 Evasion and side effects

Defenses change the costs and constraints faced by participants, but they do not necessarily remove the underlying incentive to use or optimize AI. Once a particular signal, tool, or behavior becomes subject to detection or restriction, participants can adapt by changing how that behavior is carried out or how it appears to the evaluator. At the same time, defensive measures can create unintended consequences even when no strategic evasion occurs. We therefore distinguish two related outcomes: *evasion*, in which actors adapt in response to a control, and *side effects*, in which the control itself shifts errors, labor, or risk elsewhere in the scholarly process.

7.1 Evasion after defensive intervention

A defense can itself become a target once its output has consequences. AI-text detectors, for example, rely on statistical patterns that distinguish generated from human-written text, but these patterns can change across models, decoding strategies, domains, and editing processes. Large-scale benchmarks show substantial performance degradation under distribution shifts and adversarial modification, while studies of scientific writing similarly find that relatively small changes in wording or form can alter detector outputs (Dugan et al., 2024; Odri and Yoon, 2023; Taloni et al., 2024; Liu et al., 2024; Matsubara, 2025; Shahriar et al., 2025). Once detection affects screening or enforcement, users therefore have an incentive to reduce or remove the signals on which it relies.

Controlled studies demonstrate many ways to do so. Post-generation methods use paraphrasing, rewriting, watermark modification, and transferable substitutions to change detectable traces while preserving much of the original content (Krishna et al., 2023; Sadasivan et al., 2025; Wang et al., 2024a; Fang et al., 2025; Dathathri et al., 2024). Other approaches adapt generation itself through evasive prompts, detector-guided generation, style optimization, humanization, or semantics-constrained optimization (Kumarage et al., 2023; Zhou et al., 2024b; Wang et al., 2025; Meng et al., 2025; Pedrotti et al., 2025; Gu et al., 2026; Wang et al., 2026c). Peer-review-specific detectors provide more specialized signals, but mixed human-AI writing, paraphrasing, and new generators continue to make attribution difficult (Kumar et al., 2024, 2025; Chen et al., 2026c; Saha et al., 2026). Similar rewriting can also weaken plagiarism and provenance signals while preserving substantial semantic reuse (Gupta and Pruthi, 2025; Hassanipour et al., 2024; Murdock et al., 2025; Dilber and Yoşumaz, 2026). Across these settings, the common pattern is that the underlying content can remain similar while the signal targeted by the defense changes.

Stronger defenses may in turn create new opportunities for adaptation. Randomized manuscript probes, for example, provide more direct evidence of particular forms of LLM review delegation than ordinary text detection (Rao et al., 2025). Once such mechanisms are known, sanitization, rewriting, alternative document-processing routes, or different models become plausible responses. Related studies on presentation and multimodal manipulation further show that evaluator-facing behavior can move beyond the particular textual signals targeted by a defense (Yang et al., 2026c; Zhao et al., 2026b). These results point to plausible directions in which future evasion may develop.

This distinction also highlights an important gap in the current literature. Controlled experiments provide substantial evidence that many defensive signals can be circumvented, but direct evidence of post-policy adaptation in real scholarly settings remains limited. Existing studies rarely follow the full sequence in which a venue introduces a specific defense, participants change their behavior in response, and the venue subsequently adjusts its safeguards. The mechanisms of evasion are therefore increasingly well understood, while longitudinal evidence of defense-driven counter-adaptation remains an open empirical question.

7.2 Side effects of defensive measures

Defensive measures can also create costs even when participants are not actively trying to evade them. AI-content detection provides the clearest example. False positives and uneven error rates can cause legitimate human or AI-assisted writing to be flagged, with particular concerns for non-native-English writers and other populations whose writing differs from a detector’s training distribution (Liang et al., 2023; Pratama, 2025; Boonsuk, 2026; Ali and Shaban, 2026). When detector outputs are incorporated into editorial or integrity processes, these errors can become more than classification mistakes.

Automated evaluation can introduce a separate source of unequal treatment. Controlled studies show that LLM reviewers may respond to evaluator-irrelevant author information such as institutional affiliation, seniority, or publication history (Vasu et al., 2026). Detection bias and evaluation bias arise at different stages of the process, but both illustrate the same broader concern: automation can redistribute errors rather than simply reduce human workload or inconsistency.

The same pattern applies to labor and risk. AI assistance may reduce some drafting effort while creating new work in checking generated content, verifying evidence, resolving disagreements, investigating suspicious cases, or reviewing appeals. Studies of AI-assisted reviewing report reductions in perceived workload alongside continued verification and oversight requirements (Chen et al., 2025; Ebadi et al., 2025). The use of external AI services can also create confidentiality concerns when

unpublished manuscripts or review materials leave venue-controlled infrastructure. Automation can therefore shift work and risk to reviewers, chairs, editors, integrity staff, or authors rather than eliminating them.

Taken together, these findings show that the effects of a defense extend beyond whether it successfully detects or prevents the targeted behavior. Even a technically effective intervention can introduce false positives, additional verification work, confidentiality risks, or unequal burdens across participants. Defensive measures should therefore be evaluated not only by how well they address the original threat, but also by the costs and risks they create elsewhere in the review process.

Synthesis. Current evidence supports both sides of this dynamic. Controlled studies show that detection and other technical defenses can be circumvented through rewriting, generation-time adaptation, tool switching, and changes in the signals presented to evaluators. At the same time, deployed and observational evidence shows that defensive measures can introduce false positives, unequal errors, additional human work, and confidentiality risks. What remains under-studied is the full longitudinal sequence from a deployed defense to observed participant evasion and a subsequent institutional response. Defenses therefore do not simply remove risk: they can change both participant behavior and where the remaining costs are borne. These effects can also persist beyond a single review cycle through the papers, reviews, and decisions that enter the scholarly record, which we examine next.

8 Long-horizon ecosystem feedback

The preceding dynamics largely unfold within a single research and review cycle. Long-horizon ecosystem feedback begins when the artifacts produced in that cycle persist and are reused by later research or evaluation systems. Papers, reviews, rebuttals, decisions, corrections, and citations can enter searchable corpora, retrieval systems, training pipelines, scientific agents, and AI evaluators. Through this reuse, information and judgments produced in one round can influence the inputs, objectives, or behavior of later systems. The key question is therefore not simply whether AI-associated material enters the scholarly record, but whether that record becomes part of future production and evaluation.

8.1 Scholarly artifacts enter the persistent record

The first requirement for long-horizon feedback is persistence. Corpus studies and targeted audits show that AI-associated material is already entering searchable and published scholarship. Large-scale analyses identify substantial post-2022 changes in scientific writing associated with LLM use, while more targeted studies document visible generated fragments in indexed publications (Schmidt and Meir, 2024; Siler, 2026; Strzelecki, 2025; Haider et al., 2024; Yao et al., 2026; Spinellis, 2025). These studies differ in how they identify AI involvement, but together they establish a basic point: the contemporary scholarly record increasingly contains material shaped by generative AI.

Errors can persist as well. Studies of generated scientific writing report fabricated or inaccurate references, and large corpus audits identify nonexistent citations in published or publicly available research records (Walters and Wilder, 2023; Bhattacharyya et al., 2023; Topaz et al., 2026; Zhao et al., 2026c). Once such references are indexed or repeated by later papers, they can acquire the appearance of a citation history and become harder to correct across the broader record (Camp et al., 2025). Long-horizon feedback therefore begins with a simple but consequential fact: both useful AI-assisted content and identifiable errors can survive beyond the local research or review cycle.

8.2 Scholarly records become inputs to future systems

Persistence matters because scholarly artifacts are increasingly reused by AI systems. Published papers can become part of pretraining corpora or inference-time retrieval systems. Dolma, for example, contains tens of millions of scholarly papers used in the documented training mixture for OLMo, while OpenScholar retrieves from a large collection of open-access papers to support literature synthesis (Soldaini et al., 2024; Groeneveld et al., 2024; Asai et al., 2026). Other open corpora similarly incorporate sources such as arXiv and PubMed, making scholarly material an explicit component of later model development (Kandpal et al., 2026). Studies of memorization

further show that identifiable information acquired during continued pretraining can persist through subsequent model adaptation (Li et al., 2026a). These systems are generally designed to improve scientific reasoning and access to knowledge, but they also make the dependency on prior scholarly records explicit.

Reviews and publication outcomes provide an even more direct feedback channel because they encode judgments rather than scientific content alone. Peer-review datasets have long made papers, reviews, and decisions available for computational modeling (Kang et al., 2018; Yang et al., 2026a). More recent systems train directly on large collections of conference reviews or use review-rebuttal interactions and author responses as supervision for future feedback generation (Idahl and Ahmadi, 2025; Wu et al., 2026d; Mun et al., 2026). In this setting, historical reviews can teach future evaluators what to criticize, how to score, and which forms of feedback are associated with author response.

Some systems already connect these records back to future production or evaluation. SPARK trains an evaluator on large-scale OpenReview data and uses it to filter newly generated research ideas (Sanyal et al., 2025). Historical review patterns therefore influence which future ideas are retained before they become papers. ReviewGuard uses papers, reviews, decisions, and later citation outcomes to train and optimize a future evaluator (Rasool et al., 2026). Although citation is an imperfect measure of scientific value, the system demonstrates the relevant interface: outcomes from one round of scholarly evaluation can become objectives for a later evaluator. A related system extends this pathway back into research production. Scientific Judge learns from citation-based preferences, and Scientific Thinker then uses that evaluator as a reward model for generating new research ideas (Tong et al., 2026). This provides a more direct record-to-evaluation-to-production pathway, although the proposed ideas are evaluated primarily by LLMs.

These examples establish that scholarly records are not merely archives. Papers can become future model context, while reviews, responses, decisions, and citations can become supervision or optimization signals. The reuse pathway required for long-horizon feedback is therefore already operating in several production and evaluation systems.

8.3 From reuse to ecosystem feedback

Reuse creates the possibility that particular errors, manipulations, or broader statistical patterns propagate into later systems. Controlled studies show that even a small number of strategically inserted artifacts can sometimes have disproportionate downstream influence. In a biomedical knowledge-graph experiment, a single malicious abstract substantially altered downstream drug-disease rankings (Yang et al., 2024). Related work on retrieval and model training shows that injected or synthetic information can redirect RAG outputs, alter medical model claims, or dominate retrieval results (Zou et al., 2025; Alber et al., 2025; Yu et al., 2026). These studies do not demonstrate that manipulated scholarly papers have already produced the same effects in deployed scientific systems, but they establish mechanisms through which persistent artifacts could influence later retrieval and reasoning.

At a broader level, repeated reliance on AI may also shape the distribution of future research. Messeri and Crockett (2024) highlight the risk that heavy reliance on AI could narrow research questions, methods, and viewpoints while creating an illusion of understanding. Controlled studies provide related mechanisms: repeated training on generated content can amplify existing patterns and reduce diversity (Shumailov et al., 2024; Alemohammad et al., 2024; Seddik et al., 2024; Guo et al., 2024; Doshi and Hauser, 2024), while other work identifies correlated errors and increasing similarity across models (Kim et al., 2025; Jiang et al., 2026b). Evidence closer to scholarly production is observational: AI use has been associated with changes in citation patterns and collective topical breadth (McCreery and Naser, 2026; Hao et al., 2026). Together, these studies identify plausible mechanisms and early signals of distribution-level feedback, but they do not establish that scholarly publishing has entered a closed recursive degradation cycle.

The central empirical gap is therefore provenance across generations of systems. Current evidence shows that AI-associated material enters the scholarly record, that papers and reviews are reused by later systems, and that such reuse can influence retrieval, training, idea selection, and evaluation objectives. What has not yet been traced is a complete artifact-level path in which a specific AI-mediated paper or review enters a known corpus, materially changes a named successor system,

and then alters a later paper, review, or publication decision. Establishing such lineage remains an important open empirical problem.

Synthesis. Long-horizon ecosystem feedback is therefore partially established rather than fully closed. Record ingress is observed, and the reuse of scholarly papers, reviews, responses, and outcomes in later AI systems is already implemented. Controlled and adjacent studies further show mechanisms through which reused artifacts or repeated patterns can influence later systems. What remains missing is end-to-end evidence linking an identified artifact through a specific successor system to a later scholarly outcome. The final dynamic is thus best understood as a documented exposure and reuse pathway whose complete causal feedback loop remains to be demonstrated.

9 Cross-cutting findings and research agenda

All the previous dynamics reveal several properties of AI-mediated scholarly publishing that are difficult to see when production, evaluation, manipulation, and governance are studied separately. Three findings recur across the evidence: trustworthy evaluation remains harder to scale than research production, static evaluations become less informative once participants can adapt to AI-mediated review, and safeguards can shift rather than eliminate errors, labor, and risk. These findings also identify where the current evidence remains incomplete and where future research is most needed.

9.1 Trustworthy evaluation remains the bottleneck

Generative and agentic systems can increasingly reduce the time and effort required to produce, revise, and evaluate research artifacts. The corresponding work of verifying scientific claims, identifying consequential flaws, assessing novelty, and making accountable decisions remains much harder to scale. Evidence from journals and conferences already connects growing production volume with pressure on editorial and review capacity, while automated review systems can generate feedback at substantially lower marginal cost (Gartenberg et al., 2026; Knight, 2026; Biswas et al., 2026).

How humans and AI are combined also matters. In a randomized study of 103 teams assessing research reproducibility, human-only and AI-assisted teams achieved similar performance (94% and 91%), whereas AI-led teams achieved about 37% (Brodeur et al., 2026). Human-only teams also found more coding errors. The result further suggests that increasing AI involvement does not automatically reduce the need for human control in verification-intensive tasks.

The resulting bottleneck is therefore not simply review generation, but trustworthy scientific evaluation. Faster review generation is useful only when the resulting judgments remain grounded and when the human effort required for verification, disagreement resolution, and oversight does not grow at the same rate. Future deployments should therefore measure evaluation capacity in terms of both throughput and the human work required to maintain reliable decisions.

9.2 Adaptation weakens static evaluation

AI reviewers are commonly evaluated on fixed collections of manuscripts, where model judgments are compared with historical reviews, scores, or benchmark labels. The manipulation literature shows why these evaluations can become less informative after deployment. Once authors can observe or repeatedly query an AI-mediated evaluator, they can adapt instructions, wording, presentation, or other evaluator-facing signals in ways that change the resulting judgment (Zhou et al., 2025; Yang et al., 2026c; Li et al., 2026c).

Evaluation should therefore test not only performance on a fixed distribution, but also whether judgments remain reliable under adaptation. Relevant questions include whether semantically similar papers receive consistent evaluations, whether important scientific weaknesses remain influential after presentation changes, and whether robustness persists when participants receive feedback and can change their strategies (Dyck and Gurevych, 2026; Jiang et al., 2026a). The central issue is whether an evaluator continues to distinguish stronger scientific work from more effectively optimized presentation once its behavior becomes part of the environment.

9.3 Safeguards redistribute risk and burden

The evidence also shows that defensive measures rarely remove risk without changing where it appears. Detection and verification can make particular forms of AI use or manipulation more costly, but they can also motivate evasion. More aggressive enforcement can introduce false positives, while AI-assisted evaluation can create additional verification work, confidentiality concerns, and unequal effects across participants (Liang et al., 2023; Saha et al., 2026; Boonsuk, 2026; Vasu et al., 2026; Chen et al., 2025).

Safeguards should therefore be evaluated by more than their ability to detect or block a targeted behavior. A useful intervention must also account for the human work needed to verify its outputs, the errors it introduces, the participants who bear those errors, and how behavior changes after the intervention is deployed. This is especially important because sophisticated participants may be able to adapt to a control while ordinary users continue to bear its false-positive or compliance costs.

9.4 Research agenda

These findings point to a small number of empirical questions that remain unresolved across the current literature. Rather than requiring another set of isolated capability benchmarks, they require studies that directly test the response relations represented in the taxonomy. The four priorities below follow the open links identified by that mapping.

Production scaling and evaluation pressure. Existing studies associate AI adoption with greater scholarly output, faster production cycles, and growing submission volume, but the size and generality of the resulting evaluation burden remain uncertain. Multi-venue longitudinal studies with stronger measures of AI use could examine how changes in production affect submission volume, reviewer demand, desk-screening effort, review completion, and the distribution of evaluator attention.

AI-mediated evaluation and author adaptation. Controlled studies show that AI reviewers can be manipulated, but there is much less evidence about how authors change their behavior when they know that AI is part of a real evaluation process. Future studies could compare human-only and AI-mediated evaluation settings and measure changes in manuscript presentation, revision behavior, evaluator queries, scientific content, and review outcomes.

Defense and post-policy evasion. The literature demonstrates many mechanisms for circumventing detectors and other safeguards, but rarely observes the full sequence after a real policy intervention. Longitudinal studies should track whether a specific venue defense changes participant behavior, which evasion strategies emerge, what false positives or additional burdens result, and how the venue subsequently adjusts its response.

Scholarly records and successor systems. Long-horizon feedback remains the least directly traced part of the process. Future work should preserve versioned provenance across scholarly corpora and AI systems so that researchers can determine whether a particular paper, review, correction, or decision entered a later retrieval or training pipeline and materially changed subsequent generation or evaluation. Such studies would distinguish documented reuse from a genuinely closed feedback loop.

These directions shift the research agenda from evaluating isolated AI components toward studying the response relations that connect production, evaluation, manipulation, defense, evasion, and long-horizon feedback. Progress will depend especially on evidence that follows these interactions in real scholarly settings and over time.

10 Limitations

This survey covers a rapidly changing area, and many relevant studies remain available only as preprints. Our structured literature search covered work available through July 1, 2026. We then conducted a targeted update during the survey development. Because this update did not rerun every search query, coverage of the newest literature may be uneven across taxonomy categories.

The evidence base also includes surveys, position papers, policy documents, comments, replies, and conceptual analyses. Surveys and conceptual sources are used primarily to characterize terminology,

proposals, and ongoing debates, while official policy documents provide evidence of announced institutional responses. Neither type is treated as evidence of real-world prevalence, causal effects, or policy effectiveness unless supported by corresponding empirical data.

Real-world evidence is also uneven across the six dynamics. Large-scale deployments and institutional records are concentrated in a relatively small number of AI and computer-science conferences, OpenReview-based settings, and selected journals, while many attack, defense, and long-horizon feedback studies remain controlled experiments or early demonstrations. Findings from these settings may therefore not transfer directly to disciplines with different review structures, publication timelines, or norms of AI use.

Finally, the paper-review arms-race framing emphasizes adaptation and counter-adaptation between scholarly actors. It is intended as a lens for organizing these interactions rather than as a claim that all AI-assisted research, reviewing, revision, or institutional change is adversarial. Many uses of AI can improve access, efficiency, communication, and scientific quality, and the relevance of the framing depends on whether one actor’s behavior meaningfully changes the incentives or responses of another.

11 Conclusion

This survey presents AI-mediated scholarly publishing as a connected process rather than a set of isolated tools. We organize the literature around six linked dynamics spanning production scaling, evaluation automation, evaluation manipulation, defense mechanisms and policy responses, evasion and side effects, and long-horizon ecosystem feedback. Together, these dynamics show how changes in one part of the publishing process can reshape the incentives, constraints, and behavior of others. As AI becomes more deeply embedded in research and review, the central challenge is therefore not only whether individual systems perform well, but how they interact and adapt within the broader scholarly ecosystem. Future work should increasingly evaluate these relationships in real settings and over time.

References

- Anna Abalkina. Publication and collaboration anomalies in academic papers originating from a paper mill: Evidence from a russia-based paper mill. *Learned Publishing*, 36(4):689–702, 2023.
- ACL Rolling Review. Using the REVAS Review Assistant Tool in the ARR-May Cycle, 2026a. URL <https://aclrollingreview.org/revas-may26>.
- ACL Rolling Review. ARR Reviewer Guidelines, 2026b. URL <https://aclrollingreview.org/reviewerguidelines>.
- Osama Mohammed Afzal, Preslav Nakov, Tom Hope, and Iryna Gurevych. Beyond" not novel enough": Enriching scholarly critique with llm-assisted feedback. In *Proceedings of the 19th Conference of the European Chapter of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2648–2671, 2026.
- Alekh Agarwal, Miroslav Dudik, Sharon Li, Martin Jaggi, Nihar B. Shah, Katherine Gorman, and Gautam Kamath. On Violations of LLM Review Policies, 2026. URL <https://blog.icml.cc/2026/03/18/on-violations-of-llm-review-policies/>.
- AISTATS 2026. Call for Papers, 2026. URL <https://virtual.aistats.org/Conferences/2026/CallForPapers>.
- Lisa Alazraki, Yi-Chern Tan, Jon Ander Campos, Maximilian Mozes, Marek Rei, and Max Bartolo. Reverse engineering human preferences with reinforcement learning. *Advances in Neural Information Processing Systems*, 38:101173–101210, 2026.
- Daniel Alexander Alber, Zihao Yang, Anton Alyakin, Eunice Yang, Sumedha Rai, Aly A Valliani, Jeff Zhang, Gabriel R Rosenbaum, Ashley K Amend-Thomas, David B Kurland, et al. Medical large language models are vulnerable to data-poisoning attacks. *Nature Medicine*, 31(2):618–626, 2025.
- Sina Alemohammad, Josue Casco-Rodriguez, Lorenzo Luzi, Ahmed Imtiaz Humayun, Hossein Babaei, Daniel LeJeune, Ali Siahkoohi, and Richard Baraniuk. Self-consuming generative models go mad. In *International Conference on Learning Representations*, volume 2024, pages 53581–53608, 2024.
- Sayed Ibrahim Ali and Mostafa Shaban. Stigmatization of ai-assisted research in nursing academia: Moral judgments and epistemic anxiety among nurse academics. *Nurse Education Today*, page 107219, 2026.
- Richard Armitage. Artificial intelligence policies in disaster research: a review of journals and their publishers. *International Journal of Disaster Risk Reduction*, page 106078, 2026.
- Akari Asai, Jacqueline He, Rulin Shao, Weijia Shi, Amanpreet Singh, Joseph Chee Chang, Kyle Lo, Luca Soldaini, Sergey Feldman, Mike D’arcy, et al. Synthesizing scientific literature with retrieval-augmented language models. *Nature*, 650(8103):857–863, 2026.
- Association for Computational Linguistics. ACL Policy on Publication Ethics, 2026. URL https://www.aclweb.org/adminwiki/index.php/ACL_Policy_on_Publication_Ethics.
- Jinheon Baek, Sujay Kumar Jauhar, Silviu Cucerzan, and Sung Ju Hwang. Researchagent: Iterative research idea generation over scientific literature with large language models. In *Proceedings of the 2025 conference of the nations of the Americas chapter of the association for computational linguistics: human language technologies (volume 1: long papers)*, pages 6709–6738, 2025.
- Ranjitha Shivaprasad Ballakuraya, Arash Mahyari, and Ashok Srinivasan. Do methods support the claims? intra-paper verification for peer review. *arXiv preprint arXiv:2607.26066*, 2026.
- Joachim Baumann, Jiaxin Pei, Sanmi Koyejo, and Dirk Hovy. Stop automating peer review without rigorous evaluation. *arXiv preprint arXiv:2605.03202*, 2026.
- Mehul Bhattacharyya, Valerie M Miller, Debjani Bhattacharyya, Larry E Miller, and Valerie Miller. High rates of fabricated and inaccurate references in chatgpt-generated medical content. *Cureus*, 15(5), 2023.

- Joydeep Biswas, Sheila Schoepp, Gautham Vasan, Anthony Opiari, Arthur Zhang, Zichao Hu, Sebastian Joseph, Matthew Lease, Junyi Jessy Li, Peter Stone, et al. Ai-assisted peer review at scale: The aaai-26 ai review pilot. *arXiv preprint arXiv:2604.13940*, 2026.
- Yusop Boonsuk. “when the editor detected ai—but the ‘ai’ was me”: Algorithmic misjudgement and the crisis of authorship in tesol publishing. *Journal of Second Language Writing*, 73:101335, 2026.
- Nicolas Bougie and Narimawa Watanabe. Generative reviewer agents: Scalable simulacra of peer review. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing: Industry Track*, pages 98–116, 2025.
- Abel Brodeur, David Valenta, Alexandru Marcoci, Juan P Aparicio, Derek Mikola, Bruno Barbarioli, Rohan Alexander, Lachlan Deer, Tom Stafford, Lars Vilhuber, et al. Ai-assisted teams outperform ai-led teams but not human-only teams in assessing research reproducibility in quantitative social science. *Proceedings of the National Academy of Sciences*, 123(22):e2524747123, 2026.
- Nathan T Camp, Jason A Bengtson, and John C Sandstrom. The citation catastrophe: Propagation of ai-generated counterfeit citations in scholarship. *The Journal of Academic Librarianship*, 51(4): 103065, 2025.
- Yuefan Cao, Xiaoyu Li, Yingyu Liang, Zhizhou Sha, Zhenmei Shi, Zhao Song, and Jiahao Zhang. Dissecting submission limit in desk-rejections: A mathematical analysis of fairness in ai conference policies. *arXiv preprint arXiv:2502.00690*, 2025.
- Eric Chamoun, Michael Schlichtkrull, and Andreas Vlachos. Automated focused feedback generation for scientific writing assistance. In *Findings of the Association for Computational Linguistics: ACL 2024*, pages 9742–9763, 2024.
- Yuan Chang, Ziyue Li, Hengyuan Zhang, Yuanbo Kong, Yanru Wu, Hayden Kwok-Hay So, Zhijiang Guo, Liya Zhu, and Ngai Wong. Treereview: A dynamic tree of questions framework for deep and efficient llm-based scientific peer review. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pages 15662–15693, 2025.
- Nuo Chen, Andre Lin HuiKai, Jiaying Wu, Junyi Hou, Zining Zhang, Qian Wang, Xidong Wang, and Bingsheng He. Xtragpt: Context-aware and controllable academic paper revision via human-ai collaboration. In *Proceedings of the 64th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1045–1074, 2026a.
- Shiping Chen, Duncan Brumby, and Anna Cox. Envisioning the future of peer review: Investigating llm-assisted reviewing using chatgpt as a case study. In *Proceedings of the 4th Annual Symposium on Human-Computer Interaction for Work*, pages 1–18, 2025.
- Shiping Chen, Shu Zhong, Duncan P Brumby, and Anna L Cox. What happens when reviewers receive ai feedback in their reviews? In *Proceedings of the 2026 CHI Conference on Human Factors in Computing Systems*, pages 1–19, 2026b.
- Yihan Chen, Jiawei Chen, Guozhao Mo, Xuanang Chen, Ben He, Xianpei Han, and Le Sun. Coconuts: Concentrating on content while neglecting uninformative textual styles for ai-generated peer review detection. In *Proceedings of the 64th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 26921–26950, 2026c.
- Zeyuan Chen, Ziqing Yang, Yihan Ma, Michael Backes, and Yang Zhang. Peercheck: Enhancing llm-generated academic reviews towards human-level quality. In *Findings of the Association for Computational Linguistics: ACL 2026*, pages 23362–23386, 2026d.
- Byungjin Choi, Tae Joon Jun, Joung Won Sung, Il Woo Park, Jeong-Moo Lee, Soo Ick Cho, Hyung Jun Park, Ro Woon Lee, and Jungyo Suh. Invisible text injection and peer review by ai models. *JAMA Network Open*, 9(1):e2552099, 2026.
- J Cleland, E Driessen, K Masters, L Lingard, and LA Maggio. When and how to disclose ai use in academic publishing: Amee guide no. 192. *Medical Teacher*, 48(4):542–553, 2026.

- Matteo Gioele Collu, Umberto Salviati, Roberto Confalonieri, Mauro Conti, and Giovanni Apruzzese. Misleading large language models used (or misused) in scientific peer-reviewing via hidden prompt-injection attacks. *ACM Transactions on AI Security and Privacy*, 2026.
- Computer Vision Foundation. CVPR 2025 Changes, 2025. URL <https://cvpr.thecvf.com/Conferences/2025/CVPRChanges>.
- Conference on Language Modeling. COLM 2026: Call for Papers, 2026. URL <https://colmweb.org/cfp.html>.
- Mike D’Arcy, Tom Hope, Larry Birnbaum, and Doug Downey. Marg: Multi-agent review generation for scientific papers. *arXiv preprint arXiv:2401.04259*, 2024.
- Sumanth Dathathri, Abigail See, Sumedh Ghaisas, Po-Sen Huang, Rob McAdam, Johannes Welbl, Vandana Bachani, Alex Kaskasoli, Robert Stanforth, Tatiana Matejovicova, et al. Scalable watermarking for identifying large language model outputs. *Nature*, 634(8035):818–823, 2024.
- Caner Dilber and İsmail Yoşumaz. The impact of language translation on plagiarism rates: Evidence from turnitin, ithenticate, and grammarly. *Journal of Academic Ethics*, 24(1):4, 2026.
- Ruomeng Ding, Yifei Pang, He Sun, Yizhong Wang, Zhiwei Steven Wu, and Zhun Deng. Rubrics as an attack surface: Stealthy preference drift in llm judges. *arXiv preprint arXiv:2602.13576*, 2026a.
- Tianyu Ding, Aditya Nannapaneni, Bingfan Liu, and Ling Zhang. Autonomous research agents: A survey of ai scientists and the verification gap. *arXiv preprint arXiv:2608.05179*, 2026b.
- Anil R Doshi and Oliver P Hauser. Generative ai enhances individual creativity but reduces the collective diversity of novel content. *Science advances*, 10(28):eadn5290, 2024.
- Jiangshu Du, Yibo Wang, Wenting Zhao, Zhongfen Deng, Shuaiqi Liu, Renze Lou, Henry Peng Zou, Pranav Narayanan Venkit, Nan Zhang, Mukund Srinath, et al. Llms assist nlp researchers: Critique paper (meta-) reviewing. In *Proceedings of the 2024 conference on empirical methods in natural language processing*, pages 5081–5099, 2024.
- André V Duarte, Brian Tufts, Aditya Oke, Fei Fang, Arlindo L Oliveira, and Lei Li. Sem-detect: Semantic level detection of ai generated peer-reviews. *arXiv preprint arXiv:2605.21713*, 2026.
- Liam Dugan, Alyssa Hwang, Filip Trhlík, Andrew Zhu, Josh Magnus Ludan, Hainiu Xu, Daphne Ippolito, and Chris Callison-Burch. Raid: A shared benchmark for robust evaluation of machine-generated text detectors. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 12463–12492, 2024.
- Nils Dycke and Iryna Gurevych. Automatic reviewers fail to detect faulty reasoning in research papers: A new counterfactual evaluation framework. *Transactions of the Association for Computational Linguistics*, 14:465–488, 2026.
- Saman Ebadi, Hassan Nejadghanbar, Ahmed Rawdhan Salman, and Hassan Khosravi. Exploring the impact of generative ai on peer review: Insights from journal reviewers. *Journal of Academic Ethics*, 23(3):1383–1397, 2025.
- ECCV 2026. ECCV 2026 Reviewer FAQs, 2026. URL <https://eccv.ecva.net/Conferences/2026/ReviewerFAQs>.
- Steffen Eger, Yong Cao, Jennifer D’Souza, Andreas Geiger, Christian Greisinger, Stephanie Gross, Yufang Hou, Brigitte Krenn, Anne Lauscher, Yizhi Li, Chenchua Lin, Nafise Sadat Moosavi, Wei Zhao, and Tristan Miller. Transforming science with large language models: A survey on ai-assisted scientific discovery, experimentation, content generation, and evaluation. *arXiv preprint arXiv:2502.05151*, 2025.
- EMNLP 2026. EMNLP 2026 AI Reviewing Experiment, 2026a. URL <https://2026.emnlp.org/ai-reviewing-experiment/>.
- EMNLP 2026. EMNLP 2026 Paper Integrity Policy, 2026b. URL <https://2026.emnlp.org/paper-integrity-policy/>.

- Haishuo Fang, Yue Feng, and Iryna Gurevych. From passive generation to investigation: A proactive scientific peer review agent. *arXiv preprint arXiv:2606.13349*, 2026.
- Hao Fang, Jiawei Kong, Tianqu Zhuang, Yixiang Qiu, Kuofeng Gao, Bin Chen, Shu-Tao Xia, Yaowei Wang, and Min Zhang. Your language model can secretly write like humans: Contrastive paraphrase attacks on llm-generated text detectors. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pages 8596–8613, 2025.
- Michael R Fettiplace, Anuj Bhatia, Yian Chen, Steven L Orebaugh, Michael Gofeld, Rodney A Gabriel, Daniel I Sessler, Hannah Lonsdale, Brittani Bungart, Christopher P Cheng, et al. Recommendations for disclosure of artificial intelligence in scientific writing and publishing: a regional anesthesia and pain medicine modified delphi study. *Regional Anesthesia & Pain Medicine*, 51(6): 603–610, 2026.
- Alexander M Fichtl, Lukas Ellinger, Josefin Kelber, KryŁŁtof OlŁŁk, and Georg Groh. Ai-assisted peer review across research communities: From reviewer ai policies to llm review quality. *arXiv preprint arXiv:2608.03581*, 2026.
- Dragan Filimonovic, Christian Rutzer, and Conny Wunsch. Can genai improve academic performance? evidence from the social and behavioral sciences. *arXiv preprint arXiv:2510.02408*, 2025.
- Sasi Kiran Gaddipati, Diyana Muhammed, Farhana Keya, Gollam Rabby, and Sören Auer. Mlrepligate: Benchmarking autonomous research systems for machine learning reproducibility. *arXiv preprint arXiv:2605.16616*, 2026.
- Conner Ganjavi, Michael B Eppler, Asli Pekcan, Brett Biedermann, Andre Abreu, Gary S Collins, Inderbir S Gill, and Giovanni E Cacciamani. Publishers’ and journals’ instructions to authors on use of generative artificial intelligence in academic and scientific publishing: bibliometric analysis. *bmj*, 384, 2024.
- Xian Gao, Jiacheng Ruan, Zongyun Zhang, Jingsheng Gao, Ting Liu, and Yuzhuo Fu. Reviewagents: Bridging the gap between human and ai-generated paper reviews. *arXiv preprint arXiv:2503.08506*, 2025.
- Madhav Krishan Garg, Tejash Prasad, Tanmay Singhal, Chhavi Kirtani, Murari Mandal, and Dhruv Kumar. Revieweval: An evaluation framework for ai-generated reviews. In *EMNLP (Findings)*, pages 20542–20564, 2025.
- Claudine Gartenberg, Sharique Hasan, Alex Murray, and Lamar Pierce. More versus better: Artificial intelligence, incentives, and the emerging crisis in peer review. *Organization Science*, 37(3): 795–812, 2026.
- Ali Essam Ghareeb, Benjamin Chang, Ludovico Mitchener, Angela Yiu, Caralyn J Szostkiewicz, Dmytro Shved, Gavin J Gyimesi, Jon M Laurent, Samantha M Wright, Muhammed T Razzak, et al. A multi-agent system for automating scientific discovery. *Nature*, pages 1–3, 2026.
- Ali Ghorbanpour, Soroush Sadeghian, Alireza Daghighfarsoodeh, Sajad Ebrahimi, Negar Arabzadeh, Seyed Mohammad Hosseini, and Ebrahim Bagheri. Peerispect: Claim verification in scientific peer reviews. *arXiv preprint arXiv:2604.17667*, 2026.
- Alexander Goldberg, Ihsan Ullah, Thanh Gia Hieu Khuong, Benedictus Kent Rachmat, Zhen Xu, Isabelle Guyon, and Nihar B Shah. Usefulness of llms as an author checklist assistant for scientific papers: Neurips’24 experiment. *arXiv preprint arXiv:2411.03417*, 2024.
- Juraj Gottweis, Wei-Hung Weng, Alexander Daryin, Tao Tu, Anil Palepu, Petar Sirkovic, Artiom Myaskovsky, Felix Weissenberger, Keran Rong, Ryutaro Tanno, et al. Towards an ai co-scientist. *arXiv preprint arXiv:2502.18864*, 2, 2025.
- Palash Goyal, Mihir Parmar, Yiwen Song, Hamid Palangi, Tomas Pfister, and Jinsung Yoon. Scholarpeer: A context-aware multi-agent framework for automated peer review. *arXiv preprint arXiv:2601.22638*, 2026.

- Dirk Groeneveld, Iz Beltagy, Evan Walsh, Akshita Bhagia, Rodney Kinney, Oyvind Tafjord, Ananya Jha, Hamish Ivison, Ian Magnusson, Yizhong Wang, et al. Olmo: Accelerating the science of language models. In *Proceedings of the 62nd annual meeting of the association for computational linguistics (volume 1: long papers)*, pages 15789–15809, 2024.
- Yongtong Gu, Songze Li, and Xia Hu. Mash: Evading black-box ai-generated text detectors via style humanization. In *Findings of the Association for Computational Linguistics: ACL 2026*, pages 29749–29769, 2026.
- Yanzhu Guo, Guokan Shang, Michalis Vazirgiannis, and Chloé Clavel. The curious decline of linguistic diversity: Training language models on synthetic text. In *Findings of the Association for Computational Linguistics: NAACL 2024*, pages 3589–3604, 2024.
- Tarun Gupta and Danish Pruthi. All that glitters is not novel: Plagiarism in ai generated research. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 25721–25738, 2025.
- Hilda Hadan, Derrick M Wang, Reza Hadi Mogavi, Joseph Tu, Leah Zhang-Kennedy, and Lennart E Nacke. The great ai witch hunt: Reviewers’ perception and (mis) conception of generative ai in research writing. *Computers in Human Behavior: Artificial Humans*, 2(2):100095, 2024.
- Jutta Haider, Kristofer Rolf Söderström, Björn Ekström, and Malte Rödl. Gpt-fabricated scientific papers on google scholar: Key features, spread, and implications for preempting evidence manipulation. *Harvard Kennedy School Misinformation Review*, 5(5), 2024.
- Peixuan Han, Yingjie Yu, Jingjun Xu, and Jiaxuan You. Drpg (decompose, retrieve, plan, generate): An agentic framework for academic rebuttal. In *Findings of the Association for Computational Linguistics: ACL 2026*, pages 41232–41249, 2026.
- Qian Yue Hao, Fengli Xu, Yong Li, and James Evans. Artificial intelligence tools expand scientists’ impact but contract science’s focus. *Nature*, 649(8099):1237–1243, 2026.
- Soheil Hassanipour, Sandeep Nayak, Ali Bozorgi, Mohammad-Hossein Keivanlou, Tirth Dave, Abdulhadi Alotaibi, Farahnaz Joukar, Parinaz Mellatdoust, Arash Bakhshi, Dona Kuriyakose, et al. The ability of chatgpt in paraphrasing texts and reducing plagiarism: a descriptive analysis. *JMIR medical education*, 10(1):e53308, 2024.
- Hans Ole Hatzel, Sebastian Steindl, and Jan Strich. Review arcade: On the human alignment and gameability of llm reviews. *arXiv preprint arXiv:2605.28897*, 2026.
- Yongyuan He and Yi Bu. Academic journals’ ai policies fail to curb the surge in ai-assisted academic writing. *Proceedings of the National Academy of Sciences*, 123(9):e2526734123, 2026a.
- Yongyuan He and Yi Bu. Reply to wang: Policy failure can be assessed from observed outcomes. *Proceedings of the National Academy of Sciences*, 123(27):e2617762123, 2026b.
- Eftekhari Hossain, Sanjeev Kumar Sinha, Naman Bansal, R Alexander Knipper, Souvika Sarkar, John Salvador, Yash Mahajan, Sri Ram Pavan Kumar Guttikonda, Mousumi Akter, Md Mahadi Hassan, et al. Llms as meta-reviewers’ assistants: A case study. In *Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 7763–7803, 2025.
- You Hu, Chenzhuo Zhao, Changfa Mo, Haotian Liu, and Xiaobai Li. Scifigdetect: A benchmark for ai-generated scientific figure detection. *arXiv preprint arXiv:2604.08211*, 2026.
- Shuo Huang, Yan Pen, and Lizhen Qu. Tab-audit: Detecting ai-fabricated scientific tables via multi-view likelihood mismatch. *arXiv preprint arXiv:2603.19712*, 2026.
- ICLR 2026 Program Chairs. Policies on Large Language Model Usage at ICLR 2026, 2025a. URL <https://blog.iclr.cc/2025/08/26/policies-on-large-language-model-usage-at-iclr-2026/>.
- ICLR 2026 Program Chairs. ICLR 2026 Response to LLM-Generated Papers and Reviews, 2025b. URL <https://blog.iclr.cc/2025/11/19/iclr-2026-response-to-llm-generated-papers-and-reviews/>.

- ICLR 2026 Program Chairs. A Retrospective on the ICLR 2026 Review Process, 2026. URL <https://blog.iclr.cc/2026/03/31/a-retrospective-on-the-iclr-2026-review-process/>.
- Maximilian Idahl and Zahra Ahmadi. Openreviewer: A specialized large language model for generating critical scientific paper reviews. In *Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (System Demonstrations)*, pages 550–562, 2025.
- Tal Ifargan, Lukas Hafner, Maor Kern, Ori Alcalay, and Roy Kishony. Autonomous llm-driven research—from data to human-verifiable research papers. *NEJM AI*, 2(1):AIoa2400555, 2025.
- Peter Jansen, Oyvind Tafjord, Marissa Radensky, Pao Siangliulue, Tom Hope, Bhavana Dalvi, Bodhisattwa Prasad Majumder, Daniel S Weld, and Peter Clark. Codescientist: End-to-end semi-automated scientific discovery with code-based experimentation. In *Findings of the Association for Computational Linguistics: ACL 2025*, pages 13370–13467, 2025.
- Rajesh Jayaram, Drew Tyler, David Woodruff, Corinna Cortes, Yossi Matias, Vahab Mirrokni, and Vincent Cohen-Addad. Towards automating scientific review with google’s paper assistant tool. *arXiv preprint arXiv:2606.28277*, 2026.
- Fengqing Jiang, Yichen Feng, Yuetai Li, Luyao Niu, Basel Alomair, and Radha Poovendran. Bad-scientist: Can a research agent write convincing but unsound papers that fool llm reviewers? In *Proceedings of the 64th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 24712–24727, 2026a.
- Liwei Jiang, Yuanjun Chai, Margaret Li, Mickel Liu, Raymond Fok, Nouha Dziri, Yulia Tsvetkov, Maarten Sap, and Yejin Choi. Artificial hivemind: The open-ended homogeneity of language models (and beyond). *Advances in Neural Information Processing Systems*, 38, 2026b.
- Ramsha Kamran, Maheera Amjad, Zartasha Mustansar, Arsalan Shaukat, Salma Sherbaz, and Muhammad US Khan. Prompt-to-paper: Agentic ai system for bioinformatics. *arXiv preprint arXiv:2607.05456*, 2026.
- Nikhil Kandpal, Brian Lester, Colin Raffel, Sebastian Majstorovic, Stella Biderman, Baber Abbasi, Luca Soldaini, Enrico Shippole, A Feder Cooper, Aviya Skowron, et al. The common pile v0.1: An 8tb dataset of public domain and openly licensed text. *Advances in Neural Information Processing Systems*, 38, 2026.
- Masahiro Kaneko. Paraphrasing adversarial attack on llm-as-a-reviewer. *arXiv preprint arXiv:2601.06884*, 2026.
- Dongyeop Kang, Waleed Ammar, Bhavana Dalvi, Madeleine Van Zuylen, Sebastian Kohlmeier, Eduard Hovy, and Roy Schwartz. A dataset of peer reviews (peerread): Collection, insights and nlp applications. In *Proceedings of the 2018 conference of the North American chapter of the Association for Computational Linguistics: Human language technologies, volume 1 (long papers)*, pages 1647–1661, 2018.
- KDD 2025. Research Track: Call for Papers, 2025. URL <https://kdd.org/kdd2025/research-track-call-for-papers/>.
- Elliot Kim, Avi Garg, Kenny Peng, and Nikhil Garg. Correlated errors in large language models. *arXiv preprint arXiv:2506.07962*, 2025.
- Seungone Kim, Dongkeun Yoon, Kiril Gashteovski, Juyoung Suk, Jinheon Baek, Pranjal Aggarwal, Ian Wu, Viktor Zaverkin, Spase Petkoski, Daniel R Schrider, et al. On the limits and opportunities of ai reviewers: Reviewing the reviews of nature-family papers with 45 expert scientists. *arXiv preprint arXiv:2605.20668*, 2026.
- Peter Kirgis, Sayash Kapoor, Andrew Schwartz, Stephan Rabanser, David Africa, Konstantinos Voudouris, Viet Nguyen, Toby Pilditch, Magda Dubois, Harry Coppock, et al. Can ai agents conduct open-ended ai research? early evidence from two case studies. *arXiv preprint arXiv:2607.27191*, 2026.

- Samsun Knight. Flooded gateways. *Available at SSRN 6297213*, 2026.
- Dmitry Kobak, Rita González-Márquez, Emőke-Ágnes Horvát, and Jan Lause. Delving into llm-assisted writing in biomedical publications through excess vocabulary. *Science Advances*, 11(27): eadt3813, 2025.
- Lingdong Kong, Xian Sun, Wei Chow, Linfeng Li, Kevin Qinghong Lin, Xuan Billy Zhang, Song Wang, Rong Li, Qing Wu, Wei Gao, et al. Ai for auto-research: Roadmap & user guide. *arXiv preprint arXiv:2605.18661*, 2026.
- Kalpesh Krishna, Yixiao Song, Marzena Karpinska, John Wieting, and Mohit Iyyer. Paraphrasing evades detectors of ai-generated text, but retrieval is an effective defense. *Advances in neural information processing systems*, 36:27469–27500, 2023.
- Sandeep Kumar, Mohit Sahu, Vardhan Gacche, Tirthankar Ghosal, and Asif Ekbal. ‘quis custodiet ipsos custodes?’ who will watch the watchmen? on detecting ai-generated peer-reviews. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 22663–22679, 2024.
- Sandeep Kumar, Samarth Garg, Sagnik Sengupta, Tirthankar Ghosal, and Asif Ekbal. Mixrevdetect: Towards detecting ai-generated content in hybrid peer reviews. In *Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 2: Short Papers)*, pages 944–953, 2025.
- Tharindu Kumarage, Paras Sheth, Raha Moraffah, Joshua Garland, and Huan Liu. How reliable are ai-generated-text detectors? an assessment framework using evasive soft prompts. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 1337–1349, 2023.
- Keigo Kusumegi, Xinyu Yang, Paul Ginsparg, Mathijs De Vaan, Toby Stuart, and Yian Yin. Scientific production in the era of large language models. *Science*, 390(6779):1240–1243, 2025.
- Iliia Kuznetsov, Osama Mohammed Afzal, Koen Dercksen, Nils Dycke, Alexander Goldberg, Tom Hope, Dirk Hovy, Jonathan K Kummerfeld, Anne Lauscher, Kevin Leyton-Brown, et al. What can natural language processing do for peer review? *arXiv preprint arXiv:2405.06563*, 2024.
- Anran Li, Lingfei Qian, Mengmeng Du, Yu Yin, Yan Hu, Zihao Sun, Yihang Fu, Hyunjae Kim, Erica Stutz, Xuguang Ai, et al. Memorization in large language models in medicine prevalence characteristics and implications. *Nature Communications*, 2026a.
- Baoxue Li and Chunhui Zhao. Self-reflection enhances large language models towards substantial academic response. *npj Artificial Intelligence*, 1(1):42, 2025.
- Haowen Li, Yoichi Ishibashi, and Masafumi Oyamada. Evaluating the impact of reviewer guideline design on llm-based automated peer review. In *Findings of the Association for Computational Linguistics: ACL 2026*, pages 30223–30240, 2026b.
- Jiatao Li, Yanheng Li, Xinyu Hu, Mingqi Gao, and Xiaojun Wan. Where do llms go wrong? diagnosing automated peer review via aspect-guided multi-level perturbation. In *Proceedings of the 34th ACM International Conference on Information and Knowledge Management*, pages 1572–1581, 2025a.
- Lin Li, Qi Zhang, Xander Davies, Jianing Qiu, and Yarin Gal. Gaming ai-assisted peer reviews poses new risks to the scientific community. *arXiv preprint arXiv:2606.10159*, 2026c.
- Lingyao Li, Junjie Xiong, Changjia Zhu, Runlong Yu, Chen Chen, Junyu Wang, Renkai Ma, and Zhicong Lu. Llm-as-a-reviewer: Benchmarking their ability, divergence, and prompt injection resistance as paper reviewers. *arXiv preprint arXiv:2605.25415*, 2026d.
- Miao Li, Eduard Hovy, and Jey Lau. Summarizing multiple documents with conversational structure for meta-review generation. *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 7089–7112, 2023.
- Miao Li, Jey Han Lau, and Eduard Hovy. A sentiment consolidation framework for meta-review generation. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 10158–10177, 2024a.

- Rui Li, Jia-Chen Gu, Po-Nien Kung, Heming Xia, Xiangwen Kong, Zhifang Sui, Nanyun Peng, et al. Llm-reval: Can we trust llm reviewers yet? *arXiv preprint arXiv:2510.12367*, 2025b.
- Zhi-Qiang Li, Hui-Lin Xu, Hui-Juan Cao, Zhao-Lan Liu, Yu-Tong Fei, and Jian-Ping Liu. Use of artificial intelligence in peer review among top 100 medical journals. *JAMA Network Open*, 7(12): e2448609, 2024b.
- Zhuofeng Li, Yi Lu, Dongfu Jiang, Haoxiang Zhang, Yuyang Bai, Chuan Li, Yu Wang, Shuiwang Ji, Jianwen Xie, and Yu Zhang. Reviewgrounder: Improving review substantiveness with rubric-guided, tool-integrated agents. In *Proceedings of the 64th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 32015–32044, 2026e.
- Weixin Liang, Mert Yuksekgonul, Yining Mao, Eric Wu, and James Zou. Gpt detectors are biased against non-native english writers. *Patterns*, 4(7), 2023.
- Weixin Liang, Zachary Izzo, Yaohui Zhang, Haley Lepp, Hancheng Cao, Xuandong Zhao, Lingjiao Chen, Haotian Ye, Sheng Liu, Zhi Huang, et al. Monitoring ai-modified content at scale: A case study on the impact of chatgpt on ai conference peer reviews. *arXiv preprint arXiv:2403.07183*, 2024a.
- Weixin Liang, Yuhui Zhang, Hancheng Cao, Binglu Wang, Daisy Yi Ding, Xinyu Yang, Kailas Vodrahalli, Siyu He, Daniel Scott Smith, Yian Yin, et al. Can large language models provide useful feedback on research papers? a large-scale empirical analysis. *NEJM AI*, 1(8):AIoa2400196, 2024b.
- Weixin Liang, Yaohui Zhang, Zhengxuan Wu, Haley Lepp, Wenlong Ji, Xuandong Zhao, Hancheng Cao, Sheng Liu, Siyu He, Zhi Huang, et al. Quantifying large language model usage in scientific papers. *Nature Human Behaviour*, 9(12):2599–2609, 2025.
- Tzu-Ling Lin, Wei-Chih Chen, Teng-Fang Hsiao, Hou-I Liu, Ya-Hsin Yeh, Yu Kai Chan, Wen-Sheng Lien, Po-Yen Kuo, S Yu Philip, and Hong-Han Shuai. Breaking the reviewer: Assessing the vulnerability of large language models in automated peer review under textual adversarial attacks. In *EMNLP (Findings)*, pages 4819–4839, 2025.
- Zhicheng Lin. Hidden prompts in manuscripts exploit ai-assisted peer review. *Communications of the ACM*, 69(7):53–56, 2026.
- Jae QJ Liu, Kelvin TK Hui, Fadi Al Zoubi, Zing ZX Zhou, Dino Samartzis, Curtis CH Yu, Jeremy R Chang, and Arnold YL Wong. The great detectives: humans versus ai detectors in catching large language model-generated medical writing. *International Journal for Educational Integrity*, 20(1): 8, 2024.
- Jiaqi Liu, Shi Qiu, Mairui Li, Bingzhou Li, Haonian Ji, Siwei Han, Xinyu Ye, Peng Xia, Zihan Dong, Meng Chen, et al. Autoresearchclaw: Self-reinforcing autonomous research with human-ai collaboration. *arXiv preprint arXiv:2605.20025*, 2026.
- Ngoc Phan Phuoc Loc, La Viet, Toan Huynh, Thanh Tran Khanh, Duy A Nguyen, Tuan Anh Nguyen Pham, Thanh Nguyen, Nitesh V Chawla, Wray Buntine, Kok-Seng Wong, et al. Prism: A multi-dimensional benchmark for evaluating llm peer reviewers. *arXiv preprint arXiv:2605.26730*, 2026.
- Chris Lu, Cong Lu, Robert Tjarko Lange, Jakob Foerster, Jeff Clune, and David Ha. The ai scientist: Towards fully automated open-ended scientific discovery. *arXiv preprint arXiv:2408.06292*, 2024.
- Zihan Luo, Hong Huang, Jianxun Lian, Yu Chang, Xing Xie, and Hai Jin. Can ai revise research papers with human review feedback? an empirical study and benchmark. In *Findings of the Association for Computational Linguistics: ACL 2026*, pages 17876–17893, 2026.
- Ziming Luo, Atoosa Kasirzadeh, and Nihar B Shah. The more you automate, the less you see: Hidden pitfalls of ai scientist systems. *arXiv preprint arXiv:2509.08713*, 2025.
- Youngang Lyu, Xi Zhang, Xinhao Yi, Yuyue Zhao, Shuyu Guo, Wenxiang Hu, Jan Piotrowski, Jakub Kaliski, Jacopo Urbani, Zaiqiao Meng, et al. Evoscientist: Towards multi-agent evolving ai scientists for end-to-end scientific discovery. *arXiv preprint arXiv:2603.08127*, 2026.

- Oubo Ma, Ruixiao Lin, Jiahao Chen, Yuan Su, Yong Yang, and Shouling Ji. Shattering the echo chamber: Hidden safeguards in manuscripts against the ai takeover of peer review. *arXiv preprint arXiv:2605.05271*, 2026a.
- Qianli Ma, Chang Guo, Zhiheng Tian, Siyu Wang, Jipeng Xiao, Yuanhao Yue, and Zhipeng Zhang. Paper2rebuttal: A multi-agent framework for transparent author response assistance. In *Proceedings of the 64th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 46134–46166, 2026b.
- Sebastian Porsdam Mann, Mateo Aboy, Joel Jiehaio Seah, Zhicheng Lin, Xufei Luo, Daniel Rodger, Hazem Zohny, Timo Minssen, Julian Savulescu, and Brian D Earp. Ai and the future of academic peer review. *arXiv preprint arXiv:2509.14189*, 2025.
- Shigeki Matsubara. Large language model content detectors: Effects of human editing and chatgpt mimicking individual style—a preliminary study. *Journal of Investigative Medicine*, 73(8):666–672, 2025.
- Danny Maupin, Tulsi Suchak, Adrian Barnett, and Matt Spick. Dramatic increases in redundant publications in the generative ai era. *BMC medicine*, 24(1):29, 2026.
- Zoie McCreery and MZ Naser. Llm-assisted writing in engineering is associated with higher citations as evidenced from 1.17 million papers. *Journal of Informetrics*, 20(3):101842, 2026.
- Wenlong Meng, Shuguo Fan, Chengkun Wei, Min Chen, Yuwei Li, Yuanchao Zhang, Zhikun Zhang, and Wenzhi Chen. {GradEscape}: A {Gradient-Based} evader against {AI-Generated} text detectors. In *34th USENIX Security Symposium (USENIX Security 25)*, pages 181–200, 2025.
- Lisa Messeri and Molly J Crockett. Artificial intelligence and illusions of understanding in scientific research. *Nature*, 627(8002):49–58, 2024.
- Atsuyuki Miyai, Mashiro Toyooka, Takashi Otonari, Zaiying Zhao, and Kiyoharu Aizawa. Jr. ai scientist and its risk report: Autonomous scientific exploration from a baseline paper. *arXiv preprint arXiv:2511.04583*, 2025.
- Jimin Mun, Chani Jung, Xuhui Zhou, Hyunwoo Kim, and Maarten Sap. Goodpoint: Learning constructive scientific paper feedback from author responses. *arXiv preprint arXiv:2604.11924*, 2026.
- Michaela G Murdock, Aditya Tadinada, and Michaela Murdock. Can ai tools reliably and effectively detect plagiarism in scientific writing? *Cureus*, 17(5), 2025.
- Alexander Nemecek, Yuzhou Jiang, and Erman Ayday. The feasibility of topic-based watermarking on academic peer reviews. In *Proceedings of the 14th International Joint Conference on Natural Language Processing and the 4th Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics*, pages 616–634, 2025.
- NeurIPS 2026 Communication Chairs. AI-Generated Papers in the NeurIPS 2026 Position Paper Track, 2026. URL <https://blog.neurips.cc/2026/06/02/ai-generated-papers-in-the-neurips-2026-position-paper-track/>.
- Thi Huyen Nguyen and Zahra Ahmadi. Llm-based scientific peer review: Methods, benchmarks, and reliability challenges. *arXiv preprint arXiv:2606.25057*, 2026.
- Guillaume-Anthony Odri and Diane Ji Yun Yoon. Detecting generative artificial intelligence in scientific articles: Evasion techniques and implications for scientific integrity. *Orthopaedics & Traumatology: Surgery & Research*, 109(8):103706, 2023.
- Tazro Ohta, Nomi L Harris, and Seth Carbon. Ai-assisted pre-review of open-source software submissions: an experience report from bosc 2026. *arXiv preprint arXiv:2607.27228*, 2026.
- Jiefu Ou, William Gantt Walden, Kate Sanders, Zhengping Jiang, Kaiser Sun, Jeffrey Cheng, William Jurayj, Miriam Wanner, Shaobo Liang, Candice Morgan, et al. Claimcheck: How grounded are llm critiques of scientific papers? In *EMNLP (Findings)*, pages 21712–21735, 2025.

- Andrea Pedrotti, Michele Papucci, Cristiano Ciaccio, Alessio Miaschi, Giovanni Puccetti, Felice Dell’Orletta, and Andrea Esuli. Stress-testing machine generated text detection: Shifting language models writing style to fool detectors. In *Findings of the Association for Computational Linguistics: ACL 2025*, pages 3010–3031, 2025.
- Simon J Porter and Leslie D McIntosh. Identifying fabricated networks within authorship-for-sale enterprises. *Scientific Reports*, 14(1):29569, 2024.
- Ahmad R Pratama. The accuracy-bias trade-offs in ai text detection tools and their impact on fairness in scholarly publication. *PeerJ. Computer science*, 11:e2953–e2953, 2025.
- Minfeng Qi, Zhongmin Cao, Qin Wang, Ningran Li, and Tianqing Zhu. Does genai rewrite how we write? an empirical study on two-million preprints. *arXiv preprint arXiv:2510.17882*, 2025.
- Xinpeng Qiu, Wang Yihu, Zhifeng Liu, Xiaochen Wang, and Jimin Wang. Egtr-review: Efficient evidence-grounded scientific peer review generation via multi-agent teacher distillation. *arXiv preprint arXiv:2606.06025*, 2026.
- Vyas Raina, Adian Liusie, and Mark Gales. Is llm-as-a-judge robust? investigating universal adversarial attacks on zero-shot llm assessment. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 7499–7517, 2024.
- Vishisht Srihari Rao, Aounon Kumar, Himabindu Lakkaraju, and Nihar B Shah. Detecting llm-generated peer reviews. *PLoS One*, 20(9):e0331871, 2025.
- Abdur Rasool, Xiaohui Huang, Yanqing Hu, and Linyi Yang. Reviewguard: Aligning llm-assisted peer review with long-term scientific impact. *arXiv preprint arXiv:2606.24892*, 2026.
- Thomas Renault, Antonin Bergeaud, and Clément Bosquet. Comment on scientific production in the era of large language models. *arXiv preprint arXiv:2605.17979*, 2026.
- Qian Ruan and Iryna Gurevych. Author-in-the-loop response generation and evaluation: Integrating author expertise and intent in responses to peer review. In *Proceedings of the 64th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 12390–12418, 2026.
- Giuseppe Russo Latona, Manoel Horta Ribeiro, Tim Ruben Davidson, Veniamin Veselovsky, and Robert West. The ai review lottery: Widespread ai-assisted peer reviews boost paper scores and acceptance rates. *Proceedings of the ACM on Human-Computer Interaction*, 9(7):1–28, 2025.
- Vinu Sankar Sadasivan, Aounon Kumar, Sriram Balasubramanian, Wenxiao Wang, Soheil Feizi, et al. Can ai-generated text be reliably detected? stress testing ai text detectors under various attacks. *Transactions on Machine Learning Research*, 1, 2025.
- Rounak Saha, Gurusha Juneja, Dayita Chaudhuri, Naveeja Sajeevan, Nihar B Shah, and Danish Pruthi. Policies permitting llm use for polishing peer reviews are currently not enforceable. *arXiv preprint arXiv:2603.20450*, 2026.
- Devanshu Sahoo, Manish Prasad, Vasudev Majhi, Jahnvi Singh, Vinay Chamola, Yash Sinha, Murari Mandal, and Dhruv Kumar. When reject turns into accept: Quantifying the vulnerability of llm-based scientific reviewers to indirect prompt injection. *arXiv preprint arXiv:2512.10449*, 2025.
- Aishik Sanyal, Samuel Schapiro, Sumuk Shashidhar, Royce Moon, Lav R Varshney, and Dilek Hakkani-Tur. Spark: A system for scientifically creative idea generation. *arXiv preprint arXiv:2504.20090*, 2025.
- Samuel Schmidgall and Michael Moor. Agentrxiv: Towards collaborative autonomous research. *arXiv preprint arXiv:2503.18102*, 2025.
- Samuel Schmidgall, Yusheng Su, Ze Wang, Ximeng Sun, Jialian Wu, Xiaodong Yu, Jiang Liu, Michael Moor, Zicheng Liu, and Emad Barsoum. Agent laboratory: Using llm agents as research assistants. *Findings of the Association for Computational Linguistics: EMNLP 2025*, pages 5977–6043, 2025.

- Paul G Schmidt and Amnon J Meir. Using generative ai for literature searches and scholarly writing: is the integrity of the scientific discourse in jeopardy. *Notices of the American Mathematical Society*, 71(1):93–104, 2024.
- Mohamed El Amine Seddik, Suei-Wen Chen, Soufiane Hayou, Pierre Youssef, and Merouane Debbah. How bad is training on synthetic data? a statistical analysis of language model collapse. *arXiv preprint arXiv:2404.05090*, 2024.
- Sadat Shahriar, Navid Ayoobi, and Arjun Mukherjee. The erosion of llm signatures: Can we still distinguish human and llm-generated scientific ideas after iterative paraphrasing? In *Proceedings of the 15th International Conference on Recent Advances in Natural Language Processing-Natural Language Processing in the Generative AI Era*, pages 1118–1126, 2025.
- Vibhhu Sharma, Thorsten Joachims, and Sarah Dean. Do llms favor llms? quantifying interaction effects in peer review. In *The 2026 ACM Conference on Fairness, Accountability, and Transparency*, pages 3151–3189, 2026.
- Chenhui Shen, Liying Cheng, Ran Zhou, Lidong Bing, Yang You, and Luo Si. Mred: A meta-review dataset for structure-controllable text generation. *Findings of the Association for Computational Linguistics: ACL 2022*, pages 2521–2535, 2022.
- Jiawen Shi, Zenghui Yuan, Yinuo Liu, Yue Huang, Pan Zhou, Lichao Sun, and Neil Zhenqiang Gong. Optimization-based prompt injection attack to llm-as-a-judge. In *Proceedings of the 2024 on ACM SIGSAC Conference on Computer and Communications Security*, pages 660–674, 2024.
- Ilia Shumailov, Zakhar Shumaylov, Yiren Zhao, Nicolas Papernot, Ross Anderson, and Yarin Gal. Ai models collapse when trained on recursively generated data. *Nature*, 631(8022):755–759, 2024.
- Kyle Siler. The diffusion of large language models in published academic articles. *Proceedings of the National Academy of Sciences*, 123(22):e2605754123, 2026.
- Luca Soldaini, Rodney Kinney, Akshita Bhagia, Dustin Schwenk, David Atkinson, Russell Authur, Ben Bogin, Khyathi Chandu, Jennifer Dumas, Yanai Elazar, et al. Dolma: An open corpus of three trillion tokens for language model pretraining research. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 15725–15788, 2024.
- Guijin Son, Jiwoo Hong, Honglu Fan, Heejeong Nam, Hyunwoo Ko, Seungwon Lim, Jinyeop Song, Jinha Choi, Gonçalo Paulo, Youngjae Yu, et al. When ai co-scientists fail: Spot-a benchmark for automated verification of scientific research. *arXiv preprint arXiv:2505.11855*, 2025.
- Yiwen Song, Yale Song, Tomas Pfister, and Jinsung Yoon. Paperorchestra: A multi-agent framework for automated ai research paper writing. *arXiv preprint arXiv:2604.05018*, 2026.
- Diomidis Spinellis. False authorship: an explorative case study around an ai-generated article published under my name. *Research Integrity and Peer Review*, 10(1):8, 2025.
- Ivan Stelmakh, Charvi Rastogi, Ryan Liu, Shuchi Chawla, Federico Echenique, and Nihar B Shah. Cite-seeing and reviewing: A study on citation bias in peer review. *Plos one*, 18(7):e0283980, 2023.
- Artur Strzelecki. ‘as of my last knowledge update’: How is content generated by chatgpt infiltrating scientific papers published in premier journals? *Learned Publishing*, 38(1):e1650, 2025.
- Andrea Taloni, Vincenzo Scordia, and Giuseppe Giannaccare. Modern threats in academia: evaluating plagiarism and artificial intelligence detection scores of chatgpt. *Eye*, 38(2):397–400, 2024.
- Jiabin Tang, Lianghao Xia, Zhonghang Li, and Chao Huang. Ai-researcher: Autonomous scientific innovation. *Advances in Neural Information Processing Systems*, 38:9481–9520, 2026a.
- Qiong Tang, Xiangkun Hu, Xiangyang Liu, Yiran Chen, and Yunfan Shao. Fars: A fully automated research system deployed at scale. *arXiv preprint arXiv:2606.31651*, 2026b.

- Nitya Thakkar, Mert Yuksekgonul, Jake Silberg, Animesh Garg, Nanyun Peng, Fei Sha, Rose Yu, Carl Vondrick, and James Zou. A large-scale randomized study of large language model feedback in peer review. *Nature Machine Intelligence*, 8(3):326–336, 2026.
- Panagiotis Theocharopoulos, Ajinkya Kulkarni, and Mathew Magimai Doss. Multilingual hidden prompt injection attacks on llm-based academic reviewing. *arXiv preprint arXiv:2512.23684*, 2025.
- Jingqi Tong, Mingzhe Li, Hangcheng Li, Yongzhuo Yang, Yurong Mou, Weijie Ma, Zhiheng Xi, Hongji Chen, Xiaoran Liu, Qinyuan Cheng, et al. Ai can learn scientific taste. *arXiv preprint arXiv:2603.14473*, 2026.
- Maxim Topaz, Nir Roguin, Pallavi Gupta, Zhihong Zhang, and Laura-Maria Peltonen. Fabricated citations: an audit across 2· 5 million biomedical papers. *The Lancet*, 407(10541):1779–1781, 2026.
- Federico Torrielli, Stefano Locci, Amon Rapp, and Luigi Di Caro. Exploiting large language models in peer review: indirect prompt injection attacks and integrity probes. *Scientometrics*, pages 1–63, 2026.
- Sai Suresh Macharla Vasu, Ivaxi Sheth, Hui-Po Wang, Ruta Binkyte, and Mario Fritz. Justice in judgment: Unveiling (hidden) bias in llm-assisted peer reviews. In *Findings of the Association for Computational Linguistics: ACL 2026*, pages 307–330, 2026.
- Kiri L Wagstaff, Joydeep Biswas, Erich Merrill III, Bo An, Ida Camacho, David J Crandall, and Matthew E Taylor. Aaai-26 dual submissions: Novel challenges. *arXiv preprint arXiv:2607.11918*, 2026.
- William H Walters and Esther Isabelle Wilder. Fabrication and errors in the bibliographic citations generated by chatgpt. *Scientific reports*, 13(1):14045, 2023.
- Binglu Wang, Weixin Liang, Jiahui Xue, Yuhui Zhang, Hancheng Cao, Dashun Wang, and Yian Yin. Human–ai collaboration in science at scale: A large-scale randomized field experiment. In *Academy of Management Proceedings*, volume 2026, page 21478. Academy of Management Valhalla, NY 10595, 2026a.
- James Liyuan Wang, Ran Li, Junfeng Yang, and Chengzhi Mao. Raft: Realistic attacks to fool text detectors. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 16923–16936, 2024a.
- Jialiang Wang, Yuchen Liu, Hang Xu, Kaichun Hu, Shimin Di, Wangze Ni, Linan Yue, Min-Ling Zhang, Kui Ren, and Lei Chen. When ai reviews science: Can we trust the referee? *arXiv preprint arXiv:2604.23593*, 2026b.
- Mingyi Wang, Zhuoer Shen, Yuheng Bu, and Shaofeng Zou. Detector-evasive llm paraphrasing via constrained policy optimization. *arXiv preprint arXiv:2606.00392*, 2026c.
- Peiyi Wang, Lei Li, Liang Chen, Zefan Cai, Dawei Zhu, Binghuai Lin, Yunbo Cao, Lingpeng Kong, Qi Liu, Tianyu Liu, et al. Large language models are not fair evaluators. In *Proceedings of the 62nd annual meeting of the association for computational linguistics (volume 1: Long papers)*, pages 9440–9450, 2024b.
- Tianchun Wang, Yuanzhou Chen, Zichuan Liu, Zhanwen Chen, Haifeng Chen, Xiang Zhang, and Wei Cheng. Humanizing the machine: Proxy attacks to mislead llm detectors. In *International Conference on Learning Representations*, volume 2025, pages 68339–68367, 2025.
- Yibo Wang. Policy snapshots cannot establish journal ai-policy failure. *Proceedings of the National Academy of Sciences*, 123(27):e2616276123, 2026.
- Qiyao Wei, Samuel Holt, Jing Yang, Markus Wulfmeier, and Mihaela van der Schaar. The ai imperative: Scaling high-quality peer review in machine learning. *arXiv e-prints*, pages arXiv–2506, 2025.

- Yixuan Weng, Minjun Zhu, Guangsheng Bao, Hongbo Zhang, Jindong Wang, Yue Zhang, and Linyi Yang. Cyclereviewer: Improving automated research via automated review. In *International Conference on Learning Representations*, volume 2025, pages 3669–3709, 2025.
- Yixuan Weng, Minjun Zhu, Qiujie Xie, Zhiyuan Ning, Shichen Li, Panzhong Lu, Zhen Lin, Enhao Gu, Qiyao Sun, and Yue Zhang. Deepreviewer 2.0: A traceable agentic system for auditable scientific peer review. *arXiv preprint arXiv:2604.09590*, 2026.
- Di Wu. Can ai review improve paper drafting? an empirical study on 20 computer architecture submissions. *arXiv preprint arXiv:2606.01013*, 2026.
- Fan Wu, Cheng Chen, Zhenshan Tan, Taiyu Zhang, Xinzheng Xu, Yanyu Qian, Dingcheng Gao, Lanyun Zhu, Qi Zhu, Yi Tan, et al. Claw ai lab: An autonomous multi-agent research team. *arXiv preprint arXiv:2605.22662*, 2026a.
- Hongtao Wu, Boyun Zheng, Dingjie Song, Yu Jiang, Jianfeng Gao, Lei Xing, Lichao Sun, and Yixuan Yuan. Towards a medical ai scientist. *arXiv preprint arXiv:2603.28589*, 2026b.
- Sihong Wu, Owen Jiang, Yilun Zhao, Tiansheng Hu, Yiling Ma, Kaiyan Zhang, Manasi Patwardhan, and Arman Cohan. Can ai be a good peer reviewer? a survey of peer review process, evaluation, and the future. In *Proceedings of the 64th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 32593–32619, 2026c.
- Sihong Wu, Yiling Ma, Yilun Zhao, Tiansheng Hu, Owen Jiang, Manasi Patwardhan, and Arman Cohan. Rbtact: Rebuttal as supervision for actionable review feedback generation. In *Findings of the Association for Computational Linguistics: ACL 2026*, pages 33965–33992, 2026d.
- Yuan Xin, Yixuan Weng, Minjun Zhu, Ying Ling, Chengwei Qin, Michael Backes, Yue Zhang, and Linyi Yang. Safereview: Defending llm-based review systems against adversarial hidden prompts. *arXiv preprint arXiv:2604.26506*, 2026.
- Junjie Xiong, Changjia Zhu, Shuhang Lin, Chong Zhang, Yongfeng Zhang, Yao Liu, and Lingyao Li. Invisible prompts, visible threats: Malicious font injection in external resources for large language models. *arXiv preprint arXiv:2505.16957*, 2025.
- Yutaro Yamada, Robert Tjarko Lange, Cong Lu, Shengran Hu, Chris Lu, Jakob Foerster, Jeff Clune, and David Ha. The ai scientist-v2: Workshop-level automated scientific discovery via agentic tree search. *arXiv preprint arXiv:2504.08066*, 2025.
- Jing Yang, Qiyao Wei, and Jiaxin Pei. Paper copilot: Tracking the evolution of peer review in ai conferences. In *International Conference on Learning Representations*, volume 2026, pages 68190–68201, 2026a.
- Junwei Yang, Hanwen Xu, Srбуhi Mirzoyan, Tong Chen, Zixuan Liu, Zequn Liu, Wei Ju, Luchen Liu, Zhiping Xiao, Ming Zhang, et al. Poisoning medical knowledge using large language models. *Nature Machine Intelligence*, 6(10):1156–1168, 2024.
- Xianglin Yang, Bryan Hooi, Gelei Deng, Tianwei Zhang, and Jin Song Dong. Turning bias into bugs: Bandit-guided style manipulation attacks on llm judges. *arXiv preprint arXiv:2605.26156*, 2026b.
- Xu Yang, Zhizhou Sha, Junbo Li, Jian Yu, Yifan Sun, Matthew Zhao, Jinrui Fang, Xinyue Guo, Yining Wu, Xu Hu, et al. No hidden prompts needed! you can game ai peer review with presentation-only revisions. *arXiv preprint arXiv:2606.13044*, 2026c.
- Mingxin Yao, Ying Wei, and Hao Liu. Ai practices and ethical concerns: An analysis of undeclared uses of ai in published research articles. *Ethics & Behavior*, 36(4):312–324, 2026.
- Rui Ye, Xianghe Pang, Jingyi Chai, Jiaao Chen, Zhenfei Yin, Zhen Xiang, Xiaowen Dong, Jing Shao, and Siheng Chen. Are we there yet? revealing the risks of utilizing large language models in scholarly peer review. *arXiv preprint arXiv:2412.01708*, 2024.
- Weiwei Ye, Hangchen Liu, Dongyuan Li, and Renhe Jiang. Paperclaw: Harnessing agents for autonomous research and human-in-the-loop refinement. *arXiv preprint arXiv:2606.22610*, 2026.

- Lei You, Lele Cao, and Iryna Gurevych. Preventing the collapse of peer review requires verification-first ai. *arXiv preprint arXiv:2601.16909*, 2026.
- Hongyeon Yu, Dongchan Kim, and Young-Bum Kim. Retrieval collapses when ai pollutes the web. In *Proceedings of the ACM Web Conference 2026*, pages 8745–8748, 2026.
- Jianxiang Yu, Zichen Ding, Jiaqi Tan, Kangyang Luo, Zhenmin Weng, Chenghua Gong, Long Zeng, Renjing Cui, Chengcheng Han, Qiushi Sun, et al. Automated peer reviewing in paper sea: Standardization, evaluation, and analysis. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 10164–10184, 2024a.
- Sungduk Yu, Man Luo, Avinash Madasu, Vasudev Lal, and Phillip Howard. Is your paper being reviewed by an llm? investigating ai text detectability in peer review. *arXiv preprint arXiv:2410.03019*, 2024b.
- Chunhong Yuan, Tianshi Wei, Chuangqi Li, Xing Yi, Shengrui Liu, Zixin Zhang, Yule Cai, and Xinke Du. Paperorchestrator: An llm-orchestrated multi-agent pipeline for automated end-to-end scientific paper writing. *Journal of King Saud University Computer and Information Sciences*, 2026.
- Sihang Zeng, Kai Tian, Kaiyan Zhang, Yuru Wang, Junqi Gao, Runze Liu, Sa Yang, Jingxuan Li, Xinwei Long, Jiaheng Ma, et al. Reviewrl: towards automated scientific review with rl. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pages 16942–16954, 2025.
- Bingchen Zhao, Jenny Zhang, Chenxi Whitehouse, Minqi Jiang, Michael Shvartsman, Abhishek Charnalia, Despoina Magka, Tatiana Shavrina, Derek Dunfield, Oisin Mac Aodha, et al. Apres: An agentic paper revision and evaluation system. *arXiv preprint arXiv:2603.03142*, 2026a.
- Xinyu Zhao, Rana Muhammad Shahroz Khan, Zhen Xu, Zhen Tan, and Tianlong Chen. Does ai reviewer see the full picture? attacking and defending multimodal peer review. *arXiv preprint arXiv:2606.12716*, 2026b.
- Zhenyue Zhao, Yihe Wang, Toby Stuart, Mathijs De Vaan, Paul Ginsparg, and Yian Yin. Llm hallucinations in the wild: Large-scale evidence from non-existent citations. *arXiv preprint arXiv:2605.07723*, 2026c.
- Qin Zhou, Zhixin Zhang, Zhi Li, and Limin Sun. " give a positive review only": An early investigation into in-paper prompt injection attacks and defenses for ai reviewers. *arXiv preprint arXiv:2511.01287*, 2025.
- Ruiyang Zhou, Lu Chen, and Kai Yu. Is llm a reliable reviewer? a comprehensive evaluation of llm on automatic paper reviewing tasks. In *Proceedings of the 2024 joint international conference on computational linguistics, language resources and evaluation (LREC-COLING 2024)*, pages 9340–9351, 2024a.
- Ying Zhou, Ben He, and Le Sun. Humanizing machine-generated content: Evading ai-text detection through adversarial attack. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 8427–8437, 2024b.
- Changjia Zhu, Junjie Xiong, Renkai Ma, Zhicong Lu, Yao Liu, and Lingyao Li. When your reviewer is an llm: Biases, divergence, and prompt injection risks in peer review. *arXiv preprint arXiv:2509.09912*, 2025a.
- Minjun Zhu, Yixuan Weng, Linyi Yang, and Yue Zhang. Deepreview: Improving llm-based paper review with human-like deep thinking process. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 29330–29355, 2025b.
- Zhenzhen Zhuang, Jiandong Chen, Hongfeng Xu, Yuwen Jiang, and Jialiang Lin. Large language models for automated scholarly paper review: A survey. *Information Fusion*, 124:103332, 2025.
- Wei Zou, Runpeng Geng, Binghui Wang, and Jinyuan Jia. {PoisonedRAG}: Knowledge corruption attacks to {Retrieval-Augmented} generation of large language models. In *34th USENIX Security Symposium (USENIX Security 25)*, pages 3827–3844, 2025.

A Supplemental details

This appendix consolidates the resources and studies discussed throughout the survey. Table A.1 summarizes datasets, benchmarks, and deployments. Table A.2 organizes the surveyed studies and systems using the taxonomy in Figure 2.

A.1 Datasets, benchmarks, and deployments

Table A.1: Datasets, benchmarks, and deployments discussed in the survey, organized by resource type, scale, and primary use.

Resource	Evidence type	Reported scale	Primary use
Scholarly records and training corpora			
PeerRead (Kang et al., 2018)	Corpus	14,784 papers and 10,770 reviews	Review and decision modeling
MReD (Shen et al., 2022)	Annotated corpus	7,089 meta-reviews, 45K sentences, and 9 labels	Controllable meta-review generation
OpenReviewer (Idahl and Ahmadi, 2025)	Training corpus	36K papers, 141K collected reviews, and about 79K retained	Scientific review generation
DeepReview-13K (Zhu et al., 2025b)	Training set and benchmark	13,378 valid samples and a 1.2K-sample test set	Structured review reasoning
RMR-75K / RBTAAct (Wu et al., 2026d)	Aligned corpus	75,542 review-rebuttal mappings across 4,825 papers	Actionable feedback generation
GoodPoint (Mun et al., 2026)	Longitudinal corpus	18,936 ICLR papers, 2020–2026	Feedback validity and actionability
Evaluation benchmarks and deployments			
ReviewEval (Garg et al., 2025)	Benchmark	120 papers from NeurIPS, ICLR, and UAI	Five-dimensional review evaluation
PRISM (Loc et al., 2026)	Benchmark	1,000 papers across 5 venue-years, with 5 AI reviewers plus humans	Multi-dimensional reviewer evaluation
CLAIMCHECK (Ou et al., 2025)	Expert-annotated benchmark	41 papers, 60 reviews, and 154 target claims	Claim-grounded critique
Soundness counterfactuals (Dyck and Gurevych, 2026)	Counterfactual benchmark	133 papers and 931 variants (391 critical, 540 neutral)	Faulty-reasoning sensitivity
Full-manuscript pre-review (Liang et al., 2024b)	Retrospective and prospective study	3,096 Nature-family papers, 1,709 ICLR papers, and 308 researchers	Author-facing feedback
ICLR feedback RCT (Thakkar et al., 2026)	Randomized deployment	44,831 reviews, 11,553 papers, and 18,946 feedback recipients	Reviewer assistance
AAAI-26 AI Review (Biswas et al., 2026)	Official deployment	22,977 papers and 5,834 survey responses	Auxiliary AI review
FARS (Tang et al., 2026b)	Public deployment	166 complete papers across 67 topics and 282 reviews of 140 papers	End-to-end production and evaluation
Manipulation, defense, and evasion test beds			
Hidden-prompt test bed (Zhu et al., 2025a)	Controlled benchmark	1,441 ICLR and NeurIPS papers	Reviewer manipulation
ARO (Torrielli et al., 2026)	Controlled benchmark	5,600 experiments with ChatGPT and Gemini	Cross-role prompt injection
PaperGuard (Zhao et al., 2026b)	Multimodal benchmark	1,136 papers and 17 in-the-wild attacks with source	Text and figure attacks and defenses
SafeReview (Xin et al., 2026)	Adaptive benchmark	500 NeurIPS 2024 training papers and 1,286 ICLR 2024 test papers	Co-evolutionary defense
Random-citation probe (Rao et al., 2025)	Randomized audit	Multiple LLMs and review datasets with a test of more than 10K reviews	Outsourced-review detection
IntraGuard (Ma et al., 2026a)	Controlled benchmark	17,844 cases across 7 chatbot settings and 12 venues	Hidden defensive probes
Mixed-review attribution (Saha et al., 2026)	Detection benchmark	50,156 reviews and 5 detectors	Policy enforceability
RAID (Dugan et al., 2024)	Robustness benchmark	More than 6M generations from 11 generators, covering 8 domains, 11 attacks, and 12 detectors	Detector evasion
Long-horizon reuse and impact records			
SPARK (Sanyal et al., 2025)	Training corpus and system	600K OpenReview reviews with more than 10,000 ideas filtered	Review-informed idea filtering
ReviewGuard (Rasool et al., 2026)	Longitudinal benchmark	20,861 papers, including 2,365 rejected-then-published papers	Citation-aligned evaluation
Dolma scholarly component (Soldaini et al., 2024)	Pretraining corpus	38.8M Semantic Scholar documents and 70B LLaMA tokens	Scholarly-record ingestion
OpenScholar DataStore (Asai et al., 2026)	Retrieval corpus	45M open-access papers	Citation-backed synthesis
Common Pile (Kandpal et al., 2026)	Pretraining corpus	8TB from 30 sources	Licensed scholarly-text reuse

A.2 Taxonomy index

Table A.2: Studies and systems organized by the mechanisms in the taxonomy shown in Figure 2.

Dynamic	Mechanism	Studies and systems
Production scaling	AI-assisted and agentic production	ResearchAgent (Baek et al., 2025), The AI Scientist (Lu et al., 2024), AI Scientist-v2 (Yamada et al., 2025), Agent Laboratory (Schmidgall et al., 2025), FARS (Tang et al., 2026b), PaperOrchestra (Song et al., 2026), PaperOrchestrator (Yuan et al., 2026)
	Faster output and iteration	Corpus-level LLM estimator (Liang et al., 2025), excess-vocabulary audit (Kobak et al., 2025), LLM-adoption panel (Kusumegi et al., 2025), preprint interrupted time series (Qi et al., 2025), matched-author panel (Filimonovic et al., 2025), Organization Science panel (Gartenberg et al., 2026), ICLR submission series (Knight, 2026)
	Post-submission scaling and review pressure	CycleResearcher (Weng et al., 2025), DRPG (Han et al., 2026), Paper2Rebuttal (Ma et al., 2026b), REspGen (Ruan and Gurevych, 2026), XtraGPT (Chen et al., 2026a)
Evaluation automation	Author-facing feedback	Full-manuscript pre-review (Liang et al., 2024b), Focused Feedback Generator (Chamoun et al., 2024), Checklist Assistant (Goldberg et al., 2024), drafting-feedback study (Wu, 2026), BOSC pre-review (Ohta et al., 2026)
	Reviewer assistance and official review	ICLR Review Feedback Assistant (Thakkar et al., 2026), AAAI-26 AI Review (Biswas et al., 2026), OpenReviewer (Idahl and Ahmadi, 2025), MReD (Shen et al., 2022)
	Grounding, auditability, and reliability	ReviewAgents (Gao et al., 2025), PeerCheck (Chen et al., 2026d), ReviewGrounder (Li et al., 2026e), ScholarPeer (Goyal et al., 2026), EGTR-Review (Qiu et al., 2026), DeepReviewer 2.0 (Weng et al., 2026), PRISM (Loc et al., 2026), CLAIMCHECK (Ou et al., 2025)
Evaluation manipulation	Concealed instructions	White/microscopic-text injection (Lin, 2026), positive-review injection (Zhou et al., 2025), invisible-text injection (Choi et al., 2026), multilingual hidden injection (Theocharopoulos et al., 2025), malicious-font injection (Xiong et al., 2025), indirect prompt injection (Torrielli et al., 2026)
	Presentation and multimodal steering	Adversarial Repackaging (Yang et al., 2026c), abstract-rephrasing attack (Li et al., 2026c), paraphrasing attack (Kaneko, 2026), Figure-only attack (Zhao et al., 2026b), textual adversarial perturbations (Lin et al., 2025)
	Adaptive evaluator optimization	LLM-REVal (Li et al., 2025b), Review Arcade (Hatzel et al., 2026), aspect-guided perturbation (Li et al., 2025a), optimization-based prompt injection (Shi et al., 2024), universal adversarial prompt (Raina et al., 2024), surrogate-feedback refinement (Zhou et al., 2025)
Defense mechanisms and policy responses	Detection and targeted verification	Quis (Kumar et al., 2024), Sem-Detect (Duarte et al., 2026), MixRevDetect (Kumar et al., 2025), CoCoNUTS (Chen et al., 2026c), random-citation watermark (Rao et al., 2025), IntraGuard (Ma et al., 2026a), topic-based watermarking (Nemecek et al., 2025)
	Evaluator and artifact protection	SafeReview (Xin et al., 2026), PaperGuard (Zhao et al., 2026b), TAB-AUDIT (Huang et al., 2026), SciFigDetect (Hu et al., 2026), detection prompt (Zhou et al., 2025)
Evasion and side effects	Post-generation evasion	RAID (Dugan et al., 2024), DIPPER (Krishna et al., 2023), recursive paraphrasing (Sadasivan et al., 2025), RAFT (Wang et al., 2024a), CoPA (Fang et al., 2025)
	Generation-time and route adaptation	Adversarial humanization (Zhou et al., 2024b), proxy attack (Wang et al., 2025), GradEscape (Meng et al., 2025), style-shifting attack (Pedrotti et al., 2025), MASH (Gu et al., 2026), constrained policy optimization (Wang et al., 2026c)
	Redistributed error, labor, and risk	Non-native-writer bias (Liang et al., 2023), accuracy-bias trade-off (Pratama, 2025), authorship misclassification (Boonsuk, 2026), AI-use stigmatization (Ali and Shaban, 2026), reviewer-attribute bias (Vasu et al., 2026), verification workload (Chen et al., 2025)
Long-horizon ecosystem feedback	Record ingress and persistence	Published-article diffusion audit (Siler, 2026), knowledge-update phrase audit (Strzelecki, 2025), Google Scholar paper audit (Haider et al., 2024), undeclared-AI audit (Yao et al., 2026), false-authorship case audit (Spinellis, 2025), fabricated-citation audit (Topaz et al., 2026), citation-propagation tracing (Camp et al., 2025)
	Reuse in future AI systems	Dolma/OLMo (Soldaini et al., 2024), OpenScholar (Asai et al., 2026), Common Pile (Kandpal et al., 2026), RBTAct (Wu et al., 2026d), GoodPoint (Mun et al., 2026), SPARK (Sanyal et al., 2025), ReviewGuard (Rasool et al., 2026)
	Propagation and the open causal loop	Knowledge-graph poisoning (Yang et al., 2024), PoisonedRAG (Zou et al., 2025), medical-model data poisoning (Alber et al., 2025), retrieval-pollution test (Yu et al., 2026), recursive model collapse (Shumailov et al., 2024), self-consuming loop (Alemohammad et al., 2024), diversity-collapse test (Guo et al., 2024), Artificial Hivemind (Jiang et al., 2026b)