Can AI Be a Good Peer Reviewer? A Survey of Peer Review Process, Evaluation, and the Future

Sihong Wu[1] **Owen Jiang**[1] **Yilun Zhao**[1*] **Tiansheng Hu**[2] **Yiling Ma**[1]
Kaiyan Zhang[1] **Manasi Patwardhan**[3] **Arman Cohan**[1]
[1]Yale University [2]New York University [3]TCS Research

Abstract

Peer review is a multi-stage process involving reviews, rebuttals, meta-reviews, final decisions, and subsequent manuscript revisions. Recent advances in large language models (LLMs) have motivated methods that assist or automate different stages of this pipeline. In this survey, we synthesize techniques for (i) peer review generation, including fine-tuning strategies, agent-based systems, RL-based methods, and emerging paradigms to enhance generation; (ii) after-review tasks including rebuttals, meta-review and revision aligned to reviews; and (iii) evaluation methods spanning human-centered, reference-based, LLM-based and aspect-oriented. We catalog datasets, compare modeling choices, and discuss limitations, ethical concerns, and future directions. The survey aims to provide practical guidance for building, evaluating, and integrating LLM systems across the full peer review workflow. A collection of papers is available at the following repository: GitHub Repository.

1 Introduction

Peer review is the cornerstone of scientific publishing. The rapid advancement of AI4Research has fundamentally transformed the landscape of scholarly communication, spawning a surge of interest in automating and enhancing the peer review process (Chen et al., 2025). The academic peer review pipeline includes submission, review, rebuttal, discussion, meta-review, and final decision-making. Each stage involves the generation of critical textual artifacts. For instance, research has explored the creation of initial referee reports (Yuan et al., 2021; Dycke et al., 2023a; Tan et al., 2024) and author rebuttals that respond to reviewer feedback (Cheng et al., 2020; Purkayastha et al., 2023a). Further along the pipeline, work has focused on consolidating multiple reviews into a unified meta-review (Li et al., 2023; Zeng et al., 2023; Kumar et al., 2024a) and providing structured arguments to support the final publication decision (Sukpanichnant et al., 2024).

Early approaches in this domain were confined to sub-tasks, such as predicting paper acceptance from metadata and abstract features or regressing review scores from paper content (Kang et al., 2018; Checco et al., 2021). The advent of LLMs marked a profound paradigm shift (Vaswani et al., 2017; Radford et al., 2019; Bommasani et al., 2021; Bai et al., 2023; Zhao et al., 2023; Touvron et al., 2023; OpenAI et al., 2024). Around 2023, models began to process full manuscripts and produce credible end-to-end reviews (Robertson, 2023; Hosseini and Horbach, 2023). Current systems go beyond single-prompt generation to adopt multi-agent designs that emulate panel workflows (Jin et al., 2024a; Durante et al., 2024) and reinforcement learning (Ziegler et al., 2020; Christiano et al., 2023) to align outputs with nuanced human preferences. These developments coexist with a growing recognition that evaluation must keep pace (Zhang and Abernethy, 2025).

While other valuable surveys exist, they either cover the AI4Research landscape too broadly, with peer review as only a minor component (Chen et al., 2025), or due to the field's rapid progress, do not fully capture the latest wave of agentic and RL-based methodologies (Zhuang et al., 2025). Crucially, few existing works provide a systematic analysis of the critical landscape of evaluation for peer review generation. Our survey fills this gap by providing a holistic overview of peer review process,[1] with a dual focus on cutting-edge generation methodologies and a taxonomy of their evaluation. The review procedure is in Appendix G.

This survey is structured as follows: Section 2

*Corresponding author (yilun.zhao@yale.edu)

[1]This survey is intended to support research on AI-assisted peer review workflows, not to replace human reviewers.

Proceedings of the 64th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), pages 32593–32619
July 2-7, 2026

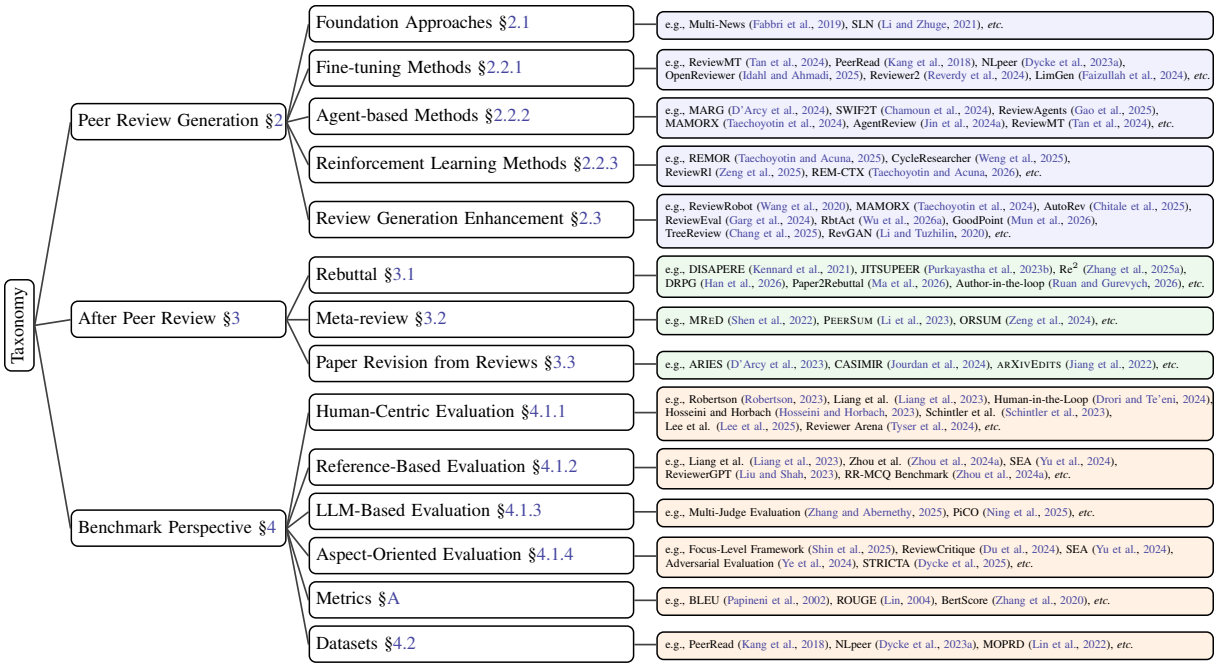

Figure 1: Taxonomy of AI for Peer Review Process and Evaluation: Key Areas and Example Systems.

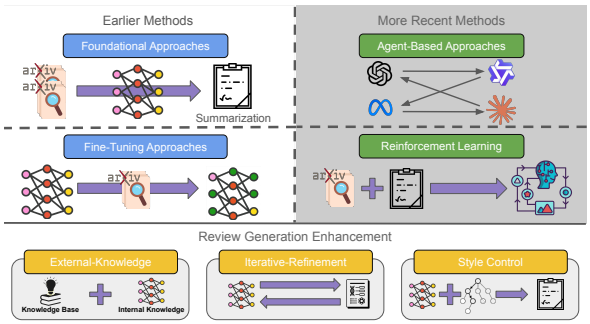

Figure 2: The methods of peer review generation: (1) Foundation approaches; (2) Fine-tuning methods; (3) Agent-based methods; (4) Reinforcement learning methods; and (5) Review Generation Enhancement.

details the evolution of peer review generation methodologies. Section 3 introduces methods of after-review tasks including rebuttal, meta-review and paper revision from reviews. Section 4 offers a systematic taxonomy of evaluation methods, metrics, and datasets. Section 5 discusses key challenges and outlines future research directions.

2 Peer Review Generation

We categorize peer review generation methodologies into five main paradigms—foundation approaches, fine-tuning methods, agent-based methods, reinforcement learning methods and generation enhancement, as illustrated in Figure 2.

2.1 Foundation Approaches

Prior to the widespread adoption of LLMs, the automated generation of complete, high-quality peer reviews was difficult (Drozdz and Ladomery, 2024). Early attempts to generate reviews often addressed a different problem, such as the multi-document summarization of existing reviews (Li et al., 2022; Fabbri et al., 2019), or relied on citation networks rather than generating novel critique directly from a source manuscript (Li and Zhuge, 2021). Consequently, research in the pre-2023 era predominantly focused on a suite of more manageable sub-tasks. These foundational efforts deconstructed the peer review process into analytical (Teufel et al., 1999), predictive (Kang et al., 2018), and narrowly generative components (Dycke et al., 2023b). While limited in their ability to produce complete reviews, these researches established critical datasets for subsequent breakthroughs.

2.2 LLM-based Approaches

We next review recent advancements in peer review generation using LLMs, including fine-tuning strategies, multi-agent systems, and reinforcement learning methods for optimizing review quality.

2.2.1 Fine-tuning Methods

To overcome the limitations of zero-shot prompting (Li, 2023; Sivarajkumar et al., 2023), such as the generation of overly positive and generic feedback and the inability to produce reviews that con-

Method & Framework	Venue/Year	Core Paradigm	Dataset(s)	Key Contributions & Features
Foundational Datasets				
PeerRead Kang et al. (2018)	NAACL 2018	Data Collection	PeerRead	Large-scale papers+reviews dataset.
NLpeer Dycke et al. (2023a)	ACL 2023	Data Collection	NLpeer	Unified, expanded peer-review resource.
MOPRD Lin et al. (2022)	NCA 2023	Data Collection	MOPRD	Multidisciplinary open peer-review set.
LLM-based Generation Approaches				
A. Fine-tuning Methods				
OpenReviewer Idahl and Ahmadi (2025)	NAACL Demo 2025	Fine-tuning	79k expert reviews	Llama-8B fine-tuned on curated expert reviews.
REVIEWER2 Gao et al. (2024b)	Preprint 2024	Fine-tuning	27k papers, 99k reviews	Two stage: prompt generation and review.
LimGen Faizullah et al. (2024)	ECML-PKDD 2024	Fine-tuning	LimGen	SLG for limitation suggestions.
B. Agent-based Systems				
MARG D'Arcy et al. (2024)	Preprint 2024	Agent-based (Dec.)	–	Leader–worker agents; iterative refinement.
ReviewAgents Gao et al. (2025)	Preprint 2025	Agent-based (Dec.)	Review-CoT	Summarize, analyze and conclude with RAG.
DeepReview Zhu et al. (2025a)	ACL 2025	Agent-based (Dec.)	DeepReview-13K	Three stages: novelty, multi-dimension, reliability.
MAMORX Taechoyotin et al. (2024)	NeurIPS 2024 WS	Agent-based (Dec.)	–	Multimodal; function-calling RAG for novelty.
SWIF²T Chamoun et al. (2024)	ACL Find. 2024	Agent-based (Dec.)	300 peer reviews	Four agents: plan/investigate/review/control.
DIAGPaper Zou et al. (2026)	Preprint 2026	Agent-based (Dec.)	AAAR, ReviewCritique	Multi-agent weakness diagnosis
AgentReview Jin et al. (2024a)	EMNLP 2024	Agent-based (Sim.)	–	Simulated peer review; authority-bias study.
ReviewMT Tan et al. (2024)	Preprint 2024	Agent-based (Sim.)	ReviewMT	SFT > zero-shot; rebuttal as dialogue.
C. Reinforcement Learning (RL)				
Remor Taechoyotin and Acuna (2025)	Preprint 2025	RL (GRPO)	PeerRT	Multi-objective RL with HPRR reward.
CycleResearcher Weng et al. (2025)	ICLR 2025	RL (SimPO)	Review-5k, Research-14k	Dual-agent research–review–refine loop.
ReviewRL Zeng et al. (2025)	EMNLP 2025	RL (Rule-based)	ICLR 2025 papers	Composite-reward RL for grounded reviews.
REM-CTX Taechoyotin and Acuna (2026)	Preprint 2026	RL (GRPO)	PeerRTEx, FCR/NCRDat	Auxiliary-context RL for grounded peer reviews.
Review Generation Enhancement				
ReviewRobot Wang et al. (2020)	INLG 2020	RAG (KG-based)	–	Pre-LLM; synthesizes three knowledge graphs.
Novelty Afzal et al. (2025)	EACL 2026	RAG (Novelty)	–	Contribution-wise novelty comparisons.
ReviewEval Garg et al. (2024)	EMNLP Find. 2025	Iterative Refinement	FullCorpus-120	Self-refinement + external improvement loops.
RbtAct Wu et al. (2026a)	ACL Find. 2026	Iterative Refinement	RMR-75K	Rebuttal-derived optimization for actionability.
GoodPoint Mun et al. (2026)	Preprint 2026	Iterative Refinement	GoodPoint-ICLR	Author-response supervision.
TreeReview Chang et al. (2025)	EMNLP 2025	Structure Control	Venue-derived benchmark	Dynamic question tree for analysis.
AutoRev Chitale et al. (2025)	Preprint 2025	Structure Control	–	Document graph to target key passages.
RevGAN Li and Tuzhilin (2020)	EMNLP 2019	Style Control (GAN)	–	Controllable, personalized reviews (pre-LLM).

Table 1: A summary of methodologies in peer review generation. The table also highlights key contributions. "(Dec.)" denotes Task Decomposition and "(Sim.)" denotes Process Simulation.

form to the required format (Tan et al., 2024; Gao et al., 2024b), research has rapidly pivoted towards fine-tuning LLMs on domain-specific data (Hu et al., 2021). Some of these datasets are directly derived from foundational datasets (Kang et al., 2018; Dycke et al., 2023a), while others are constructed by leveraging APIs and specialized tools to curate custom review datasets (we summarize methods in Appendix E) (Dycke et al., 2022).

The contemporary approach involves specializing LLMs to capture the distinct style and critical nature of peer review. To address the challenge of generating reviews in the required format, Tan et al. (2024) has fine-tuned LLMs on the ReviewMT dataset. Experimental results demonstrate that such supervised fine-tuning substantially improves the review and decision hit rates compared to zero-shot prompting. OpenReviewer (Idahl and Ahmadi, 2025) fine-tunes a Llama-8B model (AI@Meta, 2024) on a curated dataset of 79,000 expert reviews. Its high performance stemmed from curating reviews with diverse critique patterns such as methodological flaws, reproducibility issues and balanced scores. We also organize different LLM Backbone Strategies in Table 3 in Appendix.

Other frameworks decompose the generation task itself to improve specificity and control. REVIEWER2 (Reverdy et al., 2024) utilizes a two-stage pipeline where one fine-tuned model first generates a set of relevant aspect prompts for a given paper, and a second model generates review text conditioned on these specific aspects. This decomposition helps lead to more detailed reviews that better cover the range of topics. The versatility of fine-tuning is further demonstrated by its application to specific sub-tasks, as seen in LimGen (Faizullah et al., 2024), which fine-tunes models specifically for a task requires inferring potential weaknesses not explicitly stated by authors.

2.2.2 Agent-based Methods

Recognizing that peer review is a complex cognitive process, recent work has shifted towards agent-based systems. This paradigm decomposes the review process, assigning distinct roles and sub-tasks to multiple collaborating LLMs. These approaches can be broadly categorized by their primary objective: *Task Decomposition* and *Process Simulation*.

Task Decomposition. A significant body of work focuses on task decomposition to generate higher-

quality reviews. D'Arcy et al. (2024) propose MARG, a framework utilizing a leader-worker architecture where specialized expert agents are tasked with critiquing specific aspects of a paper, such as its experiments, clarity, and impact. This division of labor helps produce more targeted, helpful feedback. Similarly, SWIF²T, introduced by Chamoun et al. (2024), employs a four-component pipeline of Planner, Investigator, Reviewer, and Controller agents to generate focused and actionable feedback on specific weaknesses identified within a manuscript. Building on this concept of emulating human workflows, Gao et al. (2025) develop ReviewAgents, a framework designed to mirror the structured reasoning of human experts through summarization, analysis, and conclusion stages. DeepReview (Zhu et al., 2025a) introduces a multi-stage framework that decomposes the review into three stages: novelty verification, multi-dimension review and reliability verification, aims to mitigate issues of limited domain expertise and hallucinated reasoning in LLM-based reviewers. Pushing the boundaries of analysis beyond text, Taechoyotin et al. (2024) present MAMORX, a multi-agent system that integrates attention to text, figures, and citations together with external knowledge sources. DIAGPaper (Zou et al., 2026) further decomposes weakness identification into criterion-grounded reviewing, rebuttal-based validation, and severity prioritization, improving the validity and specificity of detected weaknesses.

Process Simulation. A distinct but complementary line of research uses multi-agent systems to simulate and study the peer review process itself. Jin et al. (2024a) introduce AgentReview, an extensible simulation testbed populated by LLM agents representing authors, reviewers, and area chairs. In a similar vein, Tan et al. (2024) reformulate peer review as a multi-turn, long-context dialogue among Reviewer, Author, and Decision Maker agents. This approach, supported by the large-scale ReviewMT dataset, explicitly models the iterative rebuttal and discussion cycles that are central to real-world academic review but are missed by generation models. Collectively, these multi-agent methodologies represent a significant leap in higher-quality reviews with dynamic simulation for deeper process understanding.

2.2.3 Reinforcement Learning Methods

Beyond supervised fine-tuning, which primarily teaches models stylistic and structural conventions, reinforcement learning (RL) has emerged as a crucial paradigm for optimizing generated content (Bai et al., 2022; Swamy et al., 2025). This approach allows models to learn from feedback signals that represent complex goals, such as producing insightful and helpful scientific critiques. The application of RL in this domain shows a clear trajectory of optimizing the entire scientific process (Rao et al., 2020; Novikov et al., 2025).

A direct application of this principle is seen in Remor (Taechoyotin and Acuna, 2025), which employs multi-objective reinforcement learning to address the common failure mode of AI-generated reviews being shallow (Wei et al., 2025). The framework's core innovation is its Human-aligned Peer Review Reward, a composite function that quantifies multiple facets of review quality, including criticism, relevance, and actionable suggestions. By optimizing this multi-objective reward using Group Relative Policy Optimization (GRPO) (Shao et al., 2024), Remor learns to generate reviews that are better aligned with nuanced human preferences.

Expanding this paradigm, other frameworks use the generated review itself as a reward signal to refine the primary artifact. CycleResearcher (Weng et al., 2025) exemplifies this with a dual-agent system where a CycleReviewer model, trained to mimic human evaluation scores, provides a reward signal for a CycleResearcher model that generates manuscripts (Meng et al., 2024). This establishes a closed "Research-Review-Refinement" loop (Tang et al., 2025). ReviewRL (Zeng et al., 2025) extends this line by combining retrieval-augmented context construction with composite-reward RL to jointly improve review quality and rating consistency. Some RL work also incorporates auxiliary context beyond the manuscript text itself, where REM-CTX (Taechoyotin and Acuna, 2026) uses correspondence-aware rewards to align generated reviews with figures and external scholarly signals.

2.3 Review Generation Enhancement

External Knowledge-Enhanced Generation. A significant challenge in peer review generation is ensuring that comments are not only coherent but also factually grounded to prevent hallucination problems (Shuster et al., 2021; Ádám Kovács and Recski, 2025). Models that generate reviews based

solely on the text of a submitted paper may produce generic critiques, lacking the necessary context of the broader scientific landscape. To address this, a key line of research has focused on enhancing generation systems with external knowledge. These approaches parallel the wider adoption of Retrieval-Augmented Generation (RAG) (Lewis et al., 2021; Gao et al., 2024a). Early, pre-LLM work in this area pioneered the use of structured knowledge synthesis to generate evidence-backed reviews (Nauta et al., 2023). ReviewRobot (Wang et al., 2020) operationalizes this by constructing three distinct knowledge graphs (KGs): one from the target paper, one from its cited works, and a background KG from a large collection of domain literature. With the advent of LLMs, methodologies have shifted towards more direct integration of external knowledge. MAMORX (Taechoyotin et al., 2024) exemplifies this modern RAG paradigm by employing a specialized agent to query external scholarly databases to assess a paper's novelty. Afzal et al. (2025) retrieves and re-ranks related literature, performing contribution-wise comparisons to produce novelty judgments.

Iterative Refinement. This paradigm replaces the single-pass generation approach with multi-step processes where an initial output is progressively improved (Kamoi et al., 2024). The refinement is driven by feedback generated either by the model itself or through interaction with other modules or agents (Gou et al., 2024; Han et al., 2024). D'Arcy et al. (2024) operationalizes this with an explicit, final refinement stage. After aspect-specific agents generate initial comments, a separate multi-agent group is convened to assess each comment for validity and clarity, collaboratively deciding whether to revise or prune it. Garg et al. (2024) proposes a self-refinement loop and an external improvement loop that iteratively optimizes its intermediate outputs and the final reviews. Taking a different approach, Tan et al. (2024) inherently creates an iterative refinement cycle where the author's rebuttal serves as feedback. Taechoyotin and Acuna (2025) also employ an RL-based refinement strategy, but focus on generating the review itself. RbtAct (Wu et al., 2026a) similarly leverages rebuttals as implicit supervision and preference signals to train generators that produce more actionable review feedback. GoodPoint (Mun et al., 2026) also uses author responses to define constructive feedback as comments that are both valid and actionable.

Structure and Style Control. Structure refers to the logical organization, format, and argumentative framework of a review. Generating a review is a complex task that requires beyond insight into technical. TreeReview (Chang et al., 2025) models peer review as a hierarchical question-answering process. The process begins with a top-down decomposition stage, where a high-level review task is recursively broken down into a tree of more fine-grained sub-questions. AutoRev (Chitale et al., 2025) leverages graph-based structural modeling of the paper itself, capturing hierarchical relationships between passages. The RevGAN model (Li and Tuzhilin, 2020), developed for generating controllable and personalized product reviews, provides a clear and powerful architecture for fine-grained stylistic control.

3 After Peer Review

Beyond generating reviews, several tasks model the post-review stage including rebuttal generation, meta-review generation and paper revision.

3.1 Rebuttal

Rebuttal generation aims to produce author responses that directly and faithfully address reviewer critiques (Gao et al., 2019; Huang et al., 2023). Early resources focused on linking reviews to rebuttals and labeling discourse functions, enabling supervised conditioning of responses on critique aspects and actions (Kennard et al., 2021) and argument-pair extraction between review claims and rebuttal (Cheng et al., 2020). The first work to explicitly formulate rebuttal generation is (Purkayastha et al., 2023b), who introduce JITSUPEER and cast the task as attitude root or theme-guided generation of canonical rebuttals conditioned on rebuttal actions. Recent datasets move from single-turn to multi-turn dialog settings by involving rebuttal: ReviewMT reframes peer review as long-context, role-based interaction (Tan et al., 2024), and Re^2 (Zhang et al., 2025a) aggregates review data into structured multi-turn rebuttal discussions. More recently, the focus is shifting from simple sentence generation to strategic, evidence-based dialog planning. DRPG (Han et al., 2026) formalizes this shift through a four-stage pipeline of concern decomposition, evidence retrieval, perspective planning, and response generation. Paper2Rebuttal (Ma et al., 2026) further reframes rebuttal assistance as a verify-then-write evidence

organization problem with explicit checkpoints for grounding and consistency. Recent author-in-the-loop work (Ruan and Gurevych, 2026) also shows that effective response generation benefits from explicit author signals, controllable planning, and evaluation-guided refinement. Even minimal author guidance can improve factual correctness and targeted refutation over direct generation baselines (Khatri and Patwardhan, 2026).

3.2 Meta-review

Meta-review generation synthesizes multiple reviewer opinions into a summary for area chairs. Early systems framed the task as extract-then-write with predicting the accept/reject and conditioning generation on it (Bhatia et al., 2020; Kumar et al., 2024b). Public datasets have enabled further study. MRED adds sentence-level functional labels to meta-reviews for structure-controllable generation (Shen et al., 2022); PEERSUM models the hierarchical conversational structure of reviews–rebuttals and introduces RAMMER for structure-aware aggregation (Li et al., 2023). ORSUM scales coverage across venues while proposing checklist-guided, multi-stage introspection for opinion consolidation and evaluation (Zeng et al., 2024). Methods increasingly leverage argument structure and rebuttal content to surface consensus vs. controversy (Wu et al., 2022), and decompose generation via facet-level judgement extraction and sentiment consolidation to improve decision quality (Li et al., 2024). Purkayastha et al. (2026) also reframes meta-reviewing as a document-grounded dialogue problem, emphasizing deliberation and decision support beyond summarization alone.

3.3 Paper Revision from Reviews

A further step is editing the manuscript itself based on peer feedback. ARIES (D'Arcy et al., 2023) introduces the task and a dataset aligning review comments to concrete paper edits, enabling edit generation. CASIMIR (Jourdan et al., 2024) compiles and analyzes papers and their reviews to understand the intent behind edits, which helps plan and evaluate future revisions. ARXIVEDITS (Jiang et al., 2022) provides gold sentence alignments across versions and fine-grained span-level edit intents, complementing review-linked corpora.

4 Benchmark Perspective

Evaluating generated reviews remains challenging due to the subjective nature of review quality. Ensuring reliability is essential for maintaining the integrity of academic publishing (Zhou et al., 2024a).

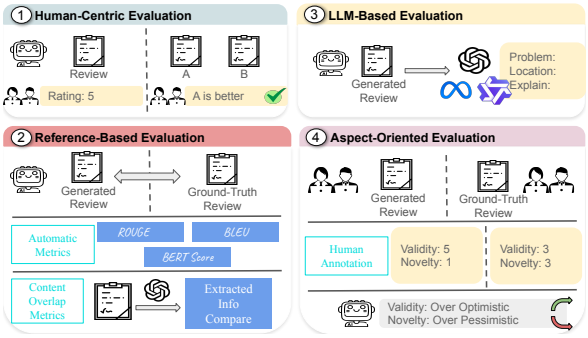

Figure 3: The main evaluation methods that we discuss are: (1) Human-centric evaluation; (2) Reference-based automated evaluation; (3) LLM-based evaluation; and (4) Aspect-oriented evaluation.

4.1 Evaluation Methods

The validity, trustworthiness, and ultimate utility of any automated review generation system hinge entirely on the rigor of its evaluation. A poorly evaluated system risks amplifying biases, providing superficial feedback, or undermining the very scientific integrity it aims to support (Zhang and Abernethy, 2025; Tyser et al., 2024; Lee et al., 2013; Bender et al., 2021). Consequently, the methods used to evaluate these systems are as important as the systems themselves. This review provides a systematic taxonomy of the evaluation methodologies, in Figure 3, that have emerged in this rapidly developing field, with their pros and cons, which is also shown in Table 2 and Appendix F.

4.1.1 Human-Centric Evaluation

The most straightforward method for evaluating a generated peer review is to ask a human expert to assess its quality. Robertson (2023) conducted the first pilot study comparing GPT-4 to human reviewers: ten participants rated the helpfulness of each review, finding identical mean scores for GPT-4 and human reviews, although GPT-4's responses showed higher variance. Expanding on this approach, Liang et al. (2023) conducted a large-scale user study involving 308 researchers who received LLM-generated feedback on their own papers, finding that users perceived the feedback as valuable. Drori and Te'eni (2024) structured GPT-4 reviews against standard review criteria and concluded that AI assistance can meaningfully reduce reviewer workload but is "not completely" reliable for all cases. However, qualitative critiques warn

of risks: Hosseini and Horbach (2023) reviews the potential benefits and risks of using LLMs in academic peer review, highlighting both efficiency gains and concerns about bias and reproducibility, and Schintler et al. (2023) emphasize ethical concerns that unchecked AI could compromise review integrity. In a domain-specific review, Lee et al. (2025) highlight that LLMs excel at language tasks (screening, summarization, language editing) but struggle with assessing scientific validity, recommending they be used cautiously as aids under clear guidelines. Recognizing the difficulty of assigning absolute numerical scores, a parallel approach has gained prominence: leveraging relative human preference. This principle has given rise to "Arena" platforms (Chiang et al., 2024; Zhao et al., 2025). Tyser et al. (2024) employs an arena-based methodology called Reviewer Arena, leveraging pairwise human (and LLM-predicted) preference comparisons to evaluate the quality of LLM-generated academic reviews.

4.1.2 Reference-Based Evaluation

To overcome the scalability limitations of human-centric evaluation, researchers have long relied on automated metrics. The evaluation assumes that a higher-quality generated review will exhibit greater lexical and semantic overlap with its human counterparts. This approach was a first step, leveraging metrics like BLEU, ROUGE, and BERTScore (Papineni et al., 2002; Lin, 2004; Ganesan, 2018; Zhang et al., 2020). We provide a detailed summary of the metrics used for peer review generation evaluation in Appendix A. Liang et al. (2023) propose a two-stage evaluation: first extracting key comments from reviews with GPT-4, then semantically matching them to compute a content overlap "hit rate", offering a more nuanced assessment than document-level metrics. Zhou et al. (2024a) employed a suite of metrics including ROUGE and BERTScore to compare generated reviews against human references. The SEA framework proposed by Yu et al. (2024) also utilizes these metrics and highlights that ROUGE is indicative of the key information. These metrics provide a scalable, automated way to assess surface-level quality but are limited in their ability to evaluate deeper aspects like correctness or constructiveness (Asano et al., 2017). Other work compares LLM outputs to known answers. Liu and Shah (2023) evaluate LLMs on targeted tasks such as error detection and checklist verification (Büyükkaramikli et al., 2019). Zhou et al. (2024a) created an RR-MCQ benchmark and tested LLMs on multiple-choice questions about papers. Both studies find while LLMs achieve reasonable accuracy, they still struggle with complex review tasks and critical feedback.

4.1.3 LLM-Based Evaluation

The LLM-Based paradigm has emerged as a scalable alternative to human-centric and reference-based evaluation (Gu et al., 2025). Zhang and Abernethy (2025) used a multi-judge pipeline where two distinct LLM judges must agree for a case to count as a "hit", reducing single-model bias. Recent work also explores unsupervised evaluation without gold labels (Lin and Chen, 2023; Allamanis et al., 2024; Ning et al., 2025).

4.1.4 Aspect-Oriented Evaluation

Moving beyond generic similarity, the aspect-oriented paradigm deconstructs review quality into specific, fine-grained dimensions such as correctness, clarity, substance, and impact (Lu et al., 2025). Some work evaluates reviews by examining specific facets or error types. Shin et al. (2025) develop a focus-level framework: they annotate review content by targets and aspects, then compare where LLM reviews concentrate versus human reviews. Du et al. (2024) released ReviewCritique, a dataset annotated with 23 fine-grained error types at the sentence level, such as "Unstated Statement", "Missing Reference", testing whether an LLM-generated review flags the same issues as humans. STRICTA (Dycke et al., 2025) decomposes peer review into a graph of aspect-level reasoning steps, annotated by experts, thereby offering interpretable evaluation beyond black-box scoring. A complementary line evaluates review utility from the author's perspective through comment-level dimensions (Sadallah et al., 2025). Adversarial testing can be seen as a specific form of aspect-oriented evaluation focused on robustness and vulnerability (Zhang et al., 2025b; Liao et al., 2025). Ye et al. (2024) found that LLMs are vulnerable to explicit manipulation through review injection attacks and to implicit manipulation by overemphasizing minor limitations. Yu et al. (2024) introduced a novel metric called the mismatch score as part of their SEA framework, predicting the degree of inconsistency between a given paper and a review.

Methodology	Pros	Cons
Human-Centric	• **Most Direct Assessment:** Best captures nuanced qualities such as helpfulness, insightfulness, and practical utility. • **Strong Reliability in Pairwise Settings:** Relative preference judgments are often more consistent than absolute ratings.	• **Limited Scalability:** Expensive and time-consuming, especially when expert reviewers are required. • **Subjectivity and Bias:** Prone to inter-annotator disagreement, personal bias, and rubric inconsistency.
Reference-Based	• **Scalable and Efficient:** Metrics such as ROUGE and BERTScore are fast and easy to apply at scale. • **Reproducible:** Produces deterministic scores that are easy to compare across systems. • **Useful for Targeted Subtasks:** Can support focused evaluation such as MCQ answering or error detection.	• **Shallow Assessment:** Often captures lexical or semantic overlap rather than scientific validity, reasoning quality, or constructiveness. • **Reference Dependency:** Relies heavily on the quality and coverage of human-written references, and may penalize valid but differently phrased feedback.
LLM-Based	• **Scalable and Nuanced:** Combines automation with the ability to assess higher-level qualities using rubric-based prompting. • **Reduced Reference Dependence:** Can evaluate outputs without requiring gold reviews for every instance. • **Flexible:** Evaluation criteria can be adapted easily through natural language instructions.	• **LLM Judge Biases:** Susceptible to position bias, verbosity bias, and style preference. • **Reliability Concerns:** Sensitive to judge model choice, prompt design, and calibration, and may still fail on complex scientific reasoning. • **Cost:** API-based evaluation can be expensive at scale.
Aspect-Oriented	• **Fine-Grained Diagnosis:** Reveals strengths and weaknesses along specific dimensions such as novelty, clarity, or evidence use. • **Supports Targeted Improvement:** Helps identify concrete failure modes for model development. • **Specialized Testing:** Enables focused evaluation such as robustness or deficiency-specific analysis.	• **High Annotation Overhead:** Often requires task-specific schemas and manually annotated data. • **Harder to Aggregate:** Produces multidimensional profiles rather than a single summary score, which can complicate direct model comparison.

Table 2: Pros and cons of evaluation methods for peer review generation. Each paradigm reflects a different trade-off between scalability, cost, and evaluative depth.

4.2 Datasets

We collected datasets for peer review across two broad eras: pre-2023 and post-2023, which is shown in Table 7. Pre-2023 resources such as PEERREAD (Kang et al., 2018) laid the groundwork for review tasks. These early datasets primarily enabled score prediction, acceptance classification, and basic review generation tasks. Since 2023, a wave of new datasets has significantly expanded the scope and depth. Compared to earlier corpora, these newer datasets are more diverse, task-specific, and comprehensive.

Despite this progress, we discuss several key challenges remain in benchmark perspective in Appendix B, along with recommendations for improving dataset quality.

5 Discussion and Future Directions

① **Novelty Evaluation.** While LLMs can fluently mimic reviewer tone, they remain weak at judging true novelty. Early systems like OPENREVIEWER (Idahl and Ahmadi, 2025) and REVIEWER2 (Gao et al., 2024b) generate coherent critiques but fail to distinguish incremental from groundbreaking work without broader scientific context. Recent efforts explicitly tackle novelty evaluation in peer review: SchNovel creates benchmark for novelty judgments (Lin et al., 2024), and structured pipelines achieve high agreement with human reviewers in distinguishing genuine contributions (Afzal et al., 2025). Other directions integrate literature graphs for contribution scoring (Rubaiat et al., 2025). NovBench (Wu et al., 2026b) also isolates textual novelty assessment as a dedicated evaluation problem, further highlighting the difficulty current LLMs face in producing reliable novelty judgments.

② **Automated Evaluation.** Review quality remains hard to quantify. While models like MARG (D'Arcy et al., 2024) and SWIF^{2}T (Chamoun et al., 2024) decompose review generation into targeted components, evalu-

ation is often limited to coarse metrics like BLEU. New datasets such as REVIEWCRITIQUE (Du et al., 2024) and SUBSTANREVIEW (Guo et al., 2023) offer fine-grained labels that allow content-aware judge. Future research could combine sentence-level assessments with reviewer consistency and informativeness benchmarks.

③ **Beyond NLP and AI Domains.** Most peer review datasets to date are sourced from NLP and ML conferences (e.g., PeerRead (Kang et al., 2018), NLPEER (Dycke et al., 2023a)), limiting generalization. Multidisciplinary resources such as MOPRD (Lin et al., 2022) are a welcome step, but broader coverage is necessary to evaluate whether current models overfit to domain-specific language.

④ **Beyond Review Generation.** While peer review generation and evaluation have received considerable attention (Zhu et al., 2025a), the automation of subsequent steps in the review process remains underexplored. Tasks such as rebuttal generation (Gao et al., 2019; Purkayastha et al., 2023b), meta-review drafting (Li et al., 2023), and paper revision from reviews (D'Arcy et al., 2023) involve richer discourse structures and complex reasoning about critique-response dynamics. Compared to review generation, these tasks demand deeper alignment with argumentative discourse and collaborative intent, yet existing benchmarks and methodologies are sparse. Addressing these underdeveloped areas opens new challenges for dataset construction, evaluation, and the design of models that can faithfully support the full peer review process.

⑤ **Multimodal Review Tasks.** Current review models operate on textual inputs, yet real submissions often include figures, tables, or code. Multimodal systems such as MAMORX (Taechoyotin et al., 2024) extend the frontier by analyzing visual and tabular components, offering richer critiques. This line of work should be expanded with datasets that pair visual artifacts and expert annotations. We further discuss this task in Appendix C.3.

⑥ **Ethical and Transparent Deployment.** A primary concern is the potential for these models to introduce and amplify biases. For instance, studies have shown that LLMs can exhibit affiliation bias, favoring authors from highly-ranked institutions in single-blinded settings (von Wedel et al., 2024). Furthermore, a corpus-level study estimates that 6.5-16.9% of review text at top AI venues was likely substantially modified by LLMs (beyond simple grammar fixes), with usage spiking near deadlines and among low-confidence reviews (Liang et al., 2024). It is essential for the research community to establish clear guidelines for the ethical use of LLMs in peer review. This includes a call for transparency, where the use of an LLM in any part of the review process is explicitly disclosed.

6 Conclusion

The landscape of peer review process has undergone a transformative shift, driven by advances in large language models and latest paradigms. This survey has provided a comprehensive synthesis of state-of-the-art methods, spanning review generation as well as after-review tasks including rebuttal, meta-review, and revision. We also present a systematic taxonomy of evaluation methods, metrics and datasets. Despite remarkable improvements in generation methods, significant challenges remain, including robust content-level evaluation and extensive capabilities. As the field advances, future research must prioritize transparent and ethical deployment of automated reviews.

Limitations

While this survey strives to provide a comprehensive and up-to-date overview of the peer review landscape, several limitations remain. First, the rapid pace of progress in large language models, agent-based systems, and evaluation protocols means that new methodologies and benchmarks may emerge soon after publication, potentially outdating some of our analyses. Second, the majority of available datasets and evaluation studies are concentrated in NLP and machine learning domains, limiting the generalizability of our findings to other scientific fields. Finally, our survey is based on publicly accessible literature and resources; proprietary systems or unpublished industrial advances may not be adequately represented. Despite these limitations, we believe this survey remains the most comprehensive and fine-grained taxonomy of peer review process and its evaluation to date.

References

2008–2025. Grobid. https://github.com/kermitt2/grobid.

2017–2025. Science parse. https://github.com/allenai/science-parse.

2023–2025. Marker. https://github.com/datalab-to/marker.

Osama Mohammed Afzal, Preslav Nakov, Tom Hope, and Iryna Gurevych. 2025. Beyond "not novel enough": Enriching scholarly critique with llm-assisted feedback.

Mistral AI. 2023. Mistral-7b-instruct-v0.2. https://huggingface.co/mistralai/Mistral-7B-Instruct-v0.2. Accessed: 2025-07-28.

Mistral AI. 2024. Mistral large. https://mistral.ai/news/mistral-large/. Accessed: 2025-07-28.

AI@Meta. 2024. Llama 3 model card.

Miltiadis Allamanis, Sheena Panthaplackel, and Pengcheng Yin. 2024. Unsupervised evaluation of code llms with round-trip correctness.

Hiroki Asano, Tomoya Mizumoto, and Kentaro Inui. 2017. Reference-based metrics can be replaced with reference-less metrics in evaluating grammatical error correction systems. In *Proceedings of the Eighth International Joint Conference on Natural Language Processing (Volume 2: Short Papers)*, pages 343–348, Taipei, Taiwan. Asian Federation of Natural Language Processing.

Jinze Bai, Shuai Bai, Yunfei Chu, Zeyu Cui, Kai Dang, Xiaodong Deng, Yang Fan, Wenbin Ge, Yu Han, Fei Huang, Binyuan Hui, Luo Ji, Mei Li, Junyang Lin, Runji Lin, Dayiheng Liu, Gao Liu, Chengqiang Lu, Keming Lu, Jianxin Ma, Rui Men, Xingzhang Ren, Xuancheng Ren, Chuanqi Tan, Sinan Tan, Jianhong Tu, Peng Wang, Shijie Wang, Wei Wang, Shengguang Wu, Benfeng Xu, Jin Xu, An Yang, Hao Yang, Jian Yang, Shusheng Yang, Yang Yao, Bowen Yu, Hongyi Yuan, Zheng Yuan, Jianwei Zhang, Xingxuan Zhang, Yichang Zhang, Zhenru Zhang, Chang Zhou, Jingren Zhou, Xiaohuan Zhou, and Tianhang Zhu. 2023. Qwen technical report.

Yuntao Bai, Andy Jones, Kamal Ndousse, Amanda Askell, Anna Chen, Nova DasSarma, Dawn Drain, Stanislav Fort, Deep Ganguli, Tom Henighan, Nicholas Joseph, Saurav Kadavath, Jackson Kernion, Tom Conerly, Sheer El-Showk, Nelson Elhage, Zac Hatfield-Dodds, Danny Hernandez, Tristan Hume, Scott Johnston, Shauna Kravec, Liane Lovitt, Neel Nanda, Catherine Olsson, Dario Amodei, Tom Brown, Jack Clark, Sam McCandlish, Chris Olah, Ben Mann, and Jared Kaplan. 2022. Training a helpful and harmless assistant with reinforcement learning from human feedback.

Emily M Bender, Timnit Gebru, Angelina McMillan-Major, and Shmargaret Shmitchell. 2021. On the dangers of stochastic parrots: Can language models be too big? In *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency*, pages 610–623.

Prabhat Kumar Bharti, Shashi Ranjan, Tirthankar Ghosal, Mayank Agrawal, and Asif Ekbal. 2021. PEERAssist: Leveraging on paper–review interactions to predict peer review decisions. In *Proceedings of the 23rd International Conference on Asia-Pacific Digital Libraries (ICADL 2021)*, Lecture Notes in Computer Science, pages —. Springer. Dataset and code: https://github.com/PrabhatkrBharti/PEERAssist.

Chaitanya Bhatia, Tribikram Pradhan, and Sukomal Pal. 2020. Metagen: An academic meta-review generation system. In *Proceedings of the 43rd International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '20, page 1653–1656, New York, NY, USA. Association for Computing Machinery.

Rishi Bommasani, Jack Clark Hudson, Ehsan Adeli, Sam Altman, Simran Arora, Sydney von Arx, Michael Bernstein, and et al. 2021. On the opportunities and risks of foundation models.

N. C. Büyükkaramikli, M. P. M. H. Rutten-van Mölken, J. L. Severens, and M. Al. 2019. TECH-VER: A verification checklist to reduce errors in models and improve their credibility. *Pharmacoeconomics*, 37(11):1391–1408.

Eric Chamoun, Michael Schlichtkrull, and Andreas Vlachos. 2024. Automated focused feedback generation for scientific writing assistance. In *Findings of the Association for Computational Linguistics: ACL 2024*, pages 9742–9763, Bangkok, Thailand. Association for Computational Linguistics.

Yuan Chang, Ziyue Li, Hengyuan Zhang, Yuanbo Kong, Yanru Wu, Zhijiang Guo, and Ngai Wong. 2025. Treereview: A dynamic tree of questions framework for deep and efficient llm-based scientific peer review.

A. Checco, L. Bracciale, P. Loreti, et al. 2021. Ai-assisted peer review. *Humanities and Social Sciences Communications*, 8:25.

Qiguang Chen, Mingda Yang, Libo Qin, Jinhao Liu, Zheng Yan, Jiannan Guan, Dengyun Peng, Yiyan Ji, Hanjing Li, Mengkang Hu, Yimeng Zhang, Yihao Liang, Yuhang Zhou, Jiaqi Wang, Zhi Chen, and Wanxiang Che. 2025. Ai4research: A survey of artificial intelligence for scientific research.

Liying Cheng, Lidong Bing, Qian Yu, Wei Lu, and Luo Si. 2020. APE: Argument pair extraction from peer review and rebuttal via multi-task learning. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 7000–7011, Online. Association for Computational Linguistics.

Wei-Lin Chiang, Lianmin Zheng, Ying Sheng, Anastasios Nikolas Angelopoulos, Tianle Li, Dacheng Li, Hao Zhang, Banghua Zhu, Michael Jordan, Joseph E. Gonzalez, and Ion Stoica. 2024. Chatbot arena: An open platform for evaluating llms by human preference.

Maitreya Prafulla Chitale, Ketaki Mangesh Shetye, Harshit Gupta, Manav Chaudhary, and Vasudeva Varma. 2025. Autorev: Automatic peer review system for academic research papers.

Gautam Choudhary, Natwar Modani, and Nitish Maurya. 2022. React: A review comment dataset for actionability (and more).

Paul Christiano, Jan Leike, Tom B. Brown, Miljan Martic, Shane Legg, and Dario Amodei. 2023. Deep reinforcement learning from human preferences.

Magic-Doc Contributors. 2024. Magic-doc: A toolkit that converts multiple file types to markdown. https://github.com/InternLM/magic-doc.

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

Mike D'Arcy, Alexis Ross, Erin Bransom, Bailey Kuehl, Jonathan Bragg, Tom Hope, and Doug Downey. 2023. ARIES: A corpus of scientific paper edits made in response to peer reviews.

Iddo Drori and Dov Te'eni. 2024. Human-in-the-loop ai reviewing: Feasibility, opportunities, and risks. *Journal of the Association for Information Systems*, 25(1):98–109.

J. A. Drozdz and M. R. Ladomery. 2024. The peer review process: Past, present, and future. *British Journal of Biomedical Science*, 81:12054.

Jiangshu Du, Yibo Wang, Wenting Zhao, Zhongfen Deng, Shuaiqi Liu, Renze Lou, Henry Peng Zou, Pranav Narayanan Venkit, Nan Zhang, Mukund Srinath, Ranran Haoran Zhang, Vipul Gupta, Yinghui Li, Tao Li, Fei Wang, Qin Liu, Tianlin Liu, Pengzhi Gao, Congying Xia, Chen Xing, Jiayang Cheng, Zhaowei Wang, Ying Su, Raj Sanjay Shah, Ruohao Guo, Jing Gu, Haoran Li, Kangda Wei, Zihao Wang, Lu Cheng, Surangika Ranathunga, Meng Fang, Jie Fu, Fei Liu, Ruihong Huang, Eduardo Blanco, Yixin Cao, Rui Zhang, Philip S. Yu, and Wenpeng Yin. 2024. Llms assist nlp researchers: Critique paper (meta-)reviewing. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 5081–5099.

Zane Durante, Qiuyuan Huang, Naoki Wake, Ran Gong, Jae Sung Park, Bidipta Sarkar, Rohan Taori, Yusuke Noda, Demetri Terzopoulos, Yejin Choi, Katsushi Ikeuchi, Hoi Vo, Li Fei-Fei, and Jianfeng Gao. 2024. Agent ai: Surveying the horizons of multimodal interaction.

Nils Dycke, Ilia Kuznetsov, and Iryna Gurevych. 2022. Yes-yes-yes: Proactive data collection for acl rolling review and beyond.

Nils Dycke, Ilia Kuznetsov, and Iryna Gurevych. 2023a. Nlpeer: A unified resource for the computational study of peer review.

Nils Dycke, Ilia Kuznetsov, and Iryna Gurevych. 2023b. Overview of PragTag-2023: Low-resource multi-domain pragmatic tagging of peer reviews. In *Proceedings of the 10th Workshop on Argument Mining*, pages 187–196, Singapore. Association for Computational Linguistics.

Nils Dycke, Matej Zečević, Ilia Kuznetsov, Beatrix Suess, Kristian Kersting, and Iryna Gurevych. 2025. STRICTA: Structured reasoning in critical text assessment for peer review and beyond. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 22687–22727, Vienna, Austria. Association for Computational Linguistics.

Alexander R. Fabbri, Irene Li, Tianwei She, Suyi Li, and Dragomir R. Radev. 2019. Multi-news: a large-scale multi-document summarization dataset and abstractive hierarchical model.

Abdur Rahman Bin Md Faizullah, Ashok Urlana, and Rahul Mishra. 2024. Limgen: Probing the llms for generating suggestive limitations of research papers.

Kavita Ganesan. 2018. Rouge 2.0: Updated and improved measures for evaluation of summarization tasks.

Xian Gao, Jiacheng Ruan, Zongyun Zhang, Jingsheng Gao, Ting Liu, and Yuzhuo Fu. 2025. Reviewagents: Bridging the gap between human and ai-generated paper reviews.

Yang Gao, Steffen Eger, Ilia Kuznetsov, Iryna Gurevych, and Yusuke Miyao. 2019. Does my rebuttal matter? insights from a major NLP conference. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 1274–1290, Minneapolis, Minnesota. Association for Computational Linguistics.

Yunfan Gao, Yun Xiong, Xinyu Gao, Kangxiang Jia, Jinliu Pan, Yuxi Bi, Yi Dai, Jiawei Sun, Meng Wang, and Haofen Wang. 2024a. Retrieval-augmented generation for large language models: A survey.

Zhaolin Gao, Kianté Brantley, and Thorsten Joachims. 2024b. Reviewer2: Optimizing review generation through prompt generation.

Zhaolin Gao, Kianté Brantley, and Thorsten Joachims. 2024c. Reviewer2: Optimizing review generation through prompt generation.

Madhav Krishan Garg, Chhavi Kirtani, Tejash Prasad, Tanmay Singhal, Murari Mandal, and Dhruv Kumar. 2024. Revieweval: An evaluation framework for ai-generated reviews. arXiv preprint arXiv:2502.11736.

Zhibin Gou, Zhihong Shao, Yeyun Gong, Yelong Shen, Yujiu Yang, Nan Duan, and Weizhu Chen. 2024. Critic: Large language models can self-correct with tool-interactive critiquing.

Jiawei Gu, Xuhui Jiang, Zhichao Shi, Hexiang Tan, Xuehao Zhai, Chengjin Xu, Wei Li, Yinghan Shen, Shengjie Ma, Honghao Liu, Saizhuo Wang, Kun Zhang, Yuanzhuo Wang, Wen Gao, Lionel Ni, and Jian Guo. 2025. A survey on llm-as-a-judge.

Daya Guo, Dejian Yang, Haowei Zhang, Junxiao Song, Ruoyu Zhang, Runxin Xu, Qihao Zhu, Shirong Ma, Peiyi Wang, Xiao Bi, et al. 2024. Deepseek-r1: Incentivizing reasoning capability in llms via reinforcement learning.

Yanzhu Guo, Guokan Shang, Virgile Rennard, Michalis Vazirgiannis, and Chloé Clavel. 2023. Automatic analysis of substantiation in scientific peer reviews. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 10198–10216.

Haixia Han, Jiaqing Liang, Jie Shi, Qianyu He, and Yanghua Xiao. 2024. Small language model can self-correct.

Peixuan Han, Yingjie Yu, Jingjun Xu, and Jiaxuan You. 2026. Drpg (decompose, retrieve, plan, generate): An agentic framework for academic rebuttal.

Kilian Hendrickx, Lorenzo Perini, Dries Van der Plas, Wannes Meert, and Jesse Davis. 2021. Machine learning with a reject option: A survey. *IEEE Transactions on Knowledge and Data Engineering*, 33(1):43–54.

M. Hosseini and S. P. J. M. Horbach. 2023. Fighting reviewer fatigue or amplifying bias? considerations and recommendations for use of chatgpt and other large language models in scholarly peer review. *Research Integrity and Peer Review*, 8(1):4.

Edward J. Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. 2021. Lora: Low-rank adaptation of large language models.

Junjie Huang, Win-bin Huang, Yi Bu, Qi Cao, Huawei Shen, and Xueqi Cheng. 2023. What makes a successful rebuttal in computer science conferences?: A perspective on social interaction. *Journal of Informetrics*, 17(3):101427.

Maximilian Idahl and Zahra Ahmadi. 2025. Openreviewer: A specialized large language model for generating critical scientific paper reviews.

Jinbae Im, Moonki Kim, Hoyeop Lee, Hyunsouk Cho, and Sehee Chung. 2021. Self-supervised multimodal opinion summarization. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 388–403, Online. Association for Computational Linguistics.

Chao Jiang, Wei Xu, and Samuel Stevens. 2022. arxivedits: Understanding the human revision process in scientific writing.

Feng Jiang, Kuang Wang, and Haizhou Li. 2024. Bridging research and readers: A multi-modal automated academic papers interpretation system.

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

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

Leane Jourdan, Florian Boudin, Nicolas Hernandez, and Richard Dufour. 2024. Casimir: A corpus of scientific articles enhanced with multiple author-integrated revisions.

Ryo Kamoi, Yusen Zhang, Nan Zhang, Jiawei Han, and Rui Zhang. 2024. When can LLMs actually correct their own mistakes? a critical survey of self-correction of LLMs. *Transactions of the Association for Computational Linguistics*, 12:1417–1440.

Dongyeop Kang, Waleed Ammar, Bhavana Dalvi, Madeleine van Zuylen, Sebastian Kohlmeier, Eduard Hovy, and Roy Schwartz. 2018. 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, New Orleans, Louisiana. Association for Computational Linguistics.

Neha Kennard, Tim O'Gorman, Rajarshi Das, Akshay Sharma, Chhandak Bagchi, Matthew Clinton, Pranay Kumar Yelugam, Hamed Zamani, and Andrew McCallum. 2021. Disapere: A dataset for discourse structure in peer review discussions.

Jyotsana Khatri and Manasi Patwardhan. 2026. Defend: Automated rebuttals for peer review with minimal author guidance.

Rodney Michael Kinney, Chloe Anastasiades, Russell Authur, Iz Beltagy, Jonathan Bragg, Alexandra Buraczynski, Isabel Cachola, Stefan Candra, Yoganand Chandrasekhar, Arman Cohan, Miles Crawford, Doug Downey, Jason Dunkelberger, Oren Etzioni, Rob Evans, Sergey Feldman, Joseph Gorney, David W. Graham, F.Q. Hu, Regan Huff, Daniel King, Sebastian Kohlmeier, Bailey Kuehl, Michael Langan, Daniel Lin, Haokun Liu, Kyle Lo, Jaron Lochner, Kelsey MacMillan, Tyler C. Murray, Christopher Newell, Smita R Rao, Shaurya Rohatgi, Paul Sayre, Zejiang Shen, Amanpreet Singh, Luca Soldaini, Shivashankar Subramanian, A. Tanaka, Alex D Wade, Linda M. Wagner, Lucy Lu Wang, Christopher Wilhelm, Caroline Wu, Jiangjiang Yang, Angele Zamarron, Madeleine van Zuylen, and Daniel S. Weld. 2023. The semantic scholar open data platform. *ArXiv*, abs/2301.10140.

Barbara Kitchenham and Stuart Charters. 2007. Guidelines for performing systematic literature reviews in software engineering technical report. *Software Engineering Group, EBSE Technical Report, Keele University and Department of Computer Science University of Durham*, 2.

Asheesh Kumar, Amrinder Arora, Patrick Derieux, and Sandeep Purao. 2024a. Towards automated meta-review generation via an nlp/ml pipeline in different stages of the scholarly peer review process. *International Journal on Digital Libraries*, 25(3):295–314.

Asheesh Kumar, Tirthankar Ghosal, and Asif Ekbal. 2024b. A deep neural architecture for decision-aware meta-review generation. In *Proceedings of the 2021 ACM/IEEE Joint Conference on Digital Libraries*, JCDL '21, page 222–225. IEEE Press.

Sandeep Kumar, Guneet Singh Kohli, Tirthankar Ghosal, and Asif Ekbal. 2024c. Longform multimodal lay summarization of scientific papers: Towards automatically generating science blogs from research articles. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 10790–10801, Torino, Italia. ELRA and ICCL.

Mayte H Laureano, Hiram Calvo, Tania Alcantara, Omar Garcıa Vazquez, and Marco A Cardoso Moreno. 2025. Predicting Reviewers' Decisions in Scientific Submissions through Linguistic Analysis. *Journal of Scientometric Research*, 14(1):331–341.

Carole J Lee, Cassidy R Sugimoto, Guo Zhang, and Blaise Cronin. 2013. Bias in peer review. *Journal of the American Society for Information Science and Technology*, 64(1):2–17.

J Lee, J Lee, and J. J. Yoo. 2025. The role of large language models in the peer-review process: opportunities and challenges for medical journal reviewers and editors. *Journal of Educational Evaluation for Health Professions*, 22:4. Epub 2025 Jan 16.

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

Miao Li, Eduard Hovy, and Jey Lau. 2023. Summarizing multiple documents with conversational structure for meta-review generation. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 7089–7112, Singapore. Association for Computational Linguistics.

Miao Li, Jey Han Lau, and Eduard Hovy. 2024. 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, Bangkok, Thailand. Association for Computational Linguistics.

Miao Li, Jianzhong Qi, and Jey Han Lau. 2022. Peersum: A peer review dataset for abstractive multi-document summarization.

Pan Li and Alexander Tuzhilin. 2020. Towards controllable and personalized review generation.

Wei Li and Hai Zhuge. 2021. Abstractive multi-document summarization based on semantic link network. *IEEE Transactions on Knowledge and Data Engineering*, 33(1):43–54.

Yinheng Li. 2023. A practical survey on zero-shot prompt design for in-context learning. In *Proceedings of the Conference Recent Advances in Natural Language Processing - Large Language Models for Natural Language Processings*, RANLP, page 641–647. INCOMA Ltd., Shoumen, BULGARIA.

Weixin Liang, Zachary Izzo, Yaohui Zhang, Haley Lepp, Hancheng Cao, Xuandong Zhao, Lingjiao Chen, Haotian Ye, Sheng Liu, Zhi Huang, Daniel A. McFarland, and James Y. Zou. 2024. Monitoring ai-modified content at scale: A case study on the impact of chatgpt on ai conference peer reviews.

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

Zeyi Liao, Jaylen Jones, Linxi Jiang, Eric Fosler-Lussier, Yu Su, Zhiqiang Lin, and Huan Sun. 2025. Redteamcua: Realistic adversarial testing of computer-use agents in hybrid web-os environments.

Chin-Yew Lin. 2004. ROUGE: A package for automatic evaluation of summaries. In *Text Summarization Branches Out*, pages 74–81, Barcelona, Spain. Association for Computational Linguistics.

Ethan Lin, Zhiyuan Peng, and Yi Fang. 2024. Evaluating and enhancing large language models for novelty assessment in scholarly publications.

Jialiang Lin, Jiaxin Song, Zhangping Zhou, Yidong Chen, and Xiaodong Shi. 2022. Moprd: A multidisciplinary open peer review dataset. arXiv preprint arXiv:2212.04972.

Yen-Ting Lin and Yun-Nung Chen. 2023. Llm-eval: Unified multi-dimensional automatic evaluation for open-domain conversations with large language models.

Ryan Liu and Nihar B. Shah. 2023. Reviewergpt? an exploratory study on using large language models for paper reviewing.

Sheng Lu, Ilia Kuznetsov, and Iryna Gurevych. 2025. Identifying aspects in peer reviews.

Qianli Ma, Chang Guo, Zhiheng Tian, Siyu Wang, Jipeng Xiao, Yuanhao Yue, and Zhipeng Zhang. 2026. Paper2rebuttal: A multi-agent framework for transparent author response assistance.

Yu Meng, Mengzhou Xia, and Danqi Chen. 2024. Simpo: Simple preference optimization with a reference-free reward.

Panitan Muangkammuen, Fumiyo Fukumoto, Jiyi Li, and Yoshimi Suzuki. 2022. Exploiting labeled and unlabeled data via transformer fine-tuning for peer-review score prediction. In *Findings of the Association for Computational Linguistics: EMNLP 2022*, pages 2233–2240.

Jimin Mun, Chani Jung, Xuhui Zhou, Hyunwoo Kim, and Maarten Sap. 2026. Goodpoint: Learning constructive scientific paper feedback from author responses.

Meike Nauta, Jan Trienes, Shreyasi Pathak, Elisa Nguyen, Michelle Peters, Yasmin Schmitt, Jörg Schlötterer, Maurice van Keulen, and Christin Seifert. 2023. From anecdotal evidence to quantitative evaluation methods: A systematic review on evaluating explainable ai. *ACM Computing Surveys*, 55(13s):1–42.

Kun-Peng Ning, Shuo Yang, Yu-Yang Liu, Jia-Yu Yao, Zhen-Hui Liu, Yong-Hong Tian, Yibing Song, and Li Yuan. 2025. Pico: Peer review in llms based on the consistency optimization.

Alexander Novikov, Ngân Vũ, Marvin Eisenberger, Emilien Dupont, Po-Sen Huang, Adam Zsolt Wagner, Sergey Shirobokov, Borislav Kozlovskii, Francisco J. R. Ruiz, Abbas Mehrabian, M. Pawan Kumar, Abigail See, Swarat Chaudhuri, George Holland, Alex Davies, Sebastian Nowozin, Pushmeet Kohli, and Matej Balog. 2025. Alphaevolve: A coding agent for scientific and algorithmic discovery.

OpenAI, Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, Red Avila, Igor Babuschkin, Suchir Balaji, Valerie Balcom, Paul Baltescu, Haiming Bao, Mohammad Bavarian, Jeff Belgum, Irwan Bello, Jake Berdine, Gabriel Bernadett-Shapiro, Christopher Berner, Lenny Bogdonoff, Oleg Boiko, Madelaine Boyd, Anna-Luisa Brakman, Greg Brockman, Tim Brooks, Miles Brundage, Kevin Button, Trevor Cai, Rosie Campbell, Andrew Cann, Brittany Carey, Chelsea Carlson, Rory Carmichael, Brooke Chan, Che Chang, Fotis Chantzis, Derek Chen, Sully Chen, Ruby Chen, Jason Chen, Mark Chen, Ben Chess, Chester Cho, Casey Chu, Hyung Won Chung, Dave Cummings, Jeremiah Currier, Yunxing Dai, Cory Decareaux, Thomas Degry, Noah Deutsch, Damien Deville, Arka Dhar, David Dohan, Steve Dowling, Sheila Dunning, Adrien Ecoffet, Atty Eleti, Tyna Eloundou, David Farhi, Liam Fedus, Niko Felix, Simón Posada Fishman, Juston Forte, Isabella Fulford, Leo Gao, Elie Georges, Christian Gibson, Vik Goel, Tarun Gogineni, Gabriel Goh, Rapha Gontijo-Lopes, Jonathan Gordon, Morgan Grafstein, Scott Gray, Ryan Greene, Joshua Gross, Shixiang Shane Gu, Yufei Guo, Chris Hallacy, Jesse Han, Jeff Harris, Yuchen He, Mike Heaton, Johannes Heidecke, Chris Hesse, Alan Hickey, Wade Hickey, Peter Hoeschele, Brandon Houghton, Kenny Hsu, Shengli Hu, Xin Hu, Joost Huizinga, Shantanu Jain, Shawn Jain, Joanne Jang, Angela Jiang, Roger Jiang, Haozhun Jin, Denny Jin, Shino Jomoto, Billie Jonn, Heewoo Jun, Tomer Kaftan, Łukasz Kaiser, Ali Kamali, Ingmar Kanitscheider, Nitish Shirish Keskar, Tabarak Khan, Logan Kilpatrick, Jong Wook Kim, Christina Kim, Yongjik Kim, Jan Hendrik Kirchner, Jamie Kiros, Matt Knight, Daniel Kokotajlo, Łukasz Kondraciuk, Andrew Kondrich, Aris Konstantinidis, Kyle Kosic, Gretchen Krueger, Vishal Kuo, Michael Lampe, Ikai Lan, Teddy Lee, Jan Leike, Jade Leung, Daniel Levy, Chak Ming Li, Rachel Lim, Molly Lin, Stephanie Lin, Mateusz Litwin, Theresa Lopez, Ryan Lowe, Patricia Lue, Anna Makanju, Kim Malfacini, Sam Manning, Todor Markov, Yaniv Markovski, Bianca Martin, Katie Mayer, Andrew Mayne, Bob McGrew, Scott Mayer McKinney, Christine McLeavey, Paul McMillan, Jake McNeil, David Medina, Aalok Mehta, Jacob Menick, Luke Metz, Andrey Mishchenko, Pamela Mishkin, Vinnie Monaco, Evan Morikawa, Daniel Mossing, Tong Mu, Mira Murati, Oleg Murk, David Mély, Ashvin Nair, Reiichiro Nakano, Rajeev Nayak, Arvind Neelakantan, Richard Ngo, Hyeonwoo Noh, Long Ouyang, Cullen O'Keefe, Jakub Pachocki, Alex Paino, Joe Palermo, Ashley Pantuliano, Giambattista Parascandolo, Joel Parish, Emy Parparita, Alex Passos, Mikhail Pavlov, Andrew Peng, Adam Perelman, Filipe de Avila Belbute Peres, Michael Petrov, Henrique Ponde de Oliveira Pinto, Michael, Pokorny, Michelle Pokrass, Vitchyr H. Pong, Tolly Powell, Alethea Power, Boris Power, Elizabeth Proehl, Raul Puri, Alec Radford, Jack Rae, Aditya Ramesh, Cameron Raymond, Francis Real, Kendra Rimbach, Carl Ross, Bob Rotsted, Henri Roussez, Nick Ryder, Mario Saltarelli, Ted Sanders, Shibani Santurkar, Girish Sastry, Heather Schmidt, David Schnurr, John Schulman, Daniel Selsam, Kyla Sheppard, Toki Sherbakov, Jessica Shieh, Sarah Shoker, Pranav Shyam, Szymon Sidor, Eric Sigler, Maddie Simens, Jordan Sitkin, Katarina Slama, Ian Sohl, Benjamin Sokolowsky, Yang Song, Natalie Staudacher, Felipe Petroski Such, Natalie Summers, Ilya Sutskever, Jie Tang, Nikolas Tezak, Madeleine B. Thompson, Phil Tillet, Amin Tootoonchian, Elizabeth Tseng, Preston Tuggle, Nick Turley, Jerry Tworek, Juan Felipe Cerón Uribe, Andrea Vallone, Arun Vijayvergiya, Chelsea Voss, Carroll Wainwright, Justin Jay Wang, Alvin Wang, Ben Wang, Jonathan Ward, Jason Wei, CJ Weinmann, Akila Welihinda, Peter Welinder, Jiayi Weng, Lilian Weng, Matt Wiethoff, Dave Willner, Clemens Winter, Samuel Wolrich, Hannah Wong, Lauren Workman, Sherwin Wu, Jeff Wu, Michael Wu, Kai Xiao, Tao Xu, Sarah Yoo, Kevin Yu, Qiming Yuan, Wojciech Zaremba, Rowan Zellers, Chong Zhang, Marvin Zhang, Shengjia Zhao, Tianhao Zheng, Juntang Zhuang, William Zhuk, and Barret Zoph. 2024. Gpt-4 technical report.

Matthew J Page, Joanne E McKenzie, Patrick M Bossuyt, Isabelle Boutron, Tammy C Hoffmann, Cynthia D Mulrow, Larissa Shamseer, Jennifer M

Tetzlaff, Elie A Akl, Sue E Brennan, Roger Chou, Julie Glanville, Jeremy M Grimshaw, Asbjørn Hróbjartsson, Manoj M Lalu, Tianjing Li, Elizabeth W Loder, Evan Mayo-Wilson, Steve McDonald, Luke A McGuinness, Lesley A Stewart, James Thomas, Andrea C Tricco, Vivian A Welch, Penny Whiting, and David Moher. 2021. The prisma 2020 statement: an updated guideline for reporting systematic reviews. *BMJ*, 372.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th Annual Meeting on Association for Computational Linguistics*, ACL '02, page 311–318, USA. Association for Computational Linguistics.

Sukannya Purkayastha, Nils Dycke, Anne Lauscher, and Iryna Gurevych. 2026. Decision-making with deliberation: Meta-reviewing as a document-grounded dialogue.

Sukannya Purkayastha, Anne Lauscher, and Iryna Gurevych. 2023a. Exploring jiu-jitsu argumentation for writing peer review rebuttals. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 14479–14495, Singapore. Association for Computational Linguistics.

Sukannya Purkayastha, Anne Lauscher, and Iryna Gurevych. 2023b. Exploring jiu-jitsu argumentation for writing peer review rebuttals.

Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, and Ilya Sutskever. 2019. Language models are unsupervised multitask learners. OpenAI Technical Report. https://cdn.openai.com/better-language-models/language_models_are_unsupervised_multitask_learners.pdf.

Kanishka Rao, Chris Harris, Alex Irpan, Sergey Levine, Julian Ibarz, and Mohi Khansari. 2020. Rl-cyclegan: Reinforcement learning aware simulation-to-real.

Clément Reverdy, Sylvie Gibet, and Thibaut Le Naour. 2024. STK LSF: A motion capture dataset in LSF for SignToKids. In *Proceedings of the LREC-COLING 2024 11th Workshop on the Representation and Processing of Sign Languages: Evaluation of Sign Language Resources*, pages 315–322, Torino, Italia. ELRA and ICCL.

Zachary Robertson. 2023. Gpt4 is slightly helpful for peer-review assistance: A pilot study.

Qian Ruan and Iryna Gurevych. 2026. Author-in-the-loop response generation and evaluation: Integrating author expertise and intent in responses to peer review.

Sajratul Y. Rubaiat, Syed N. Sakib, and Hasan M. Jamil. 2025. Mapping the evolution of research contributions using knovo.

Abdelrahman Sadallah, Tim Baumgärtner, Iryna Gurevych, and Ted Briscoe. 2025. The good, the bad and the constructive: Automatically measuring peer review's utility for authors.

Laurie A. Schintler, Connie L. McNeely, and James Witte. 2023. A critical examination of the ethics of ai-mediated peer review.

Zhihong Shao, Peiyi Wang, Qihao Zhu, Runxin Xu, Junxiao Song, Xiao Bi, Haowei Zhang, Mingchuan Zhang, Y. K. Li, Y. Wu, and Daya Guo. 2024. Deepseekmath: Pushing the limits of mathematical reasoning in open language models.

Chenhui Shen, Liying Cheng, Ran Zhou, Lidong Bing, Yang You, and Luo Si. 2021. MReD: A meta-review dataset for structure-controllable text generation.

Chenhui Shen, Liying Cheng, Ran Zhou, Lidong Bing, Yang You, and Luo Si. 2022. MReD: A meta-review dataset for structure-controllable text generation. In *Findings of the Association for Computational Linguistics: ACL 2022*, pages 2521–2535, Dublin, Ireland. Association for Computational Linguistics.

Hyungyu Shin, Jingyu Tang, Yoonjoo Lee, Nayoung Kim, Hyunseung Lim, Ji Yong Cho, Hwajung Hong, Moontae Lee, and Juho Kim. 2025. Mind the blind spots: A focus-level evaluation framework for llm reviews.

Kurt Shuster, Spencer Poff, Moya Chen, Douwe Kiela, and Jason Weston. 2021. Retrieval augmentation reduces hallucination in conversation.

Sonish Sivarajkumar, Mark Kelley, Alyssa Samolyk-Mazzanti, Shyam Visweswaran, and Yanshan Wang. 2023. An empirical evaluation of prompting strategies for large language models in zero-shot clinical natural language processing.

Purin Sukpanichnant, Anna Rapberger, and Francesca Toni. 2024. Peerarg: Argumentative peer review with llms. arXiv preprint arXiv:2409.16813.

Gokul Swamy, Sanjiban Choudhury, Wen Sun, Zhiwei Steven Wu, and J. Andrew Bagnell. 2025. All roads lead to likelihood: The value of reinforcement learning in fine-tuning.

Pawin Taechoyotin and Daniel Acuna. 2025. Remor: Automated peer review generation with llm reasoning and multi-objective reinforcement learning.

Pawin Taechoyotin and Daniel E. Acuna. 2026. Remctx: Automated peer review via reinforcement learning with auxiliary context.

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

Cheng Tan, Dongxin Lyu, Siyuan Li, Zhangyang Gao, Jingxuan Wei, Siqi Ma, Zicheng Liu, and Stan Z. Li. 2024. Peer review as a multi-turn and long-context dialogue with role-based interactions.

Jiabin Tang, Lianghao Xia, Zhonghang Li, and Chao Huang. 2025. Ai-researcher: Autonomous scientific innovation.

Qwen Team. 2024. Qwen2.5: A party of foundation models.

Simone Teufel, Jean Carletta, and Marc Moens. 1999. An annotation scheme for discourse-level argumentation in research articles. In *Ninth Conference of the European Chapter of the Association for Computational Linguistics*, pages 110–117, Bergen, Norway. Association for Computational Linguistics.

Mike Thelwall and Abdullah Yaghi. 2024. Evaluating the predictive capacity of chatgpt for academic peer review outcomes across multiple platforms.

Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Jing Fu, Thomas Wolf, et al. 2023. Llama: Open and efficient foundation language models.

Keith Tyser, Ben Segev, Gaston Longhitano, Xin-Yu Zhang, Zachary Meeks, Jason Lee, Uday Garg, Nicholas Belsten, Avi Shporer, Madeleine Udell, Dov Te'eni, and Iddo Drori. 2024. Ai-driven review systems: Evaluating llms in scalable and bias-aware academic reviews.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In *Advances in Neural Information Processing Systems*, pages 5998–6008.

D. von Wedel, R. A. Schmitt, M. Thiele, R. Leuner, D. Shay, S. Redaelli, and M. S. Schaefer. 2024. Affiliation bias in peer review of abstracts by a large language model. *JAMA*, 331(3):252–253.

Bin Wang, Chao Xu, Xiaomeng Zhao, Linke Ouyang, Fan Wu, Zhiyuan Zhao, Rui Xu, Kaiwen Liu, Yuan Qu, Fukai Shang, Bo Zhang, Liqun Wei, Zhihao Sui, Wei Li, Botian Shi, Yu Qiao, Dahua Lin, and Conghui He. 2024. Mineru: An open-source solution for precise document content extraction.

Chengye Wang, Yifei Shen, Zexi Kuang, Arman Cohan, and Yilun Zhao. 2025. Sciver: Evaluating foundation models for multimodal scientific claim verification.

Qingyun Wang, Qi Zeng, Lifu Huang, Kevin Knight, Heng Ji, and Nazneen Fatema Rajani. 2020. Reviewrobot: Explainable paper review generation based on knowledge synthesis. In *Proceedings of the 13th International Conference on Natural Language Generation*. Association for Computational Linguistics.

Qiyao Wei, Samuel Holt, Jing Yang, Markus Wulfmeier, and Mihaela van der Schaar. 2025. The ai imperative: Scaling high-quality peer review in machine learning.

Yixuan Weng, Minjun Zhu, Guangsheng Bao, Hongbo Zhang, Jindong Wang, Yue Zhang, and Linyi Yang. 2025. Cycleresearcher: Improving automated research via automated review.

Po-Cheng Wu, An-Zi Yen, Hen-Hsen Huang, and Hsin-Hsi Chen. 2022. Incorporating peer reviews and rebuttal counter-arguments for meta-review generation. In *Proceedings of the 31st ACM International Conference on Information & Knowledge Management (CIKM 2022)*, pages 2189–2198. Association for Computing Machinery. Introduces the *PRRCA* dataset.

Sihong Wu, Yiling Ma, Yilun Zhao, Tiansheng Hu, Owen Jiang, Manasi Patwardhan, and Arman Cohan. 2026a. Rbtact: Rebuttal as supervision for actionable review feedback generation.

Wenqing Wu, Yi Zhao, Yuzhuo Wang, Siyou Li, Juexi Shao, Yunfei Long, and Chengzhi Zhang. 2026b. Novbench: Evaluating large language models on academic paper novelty assessment.

Silei Xu, Wenhao Xie, Lingxiao Zhao, and Pengcheng He. 2025. Chain of draft: Thinking faster by writing less.

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

Jianxiang Yu, Zichen Ding, Jiaqi Tan, Kangyang Luo, Zhenmin Weng, Chenghua Gong, Long Zeng, RenJing Cui, Chengcheng Han, Qiushi Sun, Zhiyong Wu, Yunshi Lan, and Xiang Li. 2024. Automated peer reviewing in paper sea: Standardization, evaluation, and analysis.

Weizhe Yuan, Pengfei Liu, and Graham Neubig. 2021. Can we automate scientific reviewing? In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 6269–6284.

Weizhe Yuan, Pengfei Liu, and Graham Neubig. 2022. Can We Automate Scientific Reviewing? *Journal of Artificial Intelligence Research*, 75:171–212. Introduces the *ASAP-Review* dataset.

Qi Zeng, Meng Jiang, Shasha Li, and Zhou Yu. 2023. Meta-review generation with checklist-guided iterative introspection. arXiv preprint arXiv:2305.14647.

Qi Zeng, Mankeerat Sidhu, Ansel Blume, Hou Pong Chan, Lu Wang, and Heng Ji. 2024. Scientific opinion summarization: Paper meta-review generation dataset, methods, and evaluation.

Sihang Zeng, Kai Tian, Kaiyan Zhang, Yuru wang, Junqi Gao, Runze Liu, Sa Yang, Jingxuan Li, Xinwei Long, Jiaheng Ma, Biqing Qi, and Bowen Zhou. 2025. Reviewrl: Towards automated scientific review with rl.

Daoze Zhang, Zhijian Bao, Sihang Du, Zhiyi Zhao, Kuangling Zhang, Dezheng Bao, and Yang Yang. 2025a. Re^2: A consistency-ensured dataset for full-stage peer review and multi-turn rebuttal discussions.

Jiayao Zhang, Hongming Zhang, Zhun Deng, and Dan Roth. 2022. Investigating fairness disparities in peer review: A language model enhanced approach. Introduces the *ICLR-DB* dataset (10k submissions, 36k reviews, 68k responses from ICLR 2017–2022).

Lili Zhang, Haomiaomiao Wang, Long Cheng, Libao Deng, and Tomas Ward. 2025b. Adversarial testing in llms: Insights into decision-making vulnerabilities.

Tianmai M. Zhang and Neil F. Abernethy. 2025. Reviewing scientific papers for critical problems with reasoning llms: Baseline approaches and automatic evaluation.

Tianyi Zhang, Varsha Kishore, Felix Wu, Kilian Q. Weinberger, and Yoav Artzi. 2020. Bertscore: Evaluating text generation with bert.

Wayne Xin Zhao, Kun Zhou, Junlin Yao, Zhicheng Dou, and Ji-Rong Wen. 2023. A survey of large language models.

Yilun Zhao, Kaiyan Zhang, Tiansheng Hu, Sihong Wu, Ronan Le Bras, Taira Anderson, Jonathan Bragg, Joseph Chee Chang, Jesse Dodge, Matt Latzke, Yixin Liu, Charles McGrady, Xiangru Tang, Zihang Wang, Chen Zhao, Hannaneh Hajishirzi, Doug Downey, and Arman Cohan. 2025. Sciarena: An open evaluation platform for foundation models in scientific literature tasks.

Ruiyang Zhou, Lu Chen, and Kai Yu. 2024a. 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, Torino, Italia. ELRA and ICCL.

Ruiyang Zhou, Lu Chen, and Kai Yu. 2024b. 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)*. Introduces the *RR-MCQ* dataset of 197 review–revision multiple-choice questions.

Minjun Zhu, Yixuan Weng, Linyi Yang, and Yue Zhang. 2025a. Deepreview: Improving llm-based paper review with human-like deep thinking process.

Minjun Zhu, Yixuan Weng, Linyi Yang, and Yue Zhang. 2025b. Deepreview: Improving LLMbased paper review with humanlike deep thinking process. Introduces the *DeepReview-13K* dataset.

Zhenzhen Zhuang, Jiandong Chen, Hongfeng Xu, Yuwen Jiang, and Jialiang Lin. 2025. Large language models for automated scholarly paper review: A survey. *Information Fusion*, 124:103332.

Daniel M. Ziegler, Nisan Stiennon, Jeffrey Wu, Tom B. Brown, Alec Radford, Dario Amodei, Paul Christiano, and Geoffrey Irving. 2020. Fine-tuning language models from human preferences.

Zhuoyang Zou, Abolfazl Ansari, Delvin Ce Zhang, Dongwon Lee, and Wenpeng Yin. 2026. Diagpaper: Diagnosing valid and specific weaknesses in scientific papers via multi-agent reasoning.

Ádám Kovács and Gábor Recski. 2025. Lettucedetect: A hallucination detection framework for rag applications.

A Evaluation Metrics

Evaluating the quality of generated peer reviews is a multifaceted and non-trivial task. The field has developed a diverse array of evaluation metrics, which have evolved from standard text-similarity measures to more sophisticated, multi-dimensional frameworks that aim to capture the functional utility of a review. These metrics can be broadly categorized into five facets: reference-based text similarity, set overlap, aspect- and focus-level analysis, task-based proxies, and unsupervised or reward-based methods.

A.1 Reference-based Text Similarity

Initial approaches to evaluating generated reviews, particularly those framing the task as a form of summarization, relied on standard metrics that measure the similarity between a generated review and a ground-truth human review. These metrics provide a scalable, automated way to assess surface-level quality.

BLEU (Bilingual Evaluation Understudy): BLEU is a widely used metric for evaluating the quality of generated text by comparing it to one or more reference texts. It computes the degree of n-gram overlap, prioritizing precision (the proportion of n-grams in the candidate that appear in the reference) and incorporates a brevity penalty to discourage overly short outputs. The BLEU score is calculated as:

$$\text{BLEU} = \text{BP} \cdot \exp\left(\sum_{n=1}^{N} w_n \log p_n\right)$$

where p_n denotes the modified n-gram precision for n-grams of size n, w_n are the weights (usually uniform), and BP is the brevity penalty, defined as 1 if $c > r$ or $\exp(1 - r/c)$ if $c \leq r$, with c and r being the lengths of the candidate and reference texts, respectively. BLEU scores range from 0 to 1, with higher values indicating greater similarity to the reference.

BERTScore: BERTScore evaluates the similarity between generated and reference texts using contextual embeddings from pre-trained language models such as BERT. Instead of relying on exact n-gram matches, it computes token-level similarity via cosine similarity in the embedding space, capturing semantic similarity. The score is calculated by aligning each token in the candidate with the most similar token in the reference and averaging the resulting similarities. BERTScore reports precision, recall, and F1 scores, offering a more nuanced assessment of meaning preservation than traditional lexical overlap metrics.

ROUGE (Recall-Oriented Understudy for Gisting Evaluation): ROUGE-L evaluates the correspondence between a generated sequence and a reference sequence by identifying their Longest Common Subsequence (LCS). This metric emphasizes the preservation of word order without requiring strict consecutive matches.

Recall is calculated as the fraction of the LCS length over the total length of the reference sentence:

$$R = \frac{\text{LCS(Generated, Reference)}}{|\text{Reference}|}$$

Precision is similarly defined, using the length of the generated sequence as the denominator:

$$P = \frac{\text{LCS(Generated, Reference)}}{|\text{Generated}|}$$

To balance both precision and recall, the F_1 score is computed as follows:

$$F_1 = \frac{(1 + \beta^2)\, P\, R}{\beta^2 P + R}$$

where β allows for adjustable weighting between recall and precision. ROUGE-L scores reflects the degree of sequence similarity: a score closer to 1 indicates a higher overlap between the candidate and reference sequences. ROUGE scores are typically reported as recall, precision, and F1.

METEOR (Metric for Evaluation of Translation with Explicit ORdering): METEOR evaluates machine-generated text by aligning it with reference texts at the word level, accounting for exact matches, stem matches, synonym matches, and paraphrase matches. The metric calculates unigram precision and recall, combines them into an F-score, and applies a fragmentation penalty to account for differences in word order. The METEOR score is given by:

$$\text{METEOR} = F_{\text{mean}} \cdot (1 - \text{Pen})$$

where F_{mean} is a weighted harmonic mean of unigram precision and recall, and Pen is a penalty based on the number of chunks (contiguous matched subsequences). METEOR is designed to better correlate with human judgments than simple n-gram-based metrics.

A.2 Set Overlap Metrics

Recognizing that a review is a collection of distinct ideas or comments, some evaluation frameworks have moved beyond holistic text similarity to measure the overlap of these discrete points. This approach treats the generated and reference reviews as sets of comments and assesses their intersection.

Recall: Recall measures the proportion of relevant items in the reference set that are successfully retrieved by the model. For review generation, it reflects the fraction of human-written comments that are also captured by the generated review. The formula is:

$$\text{Recall} = \frac{|\text{Matched}_{\text{Gen}\to\text{Ref}}|}{|\text{Reference}|}$$

where $|\text{Matched}_{\text{Gen}\to\text{Ref}}|$ is the number of reference comments matched by generated comments, and $|\text{Reference}|$ is the total number of human-written reference comments. Recall ranges from 0 to 1, with higher values indicating more comprehensive coverage of reference content.

Precision: Precision evaluates the proportion of generated items that are relevant, i.e., the fraction of generated comments that can be aligned with human reference comments. The formula is:

$$\text{Precision} = \frac{|\text{Matched}_{\text{Gen}\to\text{Ref}}|}{|\text{Generated}|}$$

where $|\text{Matched}_{\text{Gen}\to\text{Ref}}|$ is the number of generated comments matched to the reference, and $|\text{Generated}|$ is the total number of generated comments. Higher precision indicates that most generated comments are meaningful and consistent with human feedback.

Jaccard Index: The Jaccard index measures the similarity between the generated and reference sets by dividing the size of their intersection by the size of their union. In the review context, it quantifies how many unique comments are shared between the two sets. The formula is:

$$\text{Jaccard} = \frac{|\text{Matched}_{\text{Gen}\to\text{Ref}}|}{|\text{Reference} \cup \text{Generated}|}$$

Here, $|\text{Matched}_{\text{Gen}\to\text{Ref}}|$ is the number of matched comments, and $|\text{Reference} \cup \text{Generated}|$ is the total number of unique comments across both sets. The Jaccard index ranges from 0 to 1, with higher values indicating greater overlap.

Comment Matching (LLM-based): To determine whether a generated comment matches a reference comment, a large language model (e.g., GPT-4) may be employed to assess semantic relatedness and relative specificity. Only pairs with high relatedness and sufficient specificity are counted as matches, ensuring that low-utility or generic comments do not inflate metric scores. This LLM-based matching provides a more nuanced, semantically grounded evaluation than traditional surface-level overlap.

A.3 Aspect and Focus Level Metrics

A significant advancement in evaluation has been the shift towards more diagnostic, fine-grained analysis. Instead of producing a single quality score, these methods dissect reviews to understand what they are about and how they critique the paper.

F1 Score: The F1 score is a widely-used metric that balances precision and recall for classification tasks. In the context of aspect-based coding schemes for peer review generation, it evaluates the accuracy of a model in assigning the correct (Target, Aspect) label to each review point, compared to expert annotations. The F1 score is defined as the harmonic mean of precision (P) and recall (R):

$$F_1 = \frac{2PR}{P + R}$$

where precision is the proportion of correctly predicted labels among all predicted labels, and recall is the proportion of correctly predicted labels among all true labels. The F1 score ranges from 0 to 1, with higher values indicating better classification performance.

Kullback-Leibler (KL) Divergence: KL Divergence is a distributional metric that quantifies the difference between two probability distributions. In peer review evaluation, it is used to compare the focus distribution of LLM-generated reviews with that of human-written reviews—i.e., how attention is allocated across various targets and aspects. The KL Divergence from the human distribution P to the model distribution Q is given by:

$$D_{\text{KL}}(P \parallel Q) = \sum_i P(i) \log \frac{P(i)}{Q(i)}$$

where $P(i)$ and $Q(i)$ are the probabilities assigned to target-aspect pair i by the human and model distributions, respectively. Lower values of KL Divergence indicate greater alignment between model and human focus.

A.4 Task-Based Proxy Metrics

Another evaluation strategy is to measure a model's performance on tasks that are proxies for the abilities required to write a good review. Instead of assessing the generated text directly, these metrics evaluate the model's capacity for judgment and deep comprehension.

Score Prediction: This task assesses whether a model can predict the final outcome of the peer review process. This can involve a binary classification task of predicting paper acceptance or a regression task of predicting the numerical scores for specific criteria like 'originality' or 'impact'. Performance is measured using standard metrics like accuracy for classification or Mean Absolute Error (MAE) for regression.

Multiple-Choice Question Answering: To probe deeper comprehension, specialized datasets have been created. For example, the RR-MCQ dataset contains multiple-choice questions derived from real review-rebuttal interactions. Performance is measured by accuracy, with results indicating that while current models perform better than random.

Accuracy: Accuracy measures the proportion of correctly predicted instances among all instances. In peer review generation, it is used both for classification tasks (e.g., predicting paper acceptance) and for multiple-choice question answering. Its formula is:

$$\text{Accuracy} = \frac{\text{Number of Correct Predictions}}{\text{Total Number of Predictions}}$$

Accuracy values range from 0 to 1 (or 0% to 100%), with higher values indicating better predictive performance.

Mean Absolute Error (MAE): MAE evaluates the average magnitude of errors in a set of predictions, without considering their direction. In peer review evaluation, it is typically used for regression tasks, such as predicting numerical review scores (e.g., for originality or impact). The formula is:

$$\text{MAE} = \frac{1}{n}\sum_{i=1}^{n}|y_i - \hat{y}_i|$$

where y_i is the true score, $\hat{y}_i$ is the predicted score, and n is the total number of items. Lower MAE values indicate more accurate predictions.

A.5 Unsupervised and Reward-Based Metrics

The need for scalable evaluation without a direct human reference for every generated review has driven the development of unsupervised and reward-based metrics.

Spearman's Rank Correlation Coefficient (ρ): Spearman's ρ measures the strength and direction of the monotonic relationship between two ranked variables. In peer review evaluation, it is used to compare the model-based ranking of reviews or systems with human-preference leaderboards. The formula is:

$$\rho = 1 - \frac{6\sum_{i=1}^{n} d_i^2}{n(n^2-1)}$$

where d_i is the difference between the ranks assigned by the two systems for item i, and n is the number of items. ρ ranges from -1 (perfect inverse correlation) to 1 (perfect agreement), with 0 indicating no correlation.

Kendall's Rank Correlation Coefficient (τ): Kendall's τ assesses the ordinal association between two ranked lists by counting the number of concordant and discordant pairs. Its formula is:

$$\tau = \frac{(C-D)}{\frac{1}{2}n(n-1)}$$

where C is the number of concordant pairs, D is the number of discordant pairs, and n is the number of items. Values of τ range from -1 (complete disagreement) to 1 (complete agreement).

Human-aligned Peer Review Reward (HPRR): HPRR is a composite, multi-objective reward function designed for reinforcement learning in peer review generation. It quantifies review quality as a weighted sum of several sentence-level attributes, such as Criticism (Cr), Examples (Ex), Suggestions and Solutions ($SuSo$), and overall relevance (often measured by METEOR). The formula can be abstracted as:

$$\text{HPRR} = w_1 \cdot Cr + w_2 \cdot Ex + w_3 \cdot SuSo + w_4 \cdot \text{M} + \ldots$$

where w_i are learned weights reflecting human preferences, M is METEOR. HPRR enables automated optimization toward human-aligned review usefulness and quality.

B Further Discussion of Benchmark Perspective

B.1 Open Challenge of Evaluation

Evaluating Deep Scientific Reasoning. Current methods, even the most advanced, are still limited

in their ability to assess the deep, domain-specific scientific reasoning that is the hallmark of a true expert review. Identifying logical fallacies, assessing the validity of a complex experimental design, or judging the true intellectual contribution of a novel method remains profoundly difficult to automate. Probing for critical errors in papers with known flaws is a promising first step, but significant work is needed to develop evaluations that can reliably measure these higher-order reasoning capabilities.

Towards Holistic Evaluation Frameworks. The future of evaluation does not lie in finding a single "best" metric. Instead, it requires the development of integrated, holistic frameworks that combine the strengths of multiple paradigms. A robust evaluation protocol for a new review generation system might involve using an unsupervised method like PiCO (Ning et al., 2025) for rapid, large-scale comparison against other models, employing a suite of aspect-oriented metrics (e.g., Focus-Level, SubstanScore) for deep diagnostic analysis of its specific flaws, and conducting targeted, comparative human evaluation on a small but representative sample of its outputs for final validation and bias checking.

B.2 Challenges for using the latest datasets (post-2023)

This part of the appendix discusses the practical challenges associated with using post-2023 peer review datasets and proposes mitigation strategies. Our goal is to help researchers make informed choices and inspire improvements in future dataset construction and refinement.

Challenges. Despite significant advancements in dataset scale, diversity, and annotation richness, several limitations persist across recent datasets for peer review generation. Many corpora remain optimized for auxiliary tasks such as score prediction or classification, rather than full-text review generation, which results in weak alignment between individual review sentences and specific paper content. Dialogue-based datasets often suffer from inconsistency in speaker roles, incomplete conversational threads, or ambiguous author responses, making them difficult to use in multi-agent training setups. Annotation scarcity is another common obstacle: datasets offering fine-grained critique labels are often small, limiting their utility for supervised learning. Additionally, review-level scores and aspect annotations are rarely paired with sentence-level metadata, making it hard to evaluate or optimize local review quality. Domain coverage also remains narrow, with most datasets focusing exclusively on NLP and ML venues—raising concerns about generalizability.

Mitigation. To address these issues, future dataset efforts should emphasize structural coherence and alignment. Establishing mappings between review sentences and the specific sections they critique would enable content-aware generation and interpretation. Dialogue datasets can benefit from more rigorous role standardization and filtering pipelines that remove incomplete or incoherent threads. Where annotation budgets are limited, semi-supervised expansion techniques—such as label propagation, weak supervision, or LLM-based pseudo-annotation—could help scale fine-grained labels. Additionally, pairing review-level scores with sentence-level utility judgments would enable more precise evaluation and training of critique generation. Finally, future datasets should aim to extend beyond the NLP/ML domain and incorporate reviews from diverse disciplines to better support general-purpose review generation models and cross-domain evaluation.

C Further Discussion of Methodologies

C.1 Open Challenge

Mitigating Bias in Automated Review. A critical and still under-explored challenge in automated peer review is the profound risk of these systems inheriting, amplifying the human biases present in their training data. At the same time, the very technologies for Controllable Text Generation (CTG) that could exacerbate this problem also offer a potential path toward building fairer and more objective reviewers. To effectively mitigate bias, we must first be able to measure it. This necessitates the parallel development of "bias benchmarks" designed specifically for the scientific domain.

Alignment Between Review Points and Final Decision. Recent studies expose a persistent disconnect between aspect scores and final decisions (Sukpanichnant et al., 2024). While models can predict both accurately in isolation (Muangkammuen et al., 2022; Yuan et al., 2021), bridging them remains difficult. Checklist-guided generation (Zeng et al., 2023) and decision-aware meta-reviewing (Kumar et al., 2024a) offer partial solutions by explicitly

linking critique structure to recommendations. Future work should explore causal modeling of score-to-decision transitions, perhaps with richer datasets like MOPRD (Lin et al., 2022) or REVIEWEVAL (Garg et al., 2024).

C.2 Future Direction

Reimagining Peer Review as a Process. Many generation models treat review writing as a one-shot task. Agent-based frameworks—such as REVIEWAGENTS (Gao et al., 2025), AGENTREVIEW (Jin et al., 2024a), and PEERREVIEWAS (Tan et al., 2024), instead simulate multi-turn workflows involving multiple roles. These systems better capture the iterative nature of review, rebuttal, and decision, and create opportunities for testing social factors (e.g., bias, anchoring, consensus drift) under controlled conditions. Their development offers both new modeling challenges and a tool for studying the peer review process itself.

Reinforcement Learning and Agentic Frameworks for Deeper Reasoning. To overcome the tendency of simple fine-tuned models to produce shallow, generic, and overly positive reviews, the field is rapidly adopting more complex learning paradigms. Multi-objective reinforcement learning and multi-agent or multi-stage frameworks are emerging as powerful approaches. These methods decompose the monolithic task of "writing a review" into more manageable, specialized, and logically sequenced sub-problems, thereby fostering deeper and more structured reasoning.

Peer Review as Improver. Is peer review's primary role gatekeeping acting as a filter, or should it embrace the role of an improver who provides actionable feedback to elevate the quality of the manuscript? Our survey of the literature reveals that current systems predominantly lean towards the gatekeeping paradigm. However, we believe that the future of automated scientific discovery (Novikov et al., 2025), hinges on a decisive shift towards the improvement-oriented paradigm. We hypothesize that improvement-oriented training in which the generated review directly informs the revision and is rewarded when it yields measurable improvements to the manuscript will better support automated discovery workflows. Recent frameworks link review signals to optimization loops such as CycleResearcher's RLHF loop (Weng et al., 2025) for research-review-refinement, suggesting a pathway from actionable review to iterative scientific progress.

CfP-aware Venue Fit and Policy Checking. Call for Papers (CfP) documents already encode conference scope, topical focus, and LLM-usage or desk-reject policies for venues such as NeurIPS and ARR. Existing LLM-based peer-review systems like SEA (Yu et al., 2024), OpenReviewer (Idahl and Ahmadi, 2025) and ReviewEval (Garg et al., 2024) align prompts or metrics with reviewer guidelines, but they do not treat CfPs themselves as retrieval sources. Future work could integrate CfP passages into RAG pipelines as machine-readable constraints to support venue-fit assessment, automatic flagging of policy violations, and assisted desk-reject decisions.

Remaining Non-Generative Components. Predictive tasks such as aspect-score regression (Muangkammuen et al., 2022), final-decision forecasting (Sukpanichnant et al., 2024), and desk-triage systems, i.e., automatic screening or "reject-option" classifiers (Hendrickx et al., 2021), remain critical, but their tooling is now intertwined with generation (e.g., conditioning meta-review generation on predicted scores).

Interactive Human-in-the-Loop Workflows. Rather than aiming for full automation, future systems will likely focus on creating more seamless and powerful human-AI collaboration tools. This could involve real-time, interactive interfaces where a human reviewer can guide, correct, and query the AI assistant, leveraging its ability to rapidly retrieve literature, analyze data, and draft text while retaining ultimate human oversight and judgment.

C.3 Multimodal Input

A comprehensive and high-fidelity automated peer review system must possess the capacity to understand and reason about the entirety of a scientific manuscript, including its non-textual components (Kumar et al., 2024c). The central challenge in this domain extends beyond the technical task of processing pixels or parsing table cells; it lies in bridging a profound semantic gap. Therefore, the most formidable aspect of multimodal integration is not the technical encoding of visual data, but rather teaching a model to perform this complex interpretive leap from perception to conceptual understanding. This suggests that generic vision-

language pre-training, while useful, is likely insufficient. Progress will depend on fine-tuning models on specialized tasks that explicitly reward this form of semantic interpretation (Wang et al., 2025).

A leading paradigm in this area is the Hybrid Modality Approach, exemplified by the Multi-Modal Automated Academic Papers Interpretation System (MMAPIS) (Jiang et al., 2024). For text and mathematical formula extraction, it leverages a transformer-based model like Nougat, which is adept at converting PDF pages into a structured format like Markdown. This process preserves not only the plaintext but also the intricate mathematical notations that are indispensable for understanding technical content. Concurrently, for visual elements, the system employs a separate tool, PDF-Figures 2.0. This tool reasons about the document's layout, such as the empty regions and whitespace separating blocks of text, to identify, extract, and caption figures and tables (Im et al., 2021).

C.4 Comparison Between Methodologies

Agent-Based Methods vs. Reinforcement Learning. Agent-based modeling and Reinforcement Learning represent two distinct philosophies for generating peer reviews. Agent-based approaches, such as AgentReview (Jin et al., 2024a), excel at simulating the complex, interactive social dynamics of the entire peer review ecosystem. However, the primary goal of these systems is often to study emergent behaviors rather than to optimize the quality of any single generated review. This focus on process simulation. In contrast, Reinforcement Learning directly targets the optimization of the review artifact. Frameworks like REMOR (Taechoyotin and Acuna, 2025) frame review generation as a policy that is optimized to maximize a multi-objective reward function, which is carefully designed to encapsulate human preferences for what constitutes a helpful review. This allows RL-based methods to effectively steer generation away from the shallow, overly positive feedback that vanilla LLMs often produce.

Iterative Refinement vs. Single-Pass Generation. Iterative refinement is a powerful technique that mimics the human cognitive process of drafting, critiquing, and revising (Xu et al., 2025; Kamoi et al., 2024). This approach is particularly effective for complex, multi-faceted tasks like peer review, as it allows the model to address objectives that are often missed in a single pass and can unlock the full potential of LLMS. The primary drawback is a significant increase in latency and computational cost, as each feedback-and-refine cycle constitutes an additional, expensive call.

Single-pass generation, the standard autoregressive approach, is the most computationally efficient method. However, it is inherently limited by its inability to perform global planning or correct early mistakes.

Retrieval-Augmented vs. Self-Contained Generation. Retrieval-Augmented Generation (RAG) (Dycke et al., 2023b) is critical for grounding peer reviews in verifiable, external knowledge, a necessity for tasks like assessing a paper's novelty or checking for factual inaccuracies. However, RAG introduces significant architectural complexity and a new critical point of failure: the retriever. The quality of the final generation is fundamentally bottlenecked by the quality of the retrieved context; irrelevant or low-quality retrieval leads to poor, unhelpful reviews.

Self-contained generation, which relies solely on the LLM's parametric knowledge, is architecturally simpler and faster at inference. Its primary limitation is that its knowledge is static, potentially outdated, and prone to factual invention (hallucination), making it fundamentally unsuited for rigorously evaluating a submission's contribution to a rapidly evolving field. Therefore, while self-contained generation may suffice for assessing aspects like clarity or presentation, any deep evaluation of a paper's technical substance and novelty necessitates grounding in external knowledge.

D Methodologies Summary

We summarizes the main methodologies for peer review generation in Table 1. We also organizes different LLM backbones used in methods in Table 3.

E Data Collection

We summarize in Table 4 the typical ways to gather papers together with their reviews. These range from directly using the OpenReview API to other APIs. In practice, researchers often combine multiple sources to ensure both coverage and diversity across venues.

Meanwhile, once data is collected, heterogeneous formats (PDFs, HTML pages, JSON exports, or XML metadata) need to be standardized. Ta-

System	Backbone LLM	Modification Type	Purpose / Role in the Framework	Open-source?
CycleResearcher — Weng et al. (2025)	Mistral-Nemo-12B (AI, 2023) Qwen2.5-Instruct-72B (Team, 2024) Mistral-Large-2 (123B) (AI, 2024)	SFT + RL (SimPO) + Reward model via LoRA	CycleResearcher generates papers using RLHF. CycleReviewer serves as the reward model.	Yes
DeepReview — Zhu et al. (2025a)	GPT-4 (API) (OpenAI et al., 2024)	Zero-shot (no finetuning)	Thinker agent produces reasoning traces. Writer agent generates the structured review.	No
OpenReviewer — Idahl and Ahmadi (2025)	Llama-3.1-8B-Instruct (AI@Meta, 2024)	Full finetuning (Axolotl)	Markdown-based review generation with template-specific prompts and PDF-to-text preprocessing.	Yes
REMOR — Taechoyotin and Acuna (2025)	DeepSeek-R1-Distill-Qwen-7B (Guo et al., 2024)	LoRA + RL (GRPO)	Backbone for both supervised and RL stages. Trained into REMOR-U and REMOR-H variants.	Yes

Table 3: Comparison of LLM Backbone Strategies in Recent Peer Review Generation Systems. We summarize the core base models used, how they are customized, their functional roles, and whether they are publicly available.

ble 5 lists common tools and pipelines for data type conversion.

Together, these resources provide a practical toolkit for curating review corpora that are both machine-readable and comparable across venues.

F Evaluation Summary

We summarizes the evaluation methods for peer review generation in Table 6. We also compares their pros&cons in Table 2. The datasets we organizes are shown in Table 7.

G Literature Review Procedure

To ensure transparency and rigor, we document here the procedure by which our literature review was conducted.

Databases. We searched major sources including ACL Anthology, OpenReview (ICLR, NeurIPS), arXiv, Semantic Scholar, DBLP, and Google Scholar.

Search Strategy. We applied keyword combinations such as "peer review generation," "meta-review," and "rebuttal generation" within the time range 2018–2025. In addition, we adopted a snowballing strategy by tracing the references and citations of seminal works (e.g., PeerRead (Kang et al., 2018)) and recent contributions (e.g., DeepReview (Zhu et al., 2025a)).

Inclusion Criteria (IC).

- IC1: The paper must directly address AI-based peer review generation, its evaluation, or closely related subtasks (e.g., meta-review, rebuttal generation).
- IC2: The paper must be publicly available as a journal article, conference paper, or preprint.
- IC3: The paper must be written in English.
- IC4: Focus primarily on work published after 2018, especially post-LLM developments.

Exclusion Criteria (EC).

- EC1: Studies that only predict or analyze peer review scores without generating review text (e.g., (Thelwall and Yaghi, 2024; Laureano et al., 2025)).
- EC2: Abstracts, short articles, and non-academic blog posts.
- EC3: Papers without accessible full text.

Screening and Statistics. Our initial screening retrieved approximately 500 articles. After deduplication, around 400 articles remained. Applying inclusion/exclusion criteria yielded 194 papers, of which 138 directly focus on peer review generation and its evaluation.

Methodological Rigor. Our protocol is informed by established guidelines for systematic reviews, such as the Kitchenham and Charters (Kitchenham and Charters, 2007) SLR framework and the PRISMA 2020 statement (Page et al., 2021). These emphasize transparent reporting of search strings, last search date, de-duplication, per-stage counts, and inclusion/exclusion flows. By following these standards, we ensure that our literature review process is rigorous, reproducible, and aligned with recognized best practices.

Channel / Platform	Representative Studies	Disciplines	Collection Methods & APIs (abbr.)
OpenReview	Kang et al. (2018); Dycke et al. (2023a); Weng et al. (2025)	ML / NLP	• OR-API • GQL • HTML-WC
Softconf / START	Kang et al. (2018); Dycke et al. (2023a)	NLP	• OPT-IN • SQL-DMP
F1000Research	Dycke et al. (2023a)	Life Sci	• CSV-EXP • HTML-SCR • DOI-TCK
PeerJ	Lin et al. (2022)	Bio / Chem / CS	• REST-CRL • HTML-CRL
Nature Comms PR File	Tan et al. (2024)	Multidomain	• HTTP-DL

Table 4: Channels and APIs for Harvesting Paper–Review Corpora. Abbreviation key: OR-API = *openreview-py* API, GQL = GraphQL query, HTML-WC = web crawler, OPT-IN = author/reviewer opt-in dump, SQL-DMP = SQL/CSV dump, CSV-EXP = official CSV export, HTML-SCR = HTML scraping, DOI-TCK = DOI tracking, REST-CRL = REST crawler, HTTP-DL = bulk HTTP download.

Tool / Stage	Used in Studies	Input → Output	Main Purpose & Highlights
GROBID (GRO, 2008–2025)	Lin et al. (2022)	PDF → XML	Accurate scholarly parse; keeps sections/cites; batch-ready.
Science Parse (Sci, 2017–2025)	Kang et al. (2018)	PDF → JSON	Fast metadata + rough text; web-scale.
Marker (Mar, 2023–2025)	Tan et al. (2024)	PDF → MD	Long-context MD; preserves headings and formulas.
MagicDoc (Contributors, 2024)	Weng et al. (2025)	PDF → MD	Bulk Softconf PDFs to MD; code-block safe.
MinerU (Wang et al., 2024)	Zhu et al. (2025a)	PDF → MD / JSON	High-precision; layout/reading-order faithful.
Semantic Scholar (Kinney et al., 2023)	Dycke et al. (2023a)	DOI → BibJSON	Reference enrichment; citation, year, impact.

Table 5: Tools Frequently Used to Convert / Enrich Paper–Review Data

Category	Paper	Core Method	Metrics / Framework Details	Key Finding / Contribution
Human-Centric *(Direct Rating)*	Robertson (2023)	Pilot Study	10 users rated helpfulness (1-5 scale).	GPT-4 reviews had same mean helpfulness as humans (3/5) but higher variance.
	Liang et al. (2023)	Large-Scale User Study	308 researchers rated LLM feedback on their own papers.	High perceived value: 57.4% found it helpful; 82.4% found it more beneficial than some human reviews.
	Drori and Te'eni (2024)	Structured Evaluation	GPT-4 reviews structured against standard review criteria.	AI assistance can reduce reviewer workload but is not completely reliable for all cases.
	Hosseini and Horbach (2023), Schintler et al. (2023), Lee et al. (2025)	Qualitative Critique	Review of benefits, risks, and ethical concerns.	Highlighted efficiency gains but warned of bias, compromised integrity, and weakness in assessing scientific validity.
Human-Centric *(Pairwise Comparison)*	Tyser et al. (2024)	Arena-Based Evaluation	**Reviewer Arena**: Pairwise human (and LLM) preference comparisons.	Proposes a scalable method for quality assessment based on relative judgments rather than absolute scores.
Reference-Based *(Overlap Metrics)*	Liang et al. (2023)	Semantic Matching	Two-stage process: GPT-4 extracts key comments, then semantic matching calculates a "hit rate".	Offers a nuanced content overlap metric beyond simple lexical similarity.
	Zhou et al. (2024a)	Standard Metrics	ROUGE, BERTScore against concatenated human reference reviews.	Established a baseline methodology for using standard NLP metrics in this domain.
	Yu et al. (2024)	Standard Metrics (SEA)	BLEU, ROUGE, BERTScore against concatenated human references.	Noted that ROUGE-Recall is particularly indicative of review comprehensiveness.
Reference-Based *(Task-Based)*	Liu and Shah (2023)	Targeted Tasks	Seeded error detection, checklist verification, abstract comparison.	LLMs show mixed results: successful at some tasks (checklist) but poor at others (error detection, comparative judgment).
	Zhou et al. (2024a)	MCQ Benchmark	**RR-MCQ**: A multiple-choice question dataset from ICLR review-rebuttal exchanges.	LLMs achieve passable accuracy (>60%) but struggle with long papers and providing deep critical feedback.
LLM-Based	Zhang and Abernethy (2025)	Multi-Judge Pipeline	Requires an affirmative vote from two distinct LLM judges to confirm a "hit" (finding a critical error).	Aims to reduce single-model bias and increase the reliability of automated evaluation.
	Ning et al. (2025)	Peer-Evaluation	**PiCO Framework**: LLMs both generate reviews and provide pairwise preferences on peers' outputs. Ranks models via consistency optimization.	Pioneers fully unsupervised evaluation without human labels or gold references.
Aspect-Oriented	Shin et al. (2025)	Focus-Level Framework	Annotates review content by targets (e.g., Method) and aspects (e.g., Novelty) to compare LLM vs. human focus.	Revealed LLM "blind spots": over-emphasis on technical validity, under-emphasis on novelty.
	Du et al. (2024)	Fine-Grained Annotation	**ReviewCritique**: Dataset with 23 fine-grained "deficiency" types annotated at the sentence level.	Enables testing of whether models can identify specific types of errors found by humans.
	Ye et al. (2024)	Adversarial Testing	"Review injection attacks" and testing susceptibility to authors' proactive disclosure of minor limitations.	Demonstrates LLMs' vulnerability to both explicit and implicit manipulation.
	Yu et al. (2024)	Inconsistency Metric	**Mismatch Score**: Predicts the degree of inconsistency between a paper and a given review.	Introduces a novel aspect-specific metric focused on review-paper faithfulness.

Table 6: Taxonomy of Evaluation Methodologies for Automated Peer Review Generation. This table summarizes the primary approaches, key papers, specific metrics or frameworks, and notable findings discussed in the literature.

Era	Phase	Name	Main Use	Scale	Source	Expert Annotation
Pre-2023	Reviewing	PEERREAD (Kang et al., 2018)	Review-score prediction, acceptance prediction, review generation	14.7k papers; 10.7k reviews	ICLR 2017–2019 OpenReview; ACL 2017 author-consented drafts/reviews; NeurIPS 2013–2017 drafts matched to accepted arXiv papers	–
		DISAPERE (Kennard et al., 2021)	Discourse analysis of reviewer–author discussion	506 threads	ICLR 2019–2020 OpenReview discussion threads	Sentence-level discourse relations in review–rebuttal discussions
		PEERASSIST (Bharti et al., 2021)	Acceptance prediction, review-score prediction	4,467 papers; 13.4k reviews	ICLR 2017–2020 OpenReview submissions and reviews	–
		ASAP-REVIEW (Yuan et al., 2022)	Aspect-aware review generation, aspect tagging	8,877 papers; 28.1k reviews	ICLR 2017–2020 and NeurIPS 2016–2019 OpenReview data	1k-review subset annotated with 15 aspect labels
		REACT (Choudhary et al., 2022)	Actionability classification, comment-type tagging	1.25k labeled; 52k unlabeled comments	ICLR 2018 OpenReview comments	Crowdsourced actionability and 7 comment types
	Meta-review	MRED (Shen et al., 2021)	Structure-controlled meta-review generation	7,089 meta-reviews; 45k sentences	ICLR 2018–2021 OpenReview meta-reviews	Every sentence labeled with 9 discourse tags; adjudicated
		PRRCA (Wu et al., 2022)	Meta-review generation from reviews and rebuttals	6,138 submissions	ICLR 2017–2022 OpenReview threads with reviews, rebuttals, decisions, and meta-reviews	–
	Multi-stage	ICLR-DB (Zhang et al., 2022)	Fairness analysis, decision/review prediction, review generation	10.3k submissions; 36.5k reviews; 68.7k responses	ICLR 2017–2022 OpenReview threads, enriched with author profiles and LLM-derived features	–
		MOPRD (Lin et al., 2022)	Meta-review generation, decision prediction, rebuttal generation	6,578 papers	PeerJ peer-review threads with reviews, rebuttals, meta-reviews, and decisions	Manually aligned review-process records
Since 2023	Meta-review	PEERSUM (Li et al., 2023)	Meta-review generation from discussions	14,993 triples	ICLR 2020–2022 OpenReview threads with reviews, discussions, and meta-reviews	–
	Reviewing	NLPEER (Dycke et al., 2023a)	Score prediction, pragmatic labeling, guided skimming	5,672 papers; 11k+ reviews	ICLR and NeurIPS 2017–2022 OpenReview submissions and reviews	Sentence-level pragmatic tags on an F1000 subset
		SUBSTANREVIEW (Guo et al., 2023)	Claim substantiation detection and scoring	550 reviews	ARR 2021–2022 reviews	Expert span-level claim–evidence annotation
		REVIEWCRITIQUE (Du et al., 2024)	Deficiency detection in human and LLM reviews	100 papers	100 NLP papers with human and LLM reviews	Experts annotate 23 fine-grained deficiency types
		REVIEWEVAL (Garg et al., 2024)	Evaluation of human and LLM reviews	120 papers	120 NeurIPS, ICLR, and UAI submissions with gold and candidate reviews	–
		REVIEW-5K (Weng et al., 2025)	Aspect-level score prediction, review evaluation	4,189 train + 782 test; 16k+ comments	ICLR 2023 OpenReview reviews and comments	–
		REVIEWER2 (Gao et al., 2024c)	Aspect-prompted review generation, diversity benchmarking	27k papers; 99k reviews	Reviews from 6 major ML/NLP venues, 2017–2022	LLM-generated aspect prompts
		DEEPREVIEW-13K (Zhu et al., 2025b)	Deep-thinking review generation	13k structured reviews	OpenReview papers and reviews, 2017–2023	Reasoning-step annotations for supervision
	Revision	ARIES (D'Arcy et al., 2023)	Comment–edit alignment, revision generation	3.9k comment–edit pairs; 196 test cases	ICLR, NeurIPS, and ACL OpenReview papers, 2018–2022	Expert-annotated gold test set
	Rebuttal-included	AGENTREVIEW (Jin et al., 2024b)	Peer-review simulation, bias/dynamics analysis, synthetic review generation	∼500 submissions; 10k reviews; 53.8k artifacts	LLM-agent simulations over ICLR 2018–2021 papers	–
		RR-MCQ (Zhou et al., 2024b)	Review–revision QA, reviewer reliability benchmark	197 MCQs	ICLR 2023 review–rebuttal threads	Manually written and labeled MCQs
		REVIEWMT (Tan et al., 2024)	Dialogue-style review simulation	26k+ papers; 92k+ reviews	OpenReview reviews reorganized as multi-turn dialogues	Speaker roles labeled
		RE^2 (Zhang et al., 2025a)	Full-stage review and rebuttal modeling, decision prediction, dialogue generation	19.9k submissions; 70.7k reviews; 53.8k rebuttals	OpenReview data from 24 conferences and 21 workshops, 2017–2025	Consistency-filtered multi-turn rebuttal threads

Table 7: Public peer-review datasets, grouped by era and phase.