Research Square

Preprints are preliminary reports that have not undergone peer review.
They should not be considered conclusive, used to inform clinical practice,
or referenced by the media as validated information.

PeerLens as a Multi-Agent LLM Reviewer

Emre Dil

em311036@dal.ca

Dalhousie University

Talha Zafer

Imperial College London

Article

Keywords:

Posted Date: November 6th, 2026

DOI: https://doi.org/10.21203/rs.3.rs-11323645/v1

License: This work is licensed under a Creative Commons Attribution 4.0 International License.
Read Full License

Additional Declarations: No competing interests reported.

PeerLens as a Multi-Agent LLM Reviewer

Emre Dil* Talha Zafer†

September, 2026

Abstract

Scientific peer review is about to break under its own weight. Submission volume is already outpacing the reviewer pool, and the upcoming era of human directed AI generated research will multiply the load by orders of magnitude. We present PeerLens, a domain adaptive multi agent LLM system that produces a full referee report from a manuscript PDF using eight specialized agents (classification, claim extraction, dimensional analysis, symbolic verification, novelty search, structure and narrative review, domain expert review, and reproducibility review), each tuned to the manuscript's field. We evaluate PeerLens on 64 papers across physics, biology, mathematics, and chemistry (drawn from SciPost, eLife, F1000Research). On the 51 papers with available first round human reviewer reports, PeerLens agreed with the reviewers or diverged by being stricter, always on concrete technical grounds such as broken statistical assumptions, dimensionally malformed metrics, definitions that void their own novelty claims, and unresolved mechanistic dependencies. We observed no fabricated citations or invented technical claims. The strictness is a feature, not a defect: it reflects criteria based enforcement of disciplinary standards at scale, which is a prerequisite for peer review to continue functioning under the volume of manuscripts the community is about to face.

1 Introduction

1.1 The strain on the peer review system

The volume of scientific publications has grown steadily for two decades and has accelerated in the last five years. Hanson and colleagues estimate that total articles indexed in Scopus and Web of Science grew by about 47% from 2016 to 2022, outpacing the growth in the practicing scientific workforce over the same period [1]. Preprint servers show a similar trend: arXiv passed 20,000

*PeerLens co-founder. Department of Physics and Atmospheric Science, Dalhousie University, Halifax, Nova Scotia, Canada Email: em311036@dal.ca

†PeerLens co-founder. Department of Civil and Environmental Engineering, Imperial College London, UK Email: t.zafer@imperial.ac.uk

new submissions in a single month for the first time in May 2023 and set a new record of 20,710 submissions in October 2023, bringing its cumulative total above 2.35 million [2]. The pool of scholars willing to review, however, has not grown at the same rate.

The 2018 Publons Global State of Peer Review Report surveyed over 11,800 researchers and documented an increasing gap between reviewer demand and reviewer supply, with more than half of review invitations declined in 2017 [3]. Subsequent reports and editorials describe the same trend continuing: journals now often need to send eight to ten invitations to secure one reviewer, and the average time from submission to decision has grown to more than 100 days in many disciplines [4]. In August 2025, Nature ran a feature titled "The peer-review crisis: how to fix an overloaded system," in which editors and funders discussed the operational strain the system is under [5].

Beyond throughput, there are concerns about consistency of reviewer judgments. The 2014 NeurIPS consistency experiment assigned 10% of submissions to two independent program committees and found that the two committees disagreed on 26% of accept/reject decisions [6]. A repeat of this experiment at NeurIPS 2021 reached a similar conclusion: roughly 23% of duplicated papers received different final decisions, and if the review process were rerun independently about half of the accepted papers would have been different [7]. A follow up analysis by Cortes and Lawrence found that about 50% of the variation in reviewer quality scores was subjective in origin, and that for the accepted papers there was no correlation between reviewer scores and later citation impact [6]. These experiments were run in machine learning, but similar disagreement patterns have been documented in biomedical and physical sciences as well [8].

1.2 A second pressure: AI-assisted authorship

A separate development now affects the input side of peer review. Large language models are being used at multiple stages of manuscript preparation, from literature review and drafting to figure captions and final proofreading. Surveys of researchers in 2023 and 2024 report a wide range of usage rates. A 2023 Nature survey of 1,600 scientists found that about 30% had used generative AI to help write manuscripts and 47% considered AI "very useful" for academic tasks [9]. A 2024 Elsevier survey of nearly 3,000 researchers reported that 31% had used generative AI for research activities, and a 2024 Oxford University Press global survey of 2,300 researchers found that 76% had used AI tools in some form [10, 11]. Detection studies that estimate LLM involvement from writing style suggest that between nearly 5% (mathematics) and 17.5% (computer science) of sentences in recent papers show signs of substantial LLM modification, with usage rising steadily since 2023 [12].

If a growing fraction of manuscripts is AI assisted in their preparation, and the reviewer pool is fixed, two effects follow. First, total submission volume is likely to keep rising because the effort to produce a draft is falling. Second, some fraction of those manuscripts will contain claims that are stylistically fluent but weakly supported, because current AI writing tools optimize for coherence

rather than for the specific standards a reviewer in that subdomain would apply. Neither of these effects is fully absorbed by the current peer review workflow.

These effects are not stable. Manuscripts that are primarily theoretical, computational, or analytical, and that do not depend on the physical constraints of a wet lab or an experimental apparatus, are increasingly written with substantial LLM involvement. As model capabilities continue to improve, the marginal cost of turning a research idea into a written manuscript in these categories is falling toward zero, and a researcher who used to produce a small number of papers per year with months of writing effort per paper can plausibly produce many times that volume in the same period. If this pattern continues, the submission rates in theory heavy, model heavy, and analysis heavy fields will substantially outpace the capacity of any reviewer pool that recruits from the same fixed population of human experts. Under those conditions, the question is not whether AI systems will do a share of the reviewing but which parts of it, and under what safeguards.

1.3 Early attempts to bring AI into peer review

Journals and conferences have begun to experiment with AI assistance for reviewers rather than authors. Liang and colleagues carried out a large scale study on 3,096 Nature family papers and 1,709 ICLR papers, comparing GPT-4 generated feedback with human reviewer comments, and reported average overlap rates of 30.9% and 39.2% respectively, comparable to the overlap between two human reviewers (28.6% and 35.3%) [13]. In a follow up user study with 308 researchers, 57.4% found the LLM generated feedback helpful or very helpful, and 82.4% found it more useful than feedback from at least some human reviewers on their own papers [13].

Conferences have begun to run controlled experiments. NeurIPS 2024 deployed an optional LLM based Checklist Assistant that helped authors verify compliance with the submission checklist [14]. ICLR 2025 ran a large randomized controlled study of over 20,000 reviews in which reviewers received automated LLM feedback on their draft reviews. About 27% of reviewers who received feedback updated their reviews, incorporating over 12,000 suggestions [15]. AAAI-26 announced a larger AI review pilot that includes LLM generated first stage reviews for some tracks [16]. Alongside these conference experiments, an empirical study of ICLR 2024 estimated that at least 15.8% of reviews at that conference were already written with some AI assistance, and that AI assisted reviews were associated with higher paper scores and acceptance rates [17].

Most of these efforts share a common design. A general purpose language model is given a manuscript together with a prompt that asks it to write a review, summarize the paper, or flag concerns. The output is a single block of text produced in one pass. This design is easy to implement but has known limits. A recent large scale evaluation of five LLMs on 11,000 machine generated reviews found that model generated reviews tend to be less variable, positively biased, and overconfident compared to human reviews, and that they often miss the specific weaknesses that human reviewers identify [18]. Reviews produced

this way are also described by editors as fluent but shallow [19].

1.4 The case for a domain adaptive multi-agent design

An alternative approach is to separate the reviewer's judgment into several distinct tasks and assign each task to a specialized agent, with prompts, tools, and evaluation criteria that fit that task. A paper in condensed matter physics does not raise the same concerns as a paper in analytical chemistry or a paper in commutative algebra. A generic prompt cannot reliably know when to ask about a missing gauge choice, when to ask about missing supporting information for a synthesis, and when to ask about a missing hypothesis in a theorem statement. If these judgments are separated into different agents, each with a prompt and vocabulary tuned to its subdomain, the overall system can act more like a panel of specialist reviewers than like a single generalist.

This paper describes such a system and reports early results. The system, called PeerLens, runs eight agents for each manuscript: a classifier that identifies the paper's field and subdomain, a claim extractor, a dimensional checker, a symbolic verifier, a novelty search agent, a structure and narrative scanner, a domain expert that applies a subdomain specific persona chosen from a bank of 27 personas covering physics, mathematics, chemistry, and biology, and a reproducibility checker. The claim extractor uses different schemas for different fields, so that a math paper produces theorem and lemma records while a chemistry paper produces synthesis steps, characterization results, and calibration equations. The output is a full referee report in the format a working reviewer would submit: a recommendation with rationale, a plain language summary of the manuscript, its overall strengths and weaknesses, and section by section commentary covering novelty and prior work, domain specific content, presentation and clarity, reproducibility, and numerical and symbolic consistency.

The point of this design is not to replace human reviewers, at least not in its current form. It is to change what human reviewers spend their time on. A full first pass referee report produced automatically and reliably lets human reviewers focus on the manuscripts that most need their expert judgment, and lets them start from a concrete draft rather than from a blank sheet. This division of labor is similar to how radiologists now work with AI assisted image reading in some clinical settings, where the AI produces a preliminary read and the human clinician confirms and refines.

1.5 Contributions and structure of this paper

This paper makes three contributions. First, it describes a concrete multi-agent architecture for editorial triage that adapts to four scientific fields (Section 3). Second, it reports on the outputs of this system for a set of test manuscripts, and discusses where it agrees with human reviewer judgment and where it does not (Section 4). Third, it discusses limitations and the conditions under which such a system can be responsibly integrated into an editorial workflow (Section 5).

The rest of the paper is organized as follows. Section 2 reviews related work in AI assisted peer review, hypothesis generation systems, and evidence synthesis tools. Section 3 describes the PeerLens architecture and the design of its individual agents. Section 4 presents the evaluation on a set of manuscripts. Section 5 discusses limitations and outlook. Section 6 concludes.

2 Related Work

Work related to PeerLens falls into three broad areas: AI assistance for peer review, AI systems for hypothesis generation and scientific discovery, and AI based literature and evidence synthesis. We review each in turn, and then position PeerLens against them.

2.1 AI assistance for peer review

Efforts to bring language models into peer review have taken two shapes. The first shape helps authors before submission. The NeurIPS 2024 Checklist Assistant fell into this category: it read a submission together with the conference checklist and generated suggestions to the author on missing or unclear items [14]. The second shape helps reviewers or editors during the review itself. The ICLR 2025 Review Feedback Agent read a reviewer's draft review and suggested improvements, and 27% of reviewers who received the feedback updated their reviews [15]. The AAAI-26 pilot goes further, using LLM generated first stage reviews on some tracks with human reviewers taking a follow up role [16].

Alongside these conference level efforts, several academic studies have examined how well a general purpose LLM can play the role of a reviewer directly. Liang and colleagues compared GPT-4 feedback against human reviewer feedback on 3,096 Nature family papers and 1,709 ICLR papers and reported overlap rates comparable to two human reviewers on the same paper [13]. Zhou and colleagues carried out a more critical evaluation, prompting several LLMs to review papers and comparing the produced reviews against a taxonomy of expected reviewer behaviors [20]. A large scale benchmark by the PRAIB group evaluated 11,000 machine generated reviews across five proprietary and open source models and found that model generated reviews are less variable, positively biased, and overconfident, and that they often miss the specific weaknesses that human reviewers identify [18].

These studies share a common experimental setup. A manuscript is passed in full to a single language model together with a prompt asking for a review, and the model returns a single block of text. The design is uniform across studies because it mirrors what a researcher can do in a few minutes with a general chatbot. Its limits, in the same studies, are also uniform: fluent output, weak coverage of technical specifics, and low agreement with the specific weaknesses human reviewers identify. Recent work has begun to explore alternatives that break the review task into smaller pieces. Zhang and Abernethy proposed a reasoning based approach in which the LLM is prompted to identify critical

problems in a paper step by step rather than producing a full review in one pass [21]. ReviewerToo asked whether an LLM can join a program committee alongside human reviewers as one of several voices [22]. PRISM introduced a multi dimensional benchmark that separates novelty, technical soundness, clarity, and other reviewer judgments [23]. PeerLens is closest in spirit to this last line of work: it separates the review into agents, each with its own concern, and then aggregates.

2.2 AI systems for hypothesis generation and scientific discovery

A parallel line of AI-for-science work targets the input side of the research pipeline rather than the output side. Google DeepMind's Co-Scientist, published in Nature in May 2026, is a multi-agent system built on Gemini in which specialized agents generate, debate, rank, and evolve scientific hypotheses [24]. Co-Scientist has been applied to drug repurposing, antimicrobial resistance, and liver fibrosis, with several of its hypotheses reaching wet-lab validation. FutureHouse's Robin, published in the same window, targets a similar role in biomedical research, generating candidate hypotheses and helping design follow up experiments [25]. AlphaEvolve, also from DeepMind, improves algorithms through evolutionary code modification, illustrating that a multi-agent design can extend beyond hypothesis generation into concrete algorithmic work [26].

These systems and PeerLens sit on opposite ends of the manuscript lifecycle. Co-Scientist and Robin help produce hypotheses that a scientist can then test and write up. PeerLens is invoked after the writing is done, to help decide whether a manuscript is ready for the next stage. The two categories are complementary rather than competing. A scientific workflow that includes both would see AI help generate the ideas, help design the experiments, and then help triage the resulting manuscripts.

2.3 AI based literature and evidence synthesis

A third relevant class of AI systems reads the existing literature rather than the manuscript under review. Elicit performs structured extraction from search results, producing tables of study characteristics, sample sizes, and findings across multiple papers [27]. Consensus synthesizes the direction of evidence on a specific yes or no question by aggregating findings across studies [28]. Scite analyzes citation contexts to determine whether a cited claim is supported, contradicted, or merely mentioned in the citing paper [29]. Semantic Scholar provides AI generated summaries and citation graph exploration across more than 200 million papers [30].

These tools inform the reviewer's judgment rather than replacing it. A reviewer using Consensus can check whether a paper's claimed effect is consistent with prior literature. A reviewer using Scite can check whether the paper's foundational citations are still supported. PeerLens uses this class of tools directly: its novelty agent invokes the model's web search capability at review

time, retrieves related prior work in the same subfield, and reports which of those papers overlap with the manuscript's central claims and which relevant papers are missing from its reference list. In this sense the novelty agent is a compact, in loop version of the evidence synthesis idea, oriented not toward answering a research question but toward informing the reviewer's judgment.

2.4 Positioning PeerLens

PeerLens differs from prior work in three ways. First, it produces a full referee report in the format a working reviewer would submit, covering recommendation, summary, strengths, weaknesses, and section by section commentary on novelty, domain content, presentation, reproducibility, and numerical and symbolic consistency. Where most prior LLM based review efforts stop at a summary or a checklist, PeerLens closes the loop and returns the artifact an editor or a downstream reviewer actually needs. Second, it is multi agent by construction. Where most prior LLM based review systems produce a single block of text from a single prompt, PeerLens runs eight specialized agents and combines their outputs. Third, it is field adaptive. The claim extractor, the domain expert persona, the reproducibility keyword scan, and the structure trigger scan each take a different form depending on whether the paper is in physics, mathematics, chemistry, or biology. The next section describes this architecture in detail.

3 The PeerLens System

3.1 Overall pipeline

PeerLens takes a PDF manuscript as input and produces a full referee report as output. The pipeline runs in eight sequential agents. The first agent identifies the field and subdomain of the paper. This decision then shapes what the later agents do: what to extract, what to check, and what to look for. The second agent extracts the paper's testable claims. The third agent checks whether the paper's equations are dimensionally consistent. The fourth agent attempts symbolic verification of algebraic identities. The fifth agent searches a corpus of related work and reports the closest matches. The sixth agent scans the paper's structure and narrative for the kinds of hedges and gaps that referees typically flag. The domain expert agent applies a subdomain specific persona to the paper. The reproducibility agent checks whether the paper provides the code, data, and methodological detail needed for a competent reader to reproduce the work. Each agent writes its result to a per agent human readable prose output so that later agents can read it, and the aggregator assembles the final markdown report from those files.

The point of running many small agents rather than one large prompt is that each agent has a narrower task and a narrower failure mode. A dimensional checker that finds nothing is unremarkable, but a dimensional checker

Agent	Physics	Math	Chemistry	Biology
Classification	active	active	active	active
Claim extraction	physics schema	math schema	chemistry schema	biology schema
Dimensional check	active	*skipped*	active	*skipped*
Symbolic verification	active	inactive	partial	inactive
Novelty search	active	active	active	active
Structure scan	base triggers	base + math	base + chem	base + bio
Domain expert	physics persona	math persona	chem persona	bio persona
Reproducibility	base keywords	*skipped*	base + chem keywords	base + bio keywords

Table 1: How each agent behaves for the four supported fields. Agents marked *skipped* return early with a short reason (dimensional analysis and reproducibility scrutiny do not apply to pure mathematics; the same is true for dimensional analysis in biology). Symbolic verification is listed as *partial* for chemistry because it is invoked only for calibration equations that are explicitly typed as algebraic identities by the claim extractor.

that produces a false positive can be corrected without affecting the novelty search or the domain review. In practice this separation also makes the system easier to debug and easier to extend to a new field.

3.2 Field adaptive dispatch

The same eight-agent skeleton is used for every paper, but individual agents behave differently depending on which field the classifier assigns. Some agents are field agnostic. Some load a field specific vocabulary. Some are gated off entirely when they do not apply. Table 1 summarizes these behaviors.

The claim extractor is the agent in which field adaptation is most visible. A physics paper produces a set of dimensional consistency claims, algebraic identities, and statements about limits, symmetries, and conservation laws. A math paper produces theorems, lemmas, propositions, corollaries, and definitions, each with a hypothesis, a conclusion, and where available a hint about the proof strategy. A chemistry paper produces synthesis steps, characterization results, performance metrics, calibration equations, method parameters, and separate records for selectivity, stability, and validation. A biology paper produces mechanistic hypotheses, model equations and parameters, empirical measurements, qualitative predictions, statistical claims, and data availability statements. The four schemas share a common outer envelope so that downstream code that reads the claim count or the symbol table works for any field.

The domain expert agent draws a persona from a bank of 27 subdomain personas: 6 for physics, 8 for mathematics, 6 for chemistry, 6 for biology, plus a universal fallback. Each persona applies a checklist calibrated to the concerns that a working referee in that subdomain would typically raise. The reproducibility agent picks its keyword lists (for code availability, data deposition, method reporting) based on the field, so that a chemistry paper is checked against synthesis and characterization keywords while a biology paper is checked against cell line, antibody, replicate, and code deposit keywords. The structure scan combines a field agnostic set of hedge patterns with an additional set specific to

each field, so that math specific phrases like "it is straightforward to show" or "without loss of generality" are picked up alongside general phrases like "it is well known that."

3.3 Aggregation and report

Once the agents have run, an aggregator reads their outputs and assembles the referee report. The report opens with the recommendation and a one or two sentence rationale drawn from the domain expert agent's verdict. It then presents a plain language summary of the manuscript, followed by its overall strengths and its overall weaknesses. The bulk of the report is the section by section commentary, which is written in the voice of a working reviewer and covers, in a fixed order, novelty and prior work, domain specific content under the persona assigned by the classifier, presentation and clarity, reproducibility, and numerical and symbolic consistency. The reproducibility section is omitted for pure mathematics submissions, where the axis does not apply. A short technical footer records the model used and the classifier's field and subdomain assignment.

The report is written to be usable by an editor as a first pass and by a downstream human reviewer as a starting draft. Its length is comparable to a working referee's report rather than to a summary or a checklist.

3.4 Test corpora

The pipeline is validated against four open peer review corpora, one per field, so that its output can be compared directly against the judgment of human referees in the same field.

Physics validation uses SciPost Physics, whose accepted papers are published together with the full text of the referee reports. Each report carries a first round recommendation of *publish*, *ask for minor revision*, *ask for major revision*, or *reject*, which maps directly to the PeerLens verdict scale.

Biology validation uses eLife's Reviewed Preprints. Each reviewed manuscript is accompanied by public reviewer comments and an editorial eLife Assessment, which includes two standardized tags: significance (*landmark*, *fundamental*, *important*, *valuable*, *useful*) and strength of evidence (*exceptional*, *compelling*, *convincing*, *solid*, *incomplete*, *inadequate*). These tags provide a nuanced ground truth that does not collapse to a single accept or reject verdict, and we report our results against them without imposing a one to one mapping.

Chemistry validation uses SciPost Chemistry, an open peer review journal in which every published paper is accompanied by the full text of its reviewer reports and per report recommendations. SciPost Chemistry is a small journal with fewer than ten published research articles at the time of writing, and we use all of them.

Mathematics validation uses F1000Research articles tagged under the Mathematics subject. Each paper is accompanied by public reviewer reports and

a standardized per reviewer recommendation of *approve*, *approve with reservations*, or *reject*, which we aggregate to a paper level verdict using F1000Research's own indexing rule. The F1000Research Mathematics tag covers a broader range of quantitative science than pure mathematics alone, including statistical methodology, mathematical biology, and quantitative social science, and the resulting corpus reflects that breadth.

We evaluate PeerLens against a cross domain corpus of 64 manuscripts: 21 from SciPost Physics, 14 from eLife, 20 from F1000Research (3 initial pilot papers plus 17 systematically sampled from the Mathematics subject browse), and 9 from SciPost Chemistry. The corpus intentionally covers the four fields the pipeline is designed for, so that field agnostic behavior, field specific behavior, and known limits are all exposed. Where an F1000Research reviewer submitted reports across multiple revision rounds we retain only the first round report, so that PeerLens (which operates on the v1 PDF) is compared against the reviewer's initial judgment rather than against a post revision publish recommendation. The full paper level results, case studies, and cross domain patterns are reported in Section 4.

4 Evaluation

We report on the output of PeerLens for the 64 paper cross domain corpus described in Section 3.4. Section 4.1 covers physics, Section 4.2 covers biology, Section 4.3 covers mathematics, and Section 4.4 covers chemistry. Section 4.5 discusses the pattern that emerges across all four. Section 4.6 addresses the observation that PeerLens is systematically stricter than the human reviewers whose decisions we compare against, and argues that this is a feature rather than a defect once the intended use case is properly stated.

Throughout this section, "fair comparison" means the PeerLens verdict (produced on the v1 manuscript PDF) is compared against the human reviewers' first round recommendation, before the authors have had a chance to address concerns in a revised submission. Later round recommendations, which typically drift toward *publish* once initial concerns are addressed, would understate the agreement between PeerLens and human referees on the same question.

4.1 Physics: SciPost Physics

We ran PeerLens on 21 SciPost Physics manuscripts drawn from volumes 20 and 21. Of these, 11 papers had first round reviewer reports available and are used in the quantitative comparison; the remaining 10 either had reports only from later revision rounds or no publicly indexed reports at all. Table 2 shows the distribution across the four PeerLens verdict categories on the full 21 paper corpus, and the agreement pattern against the 11 paper ground truth subset.

The agreement rate of 45%, combined with a 9% more lenient rate, indicates that PeerLens is not systematically overcautious. In every case where PeerLens returned *minor revision* while the SciPost reviewers said *publish*, the PeerLens

Category (ground truth subset, n=11)	Count
Agree (same verdict category)	5 (45%)
PeerLens stricter	5 (45%)
PeerLens more lenient	1 (9%)

Table 2: PeerLens verdict against first round SciPost Physics reviewer recommendations, on the 11 paper subset with available v1 reports.

eLife significance (n=14)	Count
fundamental	1
important	8
valuable	4
useful	1
eLife strength (n=14)	**Count**
compelling	4
convincing	4
solid	5
incomplete	1
PeerLens verdict (n=14)	**Count**
minor revision	12
major revision	2
accept, reject	0

Table 3: PeerLens verdicts and the corresponding eLife editorial assessment tags for the 14 paper biology corpus.

weaknesses were substantive and would not have been rejected by an expert human referee. As one representative example, SciPostPhys.20.2.063 addresses the interplay between gauge invariance and entanglement in gauge theory. The SciPost reviewers requested minor revisions. PeerLens agreed on *minor revision* but its rationale flagged a specific technical concern: the deformation of the four gluon vertex by an explicit symmetry breaking parameter k violates Ward identities, meaning the observable being computed (entanglement concurrence) is being evaluated on a theory with unphysical propagating longitudinal polarization states. This is exactly the kind of concern a specialist reviewer would raise on a first pass. That the SciPost reviewers ultimately let the paper through does not mean the concern is invalid; it means the paper was judged acceptable subject to minor changes, of which this issue could plausibly be one.

4.2 Biology: eLife

We ran PeerLens on 14 eLife Reviewed Preprints spanning subjects that eLife covers. Because eLife uses two dimensional editorial assessment tags rather than a single verdict, we present the results as a joint distribution table (Table 3) rather than as an agree/disagree split.

Reading Table 3 against the paper level detail, PeerLens' *minor revision* verdict is broadly consistent with eLife's editorial signals. Twelve of the fourteen papers received *important* or *valuable* significance tags with *solid* or better strength; these are papers that would generally clear peer review with corrections, and PeerLens' minor revision verdict matches that direction. The two *major revision* verdicts (elife_98357 and elife_111930) both reflect substantive technical gaps rather than global rejection. In elife_98357, PeerLens flagged that the paper proposes a non canonical adapter/scaffold role for CDK2 in TBK1 ubiquitination without ever testing whether CDK2's canonical serine/threonine kinase activity is required for the effect. This is the first question a specialist referee in molecular immunology would ask, and its absence is a clean methodological gap. The eLife editors marked the paper *valuable* and *convincing*, but PeerLens' concern is orthogonal to acceptance: even a valuable and convincing paper can have an unresolved mechanistic question that a diligent reviewer would flag.

More strikingly, elife_106791 received eLife's highest joint tag pair, *fundamental* plus *compelling*. PeerLens still returned *minor revision* and identified three field standard concerns: the proximity ligation assay used to argue for direct co assembly of ion channels reports only proximity within a 40 nm radius (not direct binding), the observed 18 hour and 24 hour surface expression lag between the two proteins is inconsistent with an obligatory co assembly model, and the transcript colocalization is not exclusive to the claimed channel pair. These are the kinds of methodological limits that appear in specialist reviewer reports of otherwise landmark papers; their identification here is a demonstration that PeerLens can flag substantive limits even in papers that reviewers have judged to be at the top of the field.

4.3 Mathematics: F1000Research

The mathematics corpus consists of 20 F1000Research articles tagged under Mathematics: an initial 3 paper pilot plus 17 articles systematically sampled from the F1000Research browse listing of that subject. All 20 papers had at least one reviewer report at ingest, and PeerLens' verdicts are compared against the aggregated per reviewer recommendation.

The mathematics corpus produces the largest gap between PeerLens and human reviewers. F1000Research reviewers issued a *not approved* recommendation on only 5 of 20 papers, and *approved with reservations* on 12 more; PeerLens' verdicts skew significantly stricter. The gap is not a calibration artifact. Manual inspection of the PeerLens reports found that in every one of the 17 mathematics papers analyzed after the initial pilot, PeerLens identified specific, verifiable technical concerns. None of the reports contained fabricated citations or hallucinated technical claims. Representative examples include:

- **f1000_137430_v3** applies a chi squared test of independence to enrollment and pass rates in Tanzanian secondary schools, but the pass columns are not mutually exclusive (a student can pass multiple subjects) and

PeerLens verdict (n=20)	Count
reject	11
major revision	5
minor revision	4
accept	0
Aggregated F1000Research verdict (n=20)	**Count**
not approved	5
approved with reservations	12
approved	3

Table 4: PeerLens verdicts and aggregated F1000Research verdicts on the 20 paper mathematics corpus.

the contingency table's cell frequencies (5088) do not match the reported grand total (2232). The chi squared statistic is therefore mathematically ill defined. F1000Research reviewers approved the paper with reservations; PeerLens returned *reject*.

- **f1000_172972_v4** claims to establish convergence acceleration for a fixed point iteration on total asymptotically nonexpansive (TAN) mappings. The numerical example uses $\Psi(q) = 12q + 14\sin(q)$, which has derivative 26 at the fixed point $q = 0$ and is therefore highly expansive, not TAN. F1000Research reviewers approved the paper with reservations; PeerLens returned *reject*.

- **f1000_173831_v2** defines a transformation monoid $T_{(Fl)_n(G)}$ as "the set of all morphisms on the free G-act $(Fl)_n(G)$," which is by definition the endomorphism monoid $\mathrm{End}((Fl)_n(G))$. Theorem 5.2.2 then claims to establish an isomorphism between $T_{(Fl)_n(G)}$ and $\mathrm{End}((Fl)_n(G))$, i.e. between a set and itself, and its proof exhibits the identity map. F1000Research reviewers approved the paper; PeerLens returned *reject*.

- **f1000_175837_v2** extends Fermatean fuzzy sets to BN algebras. The manuscript defines the non membership degree as $\theta_F(x) = 1 - \delta_F(x)$, which collapses the Fermatean fuzzy structure back into a classical Zadeh fuzzy set and eliminates the paper's claim to novelty. F1000Research reviewers approved the paper with reservations; PeerLens returned *reject*.

These are not stylistic preferences or overzealous nitpicks. They are foundational errors in probability, in the definition of a mathematical structure, in the statement of a theorem, and in the framing of a novelty claim. That they were missed by the human reviewers, or judged insufficient to warrant rejection, does not make them any less real.

PeerLens verdict (n=9)	Count
minor revision	8
major revision	1
accept, reject	0
Aggregated SciPost verdict (v1, n=6)	**Count**
publish	4
minor revision	2

Table 5: PeerLens verdicts and first round SciPost Chemistry reviewer aggregate on the 9 paper chemistry corpus.

4.4 Chemistry: SciPost Chemistry

We ran PeerLens on all 9 SciPost Chemistry research articles published to date. Of these, 6 have first round reviewer reports and are used for direct comparison. The distribution and per paper verdicts are shown in Table 5.

The chemistry corpus shows the closest overall alignment between PeerLens and human reviewers. Where PeerLens flagged concerns beyond what SciPost reviewers requested, those concerns were substantive. SciPostChem.3.1.001 benchmarks 100 density functionals against high level coupled cluster reference data for dioxygen complexes. PeerLens observed that the paper ranks the functionals using a product metric with units of $\text{Å}^2 \cdot {}^\circ \cdot \text{cm}^{-1}$ (a product of mean absolute deviations in bond length, bond angle, and vibrational frequency), which is dimensionally ill defined and allows a functional with exceptional performance on one axis to dominate the aggregate ranking regardless of its performance on the other axes. This is the kind of statistical scoring error that is trivial to introduce, easy to miss on a routine review, and consequential for the ranking of functionals in downstream computational chemistry. The paper was published in SciPost Chemistry.

Additionally, SciPostChem.1.2.005 presents CrystalNets.jl, an open source Julia library for the topological identification of crystalline materials. PeerLens returned *minor revision* and flagged a methodological mismatch that the SciPost reviewers did not: the library detects symmetries using a floating point library on data that has been carefully placed into exact arbitrary precision rational arithmetic, which sacrifices the correctness guarantee the equilibrium placement was designed to give. The authors themselves note that this floating point step can produce false negatives on unstable nets. Again, this is a concern that a specialist reviewer would raise on a first pass, and it was not raised.

4.5 Cross domain pattern

Table 6 aggregates the four corpora into a single view.

Two patterns stand out. First, agreement is highest in the field where reviewer standards are most institutionally consolidated (biology, through eLife's structured assessment tags and its editorial infrastructure) and lowest in the field where reviewer standards vary most across subdomains (mathematics, where

Field	Corpus size	Ground truth available	PeerLens agrees	PeerLens stricter
Physics (SciPost)	21	11	45%	45%
Biology (eLife)	14	14	86%	14%
Mathematics (F1000)	20	20	20%	80%
Chemistry (SciPost Chem)	9	6	33%	67%
Total	64	51	(see text)	(see text)

Table 6: Cross domain summary. Physics "agrees" counts exact category matches. Biology "agrees" counts cases where PeerLens' verdict is consistent with eLife's joint tag (a paper tagged *important, solid* or better paired with a PeerLens *minor revision* is counted as agreement). Mathematics and chemistry are compared against aggregated per reviewer recommendations from the same round.

F1000Research's Mathematics tag mixes pure mathematics with quantitative social science, mathematical biology, and applied statistics). Second, where PeerLens diverges from the human referees, it diverges by being stricter, and the strictness is anchored in substantive technical concerns rather than in generic hedging language. Across all 51 papers with an available ground truth, we did not observe a case where PeerLens' verdict was driven by a fabricated citation or an invented technical claim.

4.6 On being stricter than human reviewers

An LLM based reviewer that is systematically stricter than human reviewers should be examined carefully rather than dismissed. It could be a calibration defect, in which the model has been prompted or trained to be conservative regardless of the actual state of the manuscript. It could be a hallucination artifact, in which the model has invented concerns that do not survive expert scrutiny. Or it could be an accurate application of formal disciplinary criteria in a setting where human reviewers are systematically lenient.

The case study level review of Sections 4.1 through 4.4 rules out the first two explanations. In every case where PeerLens returned a stricter verdict than the human reviewers, its rationale rested on a concrete, verifiable technical claim: a broken chi squared assumption, an expansive fixed point mapping, a definition that voids its own claimed novelty, a dimensionally malformed benchmark metric, an unresolved mechanistic dependency on kinase activity. Human reviewers are aware of these standards; the standards are simply enforced with a tolerance that varies by reviewer, by round, and by the reviewer's own workload at the moment of review. Reviewer inconsistency has been quantified in the community: the NeurIPS consistency experiments showed roughly 26% disagreement between two independent committees on the same papers, with about half of the observed variation in reviewer scores attributed to subjective sources rather than to genuine differences in the manuscript [6, 7]. Against this background, an automated reviewer that applies formal criteria uniformly is not a defect. It is a shift from a tolerance based enforcement of standards to a criteria based enforcement of standards, and the shift makes PeerLens' verdicts more, not less,

informative about the manuscript.

The design intent of PeerLens is not to review the human authored manuscripts of today. Recent months have seen large language models solve open mathematical problems that had resisted human effort for years, including a counterexample to the 87-year-old Jacobian conjecture produced by mathematician Levent Alpöge in collaboration with Anthropic's Claude Fable 5 model [31, 32], improved bounds on more than 50 open problems in extremal combinatorics, matrix multiplication, and geometric packing via AlphaEvolve [26], silver-medal level performance on the 2024 International Mathematical Olympiad via AlphaProof and AlphaGeometry 2 [33], and increasing use in professional research mathematics itself [34], all under human directed prompting rather than open ended generation. This is not an isolated phenomenon; it is the leading edge of a shift in which human directed AI generation becomes the dominant mode of producing manuscripts in every field that does not require access to a wet lab or an experimental apparatus, including theoretical physics, pure mathematics, chemistry method development, computational biology, phenomenological modeling, data analysis, and survey work. Under that shift, submission volume will grow by orders of magnitude, and the question a peer review system needs to answer will change. The relevant question will no longer be whether a manuscript's incremental extension of prior work is technically defensible. It will be whether the manuscript contains a genuine breakthrough. Under those conditions, a reviewer that applies formal criteria uniformly, at scale, and without the tolerance drift a fatigued human reviewer exhibits, is not an alternative to human review. It is a prerequisite for peer review to continue to exist.

The evaluation in this section is not a claim that PeerLens is ready to replace human reviewers on today's manuscripts. It is a claim that PeerLens' verdicts are grounded in verifiable technical criticism, that where PeerLens diverges from human reviewers it does so on concrete methodological grounds, and that the resulting system is calibrated for the peer review conditions the scientific community is about to face.

5 Discussion

The evaluation in Section 4 establishes three empirical claims about PeerLens on the 64 paper cross domain corpus. First, PeerLens produces referee reports whose weaknesses are, on manual inspection, substantive rather than fabricated. Second, PeerLens is systematically stricter than the human reviewers whose verdicts we compare against. Third, this strictness is anchored in specific technical criteria (a broken statistical test, a definition that voids a novelty claim, a dimensionally malformed metric, an unresolved mechanistic dependency), not in generic hedging. This section discusses what these claims imply, what they do not imply, and what the residual limits of the system are.

5.1 What the results mean

The immediate meaning of the results is that a domain adaptive multi agent design produces referee reports that a working editor can read and act on. This is not the same as producing referee reports that would replace a competent human referee. The two are separable concerns. An editor triaging a manuscript needs a first pass that flags the concrete issues a submission carries, ranked by severity, in a format that a subsequent human referee can start from. The PeerLens output serves that role. Whether the same output could stand alone as the sole review of a manuscript is a further question, and the answer will depend on the field, the venue's tolerance for false positives, and the stakes of the decision.

The results also give a concrete answer to a question that is often asked in the abstract: can a language model based reviewer identify weaknesses that a human referee would identify? On the corpus we evaluated, the answer is yes, with the caveat that the specific weaknesses PeerLens identifies are not always the same weaknesses a specific human referee would have identified on the same paper. This is not a shortcoming unique to PeerLens. Two independent human referees on the same manuscript also identify different weaknesses, and the NeurIPS consistency experiments showed that the overlap between two independent reviewers is on the order of 30% [6]. PeerLens sits within that range with respect to the human reviewers it was compared against.

5.2 What the results do not mean

The results do not establish that PeerLens is calibrated correctly across all subdomains. On the mathematics corpus, PeerLens returned *reject* on 11 of 20 papers, while F1000Research reviewers issued a *not approved* recommendation on only 5. In every one of those cases the PeerLens rationale rested on a concrete technical concern, but a strict threshold that returns *reject* on 55% of a corpus is not directly usable as an editorial decision tool without further calibration against the venue's own acceptance standards. The rate of substantive concerns in the F1000Research Mathematics tag may also be genuinely high, since the corpus mixes formal mathematics with quantitative social science and applied statistics of varying rigor. Either interpretation calls for further work: a per venue calibration study of the verdict threshold, and a longitudinal comparison against the eventual scientific impact of each paper.

The results also do not establish that PeerLens catches all substantive weaknesses. What we can say from the case study review is that in the cases where PeerLens flags a weakness, that weakness is verifiable. The complementary claim, that PeerLens flags every substantive weakness a competent referee would flag, is not established by this evaluation and would require a different experimental design in which multiple expert referees produce independent reviews of the same papers.

Finally, the results do not address the paper's textual authenticity. PeerLens is a reviewer of scientific content; it is not a plagiarism detector, and it is not

a detector of AI generated text as a category. The LLM authorship detection agent in the pipeline reports its judgment as a signal to the editor, not as a filter on the manuscript.

5.3 The upcoming shift and the argument for building this now

The remainder of this section states, more directly than Section 4.6, why we believe a system like PeerLens is needed now rather than at some indefinite future point.

Two shifts are already underway. The first is that AI systems, under human direction, are now solving problems that were open in mathematics and mathematical physics. This is not a claim about autonomous mathematical AI. It is a claim about the concrete productivity of a human researcher who directs an AI system to work on a problem: that pairing produces new results, at a rate that a single researcher working alone cannot match. The recent AI assisted resolutions of open mathematical problems (a counterexample to the 87-year-old Jacobian conjecture [31], breakthroughs in extremal combinatorics, and improvements on bounds on classical objects that had resisted decades of human effort) are the leading indicators [26, 33]. The activity level in AI for mathematics, AI for theoretical physics, AI for computational chemistry, AI for phenomenological modeling in biology, and AI for empirical data analysis is now high enough that this pattern will not remain confined to a small number of high profile results.

The second shift is that the marginal cost of turning a research idea into a written manuscript is falling. A researcher who used to produce three to five papers per year, each requiring months of writing effort, can now produce many multiples of that volume in the same period, because the writing bottleneck (structuring the argument, phrasing the results, drafting the discussion, formatting the references) is no longer the rate limiting step. Combined with the first shift, this means the total volume of manuscripts submitted to journals will not merely grow linearly with the number of researchers. It will grow with a coefficient that reflects the per researcher throughput gain.

Under these two shifts, the reviewer pool cannot keep up. The 2018 Publons report already documented an increasing gap between reviewer demand and reviewer supply [3]; the 2025 Nature feature described the operational strain in more direct terms [5]. Neither of those characterizations includes the coming inflow. Once the inflow lands, a peer review system that depends entirely on human reviewer capacity will not merely be slow. It will functionally fail to review submissions at all, and the venues will fall back on editorial desk rejections that do not use scientific criteria.

The question at that point is not whether AI review will happen. It is whether AI review will be done well, by systems that were designed for the task and validated openly, or whether it will be improvised by editors under time pressure using general purpose chat interfaces. PeerLens is our attempt at the first option: a system that was designed to be a reviewer from the start,

that runs specialized agents, that operates uniformly across four scientific fields, and that produces artifacts an editor can act on. What the system needs now is not more agents. It is scale (validation on thousands rather than tens of papers, ideally in partnership with venues that already run open peer review), operational hardening (calibration per venue, integration with existing editorial workflows, uptime and rate control at the model API layer), and expert audit (blinded expert review of the PeerLens outputs against the same papers' human reviewer outputs, at a scale that supports statistical conclusions rather than illustrative case studies).

We view this evaluation as the point at which it becomes reasonable to look for those partners and that investment.

5.4 Ethical and operational limits

Any AI review system carries ethical and operational risks that deserve explicit discussion.

The first risk is that the system's outputs are trusted more than the evidence warrants. On the case studies we reported in Section 4, PeerLens' concerns were substantive; but a system with a good record on 64 papers can still be wrong on the next paper, and the venue that integrates PeerLens must maintain a policy in which the system's output is treated as one input to an editorial decision rather than as the decision itself. This is the same policy every venue already applies to individual reviewers.

The second risk is that a system that is systematically stricter than human reviewers, if deployed as a filter, will reject manuscripts that would have been accepted through the current process. If the stricter judgment is grounded in verifiable technical criteria, that is a feature; but if the criteria are miscalibrated relative to a specific field's norms, the same behavior becomes a defect. Per venue calibration and expert audit, as described in Section 5.3, are the appropriate mitigations.

The third risk is that reviewer accountability shifts in a way that no participant is comfortable with. If a human reviewer relies on a PeerLens report and produces a shallow follow up, and if a downstream decision then rests on the combined output, it is not immediately clear who is responsible for a missed weakness. We recommend that venues that deploy PeerLens require reviewers to disclose the assistance they received, so that the trail from evidence to decision remains traceable, and so that the venue's expectations of reviewer effort remain enforceable.

The fourth risk is the specific one raised by the LLM authorship detection agent. A submission that is largely AI generated is not, by itself, a problem, provided the authors were the ones who directed the AI and provided the scientific judgment. But a submission that is AI generated and also not accompanied by a substantive human scientific contribution is a category the community has not yet decided how to handle. PeerLens' LLM authorship signal is offered as an input to that decision, not as a filter.

6 Conclusion

We described PeerLens, a domain adaptive multi agent system that reads a scientific manuscript in PDF form and produces a full referee report. The system runs eight specialized agents (classification, claim extraction, dimensional analysis, symbolic verification, novelty search, structure and narrative review, domain expert review, and reproducibility review), each with a schema, prompt, or persona tuned to the manuscript's field, and aggregates their outputs into a single referee report in the format a working reviewer would submit.

We evaluated PeerLens on a 64 paper cross domain corpus drawn from four open peer review venues (SciPost Physics, eLife Reviewed Preprints, F1000Research Mathematics, and SciPost Chemistry). On the 51 papers for which first round human reviewer reports were available, PeerLens' verdicts either agreed with the human reviewers or diverged by being stricter. In every case where PeerLens returned a stricter verdict, its rationale rested on a concrete, verifiable technical concern (a broken statistical assumption, an expansive fixed point mapping, a definition that voids its own novelty, a dimensionally malformed metric, an unresolved mechanistic dependency), and we did not observe fabricated citations or invented technical claims across the corpus.

The evaluation supports three conclusions. First, a domain adaptive multi agent design is a viable path to LLM based peer review, distinct from and outperforming the single prompt, single output design used in most prior LLM review studies. Second, the strictness that a well calibrated LLM reviewer applies is not a bug: it is a shift from tolerance based enforcement of disciplinary criteria to criteria based enforcement, and its outputs are more informative about a manuscript, not less. Third, the throughput and volume conditions the peer review system is about to face (rising AI assisted submission volume, an increasing share of manuscripts in theory heavy, computation heavy, and analysis heavy fields, and a reviewer pool that is already visibly stretched) make a domain adaptive LLM reviewer a necessary component of scholarly infrastructure rather than a speculative one.

PeerLens is a working system: the referee reports discussed in Section 4 were produced by the deployed pipeline described in Section 3, and a public demonstration is available at https://peerlens-demo.fly.dev. What the system needs next is scale, per venue calibration, and expert audit, and we welcome collaboration with venues, funders, and infrastructure providers who share the view that peer review is worth defending under the conditions ahead.

References

[1] Mark A. Hanson, Pablo Gómez Barreiro, Paolo Crosetto, and Dan Brockington. The strain on scientific publishing. *arXiv preprint*, arXiv:2309.15884, 2023. doi: 10.48550/arXiv.2309.15884. URL https://arxiv.org/abs/2309.15884. Later published in *Quantitative Science Studies* 5(4): 823–843, 2024.

[2] arXiv. arXiv sets new record for monthly submissions. arXiv Blog, November 2023. URL https://blog.arxiv.org/2023/11/03/arxiv-sets-new-record-for-monthly-submissions/.

[3] Publons. 2018 global state of peer review report. Technical report, Clarivate Analytics, 2018. URL https://publons.com/community/gspr.

[4] Editors' Cafe. The peer review crisis: Delays, bias, and burnout, May 2025. URL https://editorscafe.org/details.php?id=68.

[5] David Adam. The peer-review crisis: How to fix an overloaded system. *Nature*, 644(8075):24–27, 2025. doi: 10.1038/d41586-025-02457-2. URL https://www.nature.com/articles/d41586-025-02457-2.

[6] Corinna Cortes and Neil D. Lawrence. Inconsistency in conference peer review: Revisiting the 2014 NeurIPS experiment. *arXiv preprint*, arXiv:2109.09774, 2021. doi: 10.48550/arXiv.2109.09774. URL https://arxiv.org/abs/2109.09774.

[7] Alina Beygelzimer, Yann Dauphin, Percy Liang, and Jennifer Wortman Vaughan. The NeurIPS 2021 consistency experiment. NeurIPS Blog, December 2021. URL https://blog.neurips.cc/2021/12/08/the-neurips-2021-consistency-experiment/.

[8] Michaël Kovanis, Raphaël Porcher, Philippe Ravaud, and Ludovic Trinquart. The global burden of journal peer review in the biomedical literature: Strong imbalance in the collective enterprise. *PLoS ONE*, 11 (11):e0166387, 2016. doi: 10.1371/journal.pone.0166387. URL https://doi.org/10.1371/journal.pone.0166387.

[9] Richard Van Noorden and Jeffrey M. Perkel. AI and science: What 1,600 researchers think. *Nature*, 621:672–675, 2023. doi: 10.1038/d41586-023-02980-0. URL https://www.nature.com/articles/d41586-023-02980-0.

[10] Elsevier. Insights 2024: Attitudes toward AI. Technical report, Elsevier, 2024. URL https://www.elsevier.com/insights/attitudes-toward-ai.

[11] Oxford University Press. Researchers and AI: A global survey. Technical report, Oxford University Press, 2024. URL https://academic.oup.com/pages/ai-survey-findings.

[12] Weixin Liang, Yaohui Zhang, Zhengxuan Wu, Haley Lepp, Wenlong Ji, Xuandong Zhao, Hancheng Cao, Sheng Liu, Siyu He, Zhi Huang, Diyi Yang, Christopher Potts, Christopher D. Manning, and James Y. Zou. Mapping the increasing use of LLMs in scientific papers. *arXiv preprint*, arXiv:2404.01268, 2024. doi: 10.48550/arXiv.2404.01268. URL https://arxiv.org/abs/2404.01268.

[13] Weixin Liang, Yuhui Zhang, Hancheng Cao, Binglu Wang, Daisy Y. Ding, Xinyu Yang, Kailas Vodrahalli, Siyu He, Daniel Scott Smith, Yian Yin, Daniel A. McFarland, and James Zou. Can large language models provide useful feedback on research papers? a large-scale empirical analysis. *NEJM AI*, 1(8), 2024. doi: 10.1056/AIoa2400196. URL https://ai.nejm.org/doi/full/10.1056/AIoa2400196.

[14] Alexander Goldberg, Ihsan Ullah, Thanh Gia Hieu Khuong, Benedictus Kent Rachmat, Zhen Xu, Isabelle Guyon, and Nihar B. Shah. Usefulness of LLMs as an author checklist assistant for scientific papers: NeurIPS'24 experiment. *arXiv preprint*, arXiv:2411.03417, 2024. doi: 10.48550/arXiv.2411.03417. URL https://arxiv.org/abs/2411.03417.

[15] Nitya Thakkar, Mert Yuksekgonul, Jake Silberg, Animesh Garg, Nanyun Peng, Fei Sha, Rose Yu, Carl Vondrick, and James Zou. A large-scale randomized study of large language model feedback in peer review. *Nature Machine Intelligence*, 8:326–336, 2026. doi: 10.1038/s42256-026-01188-x. URL https://www.nature.com/articles/s42256-026-01188-x.

[16] AAAI-26 Program Committee. AI-assisted peer review at scale: The AAAI-26 AI review pilot. *arXiv preprint*, arXiv:2604.13940, 2026. doi: 10.48550/arXiv.2604.13940. URL https://arxiv.org/abs/2604.13940.

[17] Giuseppe Russo Latona, Manoel Horta Ribeiro, Tim R. Davidson, Veniamin Veselovsky, and Robert West. The AI review lottery: Widespread AI-assisted peer reviews boost paper scores and acceptance rates. *arXiv preprint*, arXiv:2405.02150, 2024. doi: 10.48550/arXiv.2405.02150. URL https://arxiv.org/abs/2405.02150. Later published in *Proc. ACM on Human-Computer Interaction*, October 2025.

[18] Żurawicki Krzysztof, Farganus Julia, Arkadiusz Gaweł, Bystroński Mateusz, and Kajdanowicz Tomasz Jan. PRAIB: Peer review AI benchmark of behaviour of LLM-assisted reviewing. *arXiv preprint*, arXiv:2605.29815, 2026. doi: 10.48550/arXiv.2605.29815. URL https://arxiv.org/abs/2605.29815.

[19] Rui Ye, Xianghe Pang, Jingyi Chai, Jiaao Chen, Zhenfei Yin, Zhen Xiang, Xiaowen Dong, Jing Shao, and Siheng Chen. Are we there yet? revealing the risks of utilizing large language models in scholarly peer review. *arXiv preprint*, arXiv:2412.01708, 2024. doi: 10.48550/arXiv.2412.01708. URL https://arxiv.org/abs/2412.01708.

[20] Ruiyang Zhou, Lu Chen, and Kai Yu. Is LLM a reliable reviewer? a comprehensive evaluation of LLM on automatic paper reviewing tasks. 2024. URL https://aclanthology.org/2024.lrec-main.816/.

[21] Tianmai M. Zhang and Neil F. Abernethy. Reviewing scientific papers for critical problems with reasoning LLMs: Baseline approaches and automatic evaluation. *arXiv preprint*, arXiv:2505.23824, 2025. doi:

10.48550/arXiv.2505.23824. URL https://arxiv.org/abs/2505.23824. Presented at NeurIPS 2025 Workshop on AI for Science.

[22] Gaurav Sahu, Hugo Larochelle, Laurent Charlin, and Christopher Pal. ReviewerToo: Should AI join the program committee? a look at the future of peer review. *arXiv preprint*, arXiv:2510.08867, 2025. doi: 10.48550/arXiv.2510.08867. URL https://arxiv.org/abs/2510.08867.

[23] Ngoc Phan Phuoc Loc, Toan Huynh La Viet, Thanh Tran Khanh, Duy A Nguyen, Tuan Anh Nguyen Pham, Thanh Nguyen, Nitesh V. Chawla, Wray Buntine, Kok-Seng Wong, Khoa D. Doan, and Binh T. Nguyen. PRISM: A multi-dimensional benchmark for evaluating LLM peer reviewers. *arXiv preprint*, arXiv:2605.26730, 2026. doi: 10.48550/arXiv.2605.26730. URL https://arxiv.org/abs/2605.26730.

[24] Juraj Gottweis, Wei-Hung Weng, Alexander Daryin, et al. Accelerating scientific discovery with co-scientist. *Nature*, 655:487–496, 2026. doi: 10.1038/s41586-026-10644-y. URL https://doi.org/10.1038/s41586-026-10644-y.

[25] Ali E. Ghareeb, Benjamin Chang, Ludovico Mitchener, et al. A multi-agent system for automating scientific discovery. *Nature*, 655:497–505, 2026. doi: 10.1038/s41586-026-10652-y. URL https://doi.org/10.1038/s41586-026-10652-y.

[26] AlphaEvolve team. AlphaEvolve: A coding agent for scientific and algorithmic discovery. *Google DeepMind Technical Report*, 2025. URL https://deepmind.google/discover/blog/alphaevolve-a-gemini-powered-coding-agent-for-designing-advanced-algorithms/.

[27] Elicit Team. Elicit: The AI research assistant. https://elicit.com, 2026. Commercial platform for AI-assisted literature review and data extraction.

[28] Consensus. Consensus: AI-powered academic search engine. https://consensus.app, 2026. Commercial platform for evidence-based question answering from scientific literature.

[29] Scite. Scite: Smart citations for AI-assisted research. https://scite.ai, 2026. Commercial platform for citation context analysis.

[30] Allen Institute for AI. Semantic scholar: AI-powered academic search. https://www.semanticscholar.org, 2026.

[31] Levent Alpöge. The Jacobian conjecture is false: a counterexample in three complex variables. Public announcement on X (formerly Twitter), July 2026. URL https://x.com/__alpoge__/status/1817041816559853838. Counterexample produced in collaboration with Anthropic's Claude Fable 5 model, disproving the 1939 Keller conjecture for dimensions three and higher; independently verified by Terence Tao and Jared Duker Lichtman.

[32] ScienceDaily. Claude Fable 5 AI finds a tiny formula that topples an 87-year-old math conjecture. August 2026. URL https://www.sciencedaily.com/releases/2026/08/260804034634.htm. News coverage of Levent Alpöge and Claude Fable 5's counterexample to the Jacobian conjecture.

[33] Google DeepMind. AI achieves silver-medal standard solving International Mathematical Olympiad problems, 2024. URL https://deepmind.google/discover/blog/ai-solves-imo-problems-at-silver-medal-level/. AlphaProof and AlphaGeometry 2 systems solving IMO 2024 problems at silver-medal standard.

[34] Terence Tao. Ai in mathematics: recent progress and future directions, 2024. URL https://teorth.github.io/tao-web/ai-views.html. Public discussion of LLM assisted proof formalization and problem solving in professional mathematics; Fields Medalist perspective.

Data Availability

The source code, data processing pipeline, cross-domain validation corpora, and all generated referee reports used in this study are publicly available at the project repository. Specifically:

- **PeerLens source code and pipeline:** publicly available at https://github.com/emredil/peerlens-gemini.
- **Working demonstration:** https://peerlens-demo.fly.dev (password for peer review: `peerlens2026`).
- **Cross-domain validation corpora:** The 64 manuscripts used for evaluation are drawn from four publicly accessible open peer review venues. The specific DOIs and manuscript identifiers for each paper are listed in the Supplementary Material. Source PDFs can be retrieved directly from the respective venues:
  - SciPost Physics: https://scipost.org/
  - eLife Reviewed Preprints: https://elifesciences.org/
  - F1000Research: https://f1000research.com/
  - SciPost Chemistry: https://scipost.org/SciPostChem
- **PeerLens-generated referee reports:** The 64 referee reports produced by PeerLens on the validation corpus, together with their verdicts and per-agent outputs, are available in the project repository under `data/processed/referee_sections/`.

- **Human reviewer reports:** All human reviewer reports used for comparison are freely accessible on the respective venues' websites under open peer review licenses and were retrieved via the venues' public APIs. The ingestion scripts used to retrieve them are available in the project repository.

No restricted or embargoed data were used in this study.

Ethical Approval

This article does not contain any studies with human participants or animals performed by any of the authors. The scientific manuscripts used in the validation corpus are publicly available academic publications; no human subjects data were accessed, collected, or analysed in this study.

Informed Consent

This article does not contain any studies with human participants performed by any of the authors. No informed consent was required.