Gen-Review: A Large-scale Dataset of AI-Generated (and Human-written) Peer Reviews

Anonymous authors
Paper under double-blind review

Abstract

How does the progressive embracement of Large Language Models (LLMs) affect scientific peer reviewing? This multifaceted question is fundamental to the effectiveness—as well as to the integrity—of the scientific process. Recent evidence suggests that LLMs may have already been tacitly used in peer reviewing, e.g., at the 2024 International Conference of Learning Representations (ICLR). Furthermore, some efforts have been undertaken in an attempt to explicitly integrate LLMs in peer reviewing by various editorial boards (including that of ICLR'25). To fully understand the utility and the implications of LLMs' deployment for scientific reviewing, a comprehensive relevant dataset is strongly desirable. Despite some previous research on this topic, such dataset has been lacking so far. We fill in this gap by presenting `Gen-Review`, the hitherto largest dataset containing LLM-written reviews. Our dataset includes 81K reviews generated for all submissions to the 2018–2025 editions of the ICLR by providing the LLM with three independent prompts: a negative, a positive, and a neutral one. `Gen-Review` is also linked to the respective papers and their original reviews, thereby enabling a broad range of investigations. To illustrate the value of `Gen-Review`, we explore several intriguing research questions, namely: if LLMs favor accept-class ratings (they tend to do so) if LLM-written reviews can be automatically detected (so far, they can); if LLMs can rigorously follow reviewing instructions (not always) and whether LLM-provided ratings align with decisions on paper acceptance or rejection (holds true only for accepted papers). `Gen-Review` can be accessed at the following link: https://anonymous.4open.science/r/gen_review/.

1 Introduction

Since the release of ChatGPT in Q4 2022 [57], Large Language Models (LLMs) are revolutionizing many areas of our society [19]. For instance, enormous potential for productivity growth has been reported in fields such as healthcare, software engineering, human-computer interaction, finance, and education, to name a few [35, 14, 46, 32, 13, 37, 79, 41, 77]. From a broader perspective, LLMs are also expected to have a profound *impact on science in general*, regardless of their specific fields [11, 45].

LLMs can affect scientific work in various ways. They can be used to revise text [20], summarize prior literature [7], or implement an experimental pipeline or its parts [28]. The use of LLMs for scientific work has initially faced ample criticism [4, 33, 49]. However, LLMs are a valuable asset to researchers [11, 19] as they can facilitate routine scientific tasks, allowing researchers to focus on the scientific discovery. Consequently, efforts were made to promote a transparent disclosure of the usage of LLMs along the path leading to a scientific publication [2].

A complementary task, integral to the scientific process, is *peer-reviewing*. Some prior works have addressed the subject of using LLMs for peer-reviewing purposes, e.g., [44, 8, 40, 68, 76, 61, 39]. As an almost anecdotal finding, the study of Liang et al. [44] reported that, after the release of ChatGPT, the reviews submitted to the 2024 edition of the International Conference of Learning Representations (ICLR) included a strikingly more frequent (up to 34 times) occurrence of words such as "meticulous" or "intricate", often associated with

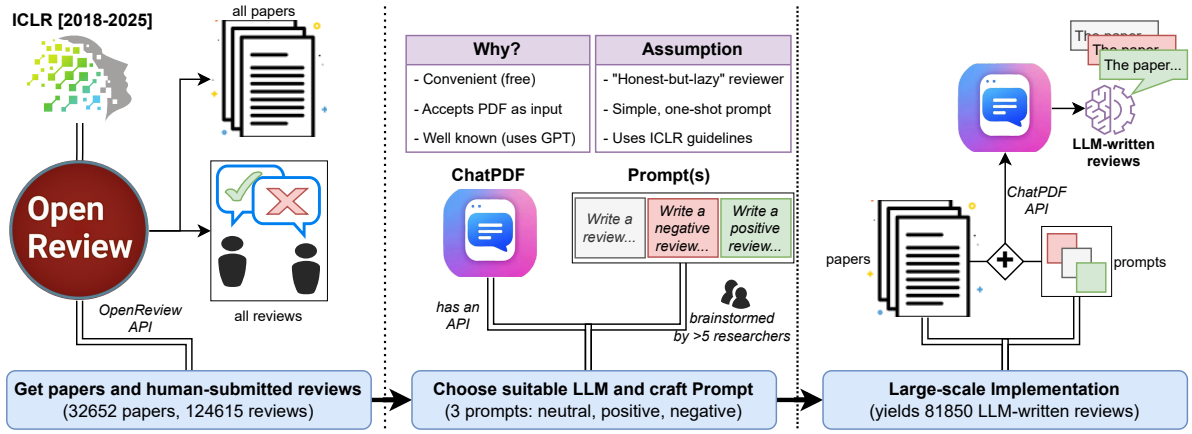


Figure 1: **The workflow to create `Gen-Review`.** We rely on the papers submitted to the [2018–2025] editions of ICLR (we also collect all of their human-submitted reviews). Then, we craft three simple prompts and we leverage the ChatPDF API to generate our large-scale dataset of LLM-written reviews. We then analyse our LLM-written reviews alongside those submitted by human reviewers.

ChatGPT, compared to the previous three ICLR conferences. Such an anomaly suggests that LLMs are likely being used for peer-review at top-tier conferences.

In fact, possibly as a response to the increasing number of papers that require peer-review, some established scientific outlets have started to actively integrate LLMs into their reviewing pipelines. For instance, ICLR'25 used LLMs to provide feedback to a subset of reviewers with suggestions for improving their reviews [80]. As a result, 27% of reviewers confronted with such feedback updated their reviews [67]. Yet, the overall sentiment towards a large-scale deployment of LLMs for reviewing remains mixed, with opinions ranging from "inevitable" to "a disaster" [50].

In light of such diverging opinions, it becomes apparent that the discourse on the impact of LLMs on scientific reviewing must be supported by fundamental data-driven research. To facilitate such research, we present `Gen-Review`, the hitherto largest publicly-available dataset of LLM-generated reviews. It contains over 80 thousand reviews generated for *all papers* submitted to the ICLR between 2018 and 2025. For each paper, three reviews were generated by issuing three independent prompts: one requesting a "positive" review, another requesting a "negative" review, and a "neutral" one without a specific instruction (our workflow is depicted in Fig. 1). We expect `Gen-Review` to foster investigations addressing LLM-driven reviewing, including but not limited to analyzing the potential bias in LLM reviews, gauging their overall quality, measuring the alignment of LLM-reviews with human-authored ones, and evaluating detectors of LLM-generated content. We illustrate the potential benefits of `Gen-Review` for such research by carrying out exemplary investigations. Specifically, after collecting all the human-submitted reviews for the same editions of the ICLR (which we provide in our dataset), we: *(i)* compare the LLM-proposed recommendation with the human-driven papers' outcome; *(ii)* investigate the presence of bias in our LLM-written reviews; and *(iii)* test state-of-the-art detectors of LLM-generated text on our collected data.

Contributions. In summary, our paper makes two key contributions:

- We create `Gen-Review`, a large-scale dataset of over 80k LLM-written reviews, related to over 32k papers submitted to the [2018–2025] editions of the ICLR.
- We use our curated data to provide quantitative insights related to the utilization of LLMs for scientific peer-review.

This paper is organized as follows. First, we define our scope and justify the need for our contributions in Section §2. We then describe the creation of `Gen-Review` in Section §3. Exploratory analyses are elucidated in Section §4. Afterwards, we expand our contributions with a small-scale experiment focusing on additional LLMs in Appendix C. We discuss our results in Section §5 and provide avenues for future work in Section §6.

2 Preliminaries, Goals, and Motivation

In this section, we outline the context of our work, which also serves to substantiate some design choices (§2.1). Then, we outline our research goals (§2.2) and compare our contributions with related work (§2.3).

2.1 Background and Context

We summarize the landscape of using Artificial Intelligence (AI), such as LLM, for content generation. Then, we focus on the core of our work, emphasizing the relevance and necessity of similar efforts.

Generative AI and LLMs. One of the most appreciated capabilities of LLMs is their content-generation ability. An LLM can interpret the instructions embedded in a given *prompt* and produce a corresponding output. Initially, both the prompt and the corresponding output were limited to textual format [57]. However, over time, LLM-related technologies substantially improved, and it is now possible to provide prompts (and requesting an output) as text, images, audio, videos, or a combination thereof [51]. Recent findings have shown that the content generated by modern LLMs is of such a high quality that people can hardly figure out if it is human- or LLM-generated [22, 69, 12, 43].

Detection of AI-generated content. In some contexts (such as in science), determining the author of any given "creation" is of paramount importance (e.g., for authorship, or accountability). Therefore, due to the (allegedly) increasing appearance of LLM-generated content—such as in online social networks [43], or in emails [55]—there has been a growing interest in the development of *automated detectors* of LLM-generated media [65]. Abundant prior works have developed various tools that can estimate whether a given input was generated by an AI (e.g., [36, 10]). For instance, Hans et al. [27] proposed *Binoculars*, an open-source detector that can infer whether a given piece of text was generated by, e.g., ChatGPT, with an accuracy of over 90% and a false-positive rate of only 0.01%. Unfortunately, attaining complete certainty on the true author of any given content is still an open problem: as stated in a recent survey [73], there is "an urgent need to strengthen detector research."

LLM-assisted generation of scientific peer-reviews. As acknowledged by the organizers/editors of various research venues [50, 80], *LLMs are being used today* in the peer-review of scientific articles. However, there are many ways in which LLMs can be used in this process [25]. For instance, LLMs can take an existing review (or parts thereof) and improve its writing quality, or check that the review is written constructively and respectfully; LLMs can also provide a short and high-level account on a work referenced in a given submission; finally, LLMs can also write an entire review on the reviewers' behalf. Such a task can be carried out by *(i)* issuing a prompt such as "write a review on this paper" and *(ii)* attaching the PDF of the paper to review in the prompt. Doing so would produce an output text of variable length that describes the content of the paper and outlines its strengths and weaknesses—according to the LLM's judgment. For instance, a popular tool to achieve such an objective is ChatPDF:[1] by using its web interface (which is free), it is possible to produce a review of a paper in mere seconds (we provide a screenshot of ChatPDF's Web interface in Fig. 6).

Concerns of AI-generated reviews. Complete reliance on LLMs for reviewing duties raises various concerns, since the LLM's judgment replaces or influences that of the human expert. This can impact both the quality of the scientific selection of published works and the quality of the feedback returned to the authors. Among the most well-known issues of using LLMs for peer-review, we mention: the risk of "hallucinations" that undermine the correctness of the review; the lack of knowledge of the state of the art which prevents assessing the originality/novelty of the paper's claimed contributions; as well as the risk of breaching confidentiality agreements—due to uploading a submitted paper to a third-party. Consequently, certain venues have begun regulating the LLM usage for peer-reviewing purposes (e.g., NeurIPS'25) while others have explicitly prohibited any usage of LLMs in the reviewing process (e.g., CVPR'25). Regardless of whether LLMs are (or not) allowed, *what is crucial is being transparent towards the recipients of the reviews*: the authors have the right to be informed about whether LLMs played a role in the peer-review process of their papers [25].

[1] https://chatpdf.com/, allegedly the #1 PDF Chat AI; ChatPDF relies on the OpenAI GPT models.

2.2 Problem Statement and Research Workflow

At a high-level, our contributions are motivated by two complementary reasons: *(i)* the potentially inescapable integration of LLMs in (parts of) the peer-review process [50], which requires improving our generic understanding of LLM-generated reviews; and *(ii)* the necessity of identifying cases of misconduct wherein reviewers relied on LLMs without disclosure (thereby failing to uphold the authors' right to be informed [25]), which calls for ad-hoc detectors of LLM-generated reviews.[2]

Therefore, our first goal is the creation of a large-scale dataset of LLM-generated reviews, i.e., `Gen-Review`. We do this by using all paper submissions to the last eight editions of the ICLR. We elect to use ICLR papers as the core of the dataset and analysis not only because of their public reviews, but also because all ICLR submissions (including rejected or withdrawn papers) are publicly available. Crucially, this enabled us to create a dataset that is based on a large variety of papers in terms of quality (i.e., a dataset whose reviews are based solely on accepted papers would not be well-suited for research on the capabilities of LLMs in assisting in the peer-review).

Our workflow is depicted in Fig. 1 (further discussed in §3). Upon taking all the 32'652 papers submitted to the last eight editions of the ICLR (i.e., 2018–2025), we use ChatPDF to generate three reviews per paper, each based on an independent one-shot prompt: *(a)* a "positive" prompt, specifically crafted to induce the model to recommend an accept-class score; *(b)* a "negative" prompt, crafted to induce the model to recommend a reject-class score; and *(c)* a "neutral" prompt, wherein we do not add any explicit instruction on the (LLM-provided) recommendation. This led to the generation of 81'850 LLM-written reviews. Next, we collect all the human-submitted reviews (124'615 in total) for our sample of papers. Finally, we use all of this data to answer four research questions (RQ):

- RQ1: *Is there any intrinsic bias in the LLM-written reviews?* (i.e., what is the general score distribution of "neutral" reviews w.r.t. "positive" and "negative" ones?)
- RQ2: *How much do "neutral" reviews align with the overall outcome of the paper?* (e.g., if the LLM recommended accepting the paper, was the paper accepted?)
- RQ3: *How much do LLMs fulfill the instructions provided in the prompt?* (e.g., if we specify a given length for the review, does the LLM follow such a requirement?)
- RQ4: *How well can a state-of-practice and open-source detector (Binoculars [27]) identify the reviews in `Gen-Review`?* (and how does it perform on the human-submitted reviews?)

Given that ChatPDF relies on GPT-4o models, and that our papers are taken from ICLR, the answers to our RQs is restricted to this specific LLM and venue (both being, objectively, very popular at the time of writing [62, 29]). However, in Appendix C we carry out additional analyses on a subset of `Gen-Review`, encompassing also models developed by Google and Anthropic, as well as a commercial (and closed-source) detector of AI-generated text, Pangram [53].

2.3 Related Work

Various prior works have addressed problems related to our contributions. However, to the best of our knowledge, no existing dataset has a scope comparable to `Gen-Review`, and our findings are also original. In what follows, we summarize and compare the most related works to this paper.

Lack of ground truth. The findings of Liang et al. [44] indicate that LLMs are likely to have been used in ICLR'24. However, there is no ground truth to verify if any given review with an anomalous utilization of certain terms (e.g., "meticulous") was indeed written by an LLM. Moreover, without such ground truth, it is also impossible to determine the extent to which an LLM has been used (e.g., was it used to generate the entire review, or only to improve the textual quality of a human-written review?). The same shortcoming (i.e., lack of ground truth) also affects the work by Latona et al. [40], where GPTZero was used on the reviews submitted to ICLR'24, finding that potentially 15% were written with AI assistance. We address this problem by constructing a large-scale dataset of LLM-generated reviews, where the level and nature of AI involvement are fully controlled. Therefore, our dataset represents a valid proxy for a wide range of investigations.

[2] Ideally, such detectors can be used *before* the authors receive the LLM-generated reviews, so that action can be taken before making a (potentially inappropriate) decision on the paper's outcome.

Small-scale analyses. Thelwall et al. [68] assess ChatGPT's ability to predict the outcome of some papers submitted to ICLR'17 (collected in [31]); Vasu et al. [70] seek to find hidden biases in the reviews of LLMs, but only consider 126 papers from ICLR'25. The authors of [61] compared the assessments of human reviewers to those of GPT-4 across 325 abstracts, finding alignment only for the best submissions; whereas Du et al. [18] and Collu et al. [15] generate a dataset of LLM-written reviews based on a set of ≈20 papers. The datasets (and corresponding analyses) of all these works are of a much smaller scale than ours. Our contributions seeks to provide a foundation for large-scale analyses.

Limited-scope datasets of LLM-written reviews. The closest works to our paper are those of Yu et al. [76], Kumar et al. [39], and the just-accepted NeurIPS'25 paper by Zhang et al. [78]. All such works entailed the generation of LLM-written revriews based on submission to top-tier ML venues, but the datasets have a much smaller scope than our proposed `Gen-Review`. For instance, Yu et al. [76] generate the reviews by selectively removing some parts of the papers (such as the bibliography and images), and even though the reviews (16K in total; we have 81K) are based on papers submitted to the ICLR from 2021–2024, the overall number of papers used as a basis is only 500; Zhang et al. [78] carry out a broader analysis across 7.1K papers (mostly from ICLR'23–24), but our `Gen-Review` is much larger with 32K papers from *all* editions of ICLR since 2018. Kumar et al. [39] also use a much smaller number of papers (i.e., 1480 in total, taken from ICLR'22 and NeurIPS'22) and the reviews are generated by providing only the paper's text (i.e., without images) as input to the prompt. In contrast, our reviews are generated by providing the entire PDF, ensuring that the LLM has access to all the information available to any human reviewer.

Orthogonal works. There are also orthogonal works that propose datasets of various AI-generated content—not necessarily peer-reviews—such as [64, 16, 74]; or works that focus on the detection of LLM-written *papers*—and not reviews—such as [49]. Finally, we stress that our work is in no manner related to the detection of "fake reviews" in online platforms (e.g., online marketplaces [34, 58]).

3 `Gen-Review`: Large-scale Dataset of Peer Reviews

We describe the creation process of our major contribution: the `Gen-Review` dataset. Our workflow (shown in Fig. 1) can be split in three phases, which we elaborate on in the remainder of this section.

3.1 Preparation: retrieving papers and human-submitted reviews

Requirements. To create a dataset of LLM-written peer-reviews, we need research papers—ideally (dozens of) thousands, since we aim to provide a dataset that enables large-scale assessments. Moreover, to provide a dataset that allows *fair* evaluations of LLM-written peer-reviews, we need papers that have been either "accepted" or "rejected": indeed, using only "accepted" papers would prevent one from gauging the quality of LLM-written reviews for those papers (theoretically of lower quality) that were not accepted to a given venue—which typically represent a large share of the submissions. Finally, we must ensure that our dataset includes also human-submitted reviews, necessary to facilitate comparison against LLM-written ones.

Collection. We determined that ICLR is the most suited venue that fulfills all of the aforementioned requirements. Aside from being a top-tier venue, it yearly receives thousands of submissions; moreover, the complete peer-review details (including each human-submitted review, as well as outcome) of each submission are publicly observable—and there is historical data available on OpenReview for all of its editions. Therefore, we used the OpenReview API to collect all relevant data for our purposes for each paper submitted to ICLR from 2018 to 2025 (8 editions in total). In this way, we obtained: 32'652 papers (spanning accepted, rejected, and even withdrawn papers) and 124'615 human-submitted reviews (including their text, recommendation, and confidence). We do not consider submissions to satellite events of ICLR (e.g., workshops or blogposts). We note that such a process complies with OpenReview's terms of use (`https://openreview.net/legal/terms`).

3.2 Design choices: selecting the LLM, and crafting the prompts

The second step involves determining which LLM to use to generate our reviews, as well as devising prompts that would make `Gen-Review` appealing for future research. To better appreciate our contributions, we must

first describe our underlying assumption. Indeed, there are virtually infinite ways to craft a prompt that asks an LLM to "review a paper", and there are also dozens (or hundreds) of LLMs that can be leveraged for such a task. Therefore, to create `Gen-Review`, we set ourselves the goal to mimic a realistic and likely common use case. Specifically, we asked ourselves: "*If I were a reviewer tasked to write a review for a paper (submitted to ICLR) and I had no time to accomplish such a task, what would be the best way to do so by leveraging LLM-based solutions?*" Essentially, we assumed the perspective of an "honest-but-lazy" reviewer, who wants to fulfill their reviewing duties but does not have enough time to do so properly, and hence decides to rely on an LLM. This is a sensible assumption, given the increasing reviewing load in many research domains [50].[3]

LLM-solution of choice: ChatPDF. The first decision that our envisioned reviewer must make is which LLM to use. From this viewpoint, the ideal solution is one that fulfills the following criteria: *(i) it is convenient*—our reviewer does not want to spend money (e.g., to use more sophisticated models) or time (e.g., to setup a local model); *(ii) it is simple to use*—our reviewer just wants to write a prompt and provide the paper as-is, i.e., without converting the PDF into other formats; *(iii) it is well-known*—given that no LLM is intrinsically perfect, the reviewer (being a scientist) wants to resort to a solution for which there is evidence that it is "good enough" to carry out such a task. We found that ChatPDF is a solution that fulfills all of these criteria. Specifically, ChatPDF is free and is provided with a Web interface (even users who are not logged in can use it); it enables PDF upload by default[4], and it is popular, since it relies on state-of-the-art GPT models. Finally, and crucially (for the sake of feasibly creating `Gen-Review`), *ChatPDF provides an API that allows to scale our workflow.* Put simply, ChatPDF was the best viable option for our goals, motivating our choice (we note that, to create `Gen-Review`, we had to purchase thousands of API queries).

Devising our prompts. Our envisioned reviewer must also determine which prompt to use. Being time-pressured, the reviewer would opt for something simple, i.e., a prompt that does not include any remark about what parts of the paper to mention in the review. The reviewer would, however, provide the generic guidelines of ICLR, since this would enable aligning the LLM-written review with the expectations of the considered venue. Furthermore, the reviewer would not try to craft a prompt that, e.g., seeks to "evade" detectors of LLM-generated content (if he/she wants to do so, they can take the output and modify it accordingly). Additionally, being "honest", the reviewer would not introduce any specific instruction about whether to accept or reject the paper. Finally, the prompt must be context-agnostic: the reviewer is not willing to engage in a long conversation with the LLM to derive the "perfect review". Therefore, to craft a prompt that resembles such a use case, more than five researchers collectively brainstormed and discussed various alternatives. We ultimately converged to the prompt reported in Prompt 1. In our prompt, which has a somewhat similar structure to that used by [39] (i.e., a summary of the paper, followed by a main review), we have added constraints on the length of the review (i.e., the summary and the review should be [100–300] and [800-1000] words in length, respectively). We have also integrated common elements taken from the CFP of each considered edition of ICLR. Finally, to enable assessment of bias in the LLM reasoning, and also to simulate a slightly different use case of a "not-very-honest" reviewer, we created two variants of our prompt: a "positive" (in Prompt 2) and a "negative" (in Prompt 3) one. We note that these two alternatives are identical to the "neutral" version, with the only difference being the word "POSITIVE" (or "NEGATIVE") mentioned twice in the respective prompt.

3.3 Implementation: overall statistics, and development challenges

The last step involves using the API provided by ChatPDF to interact with the underlying LLM[5] by providing *(i)* each of our retrieved papers alongside *(ii)* all of our prompts as input.

Overview. Specifically, for each of our 32652 retrieved papers, we use (in independent contexts) each of our three prompts, thereby generating three reviews per paper—a neutral-prompted one, a positive-prompted one,

[3]We stress that **we do not take any stance on the ethical or moral implications** of *(a)* using LLMs as a potential "shortcut" for carrying out peer-reviewing duties, or *(b)* the act of uploading papers to a third-party LLM service. Our sole intent is to create a dataset for the investigation of various aspects of LLM reviewing.

[4]At the time of designing our pipeline (i.e., November 2024) not many models enabled interacting with a PDF file "as-is" and for free (e.g., for OpenAI, this feature was added only in December 2024 [51])

[5]We issued our queries between Feb.–Apr. 2025: according to the ChatPDF documentation, the queries were routed to models of the GPT-4o family. No change was made to ChatPDF during our considered time frame.

Table 1: **`Gen-Review` in a nutshell.** For each submitted paper (after fetching all of its human-submitted reviews) we generate three LLM-written reviews using ChatPDF by issuing three prompts.

ICLR Edition		2018	2019	2020	2021	2022	2023	2024	2025	Total
Paper Submissions		935	1419	2213	2594	2618	3797	7404	11672	**32652**
Hum.-sub. Reviews		2784	5751	6721	10022	10206	14355	28028	46748	**124615**
GenAI Reviews	Neutral	929	1398	2181	2542	2544	3686	5361	8378	**81850**
	Positive	928	1397	2176	2541	2544	3686	5361	8377	
	Negative	928	1397	2176	2541	2544	3686	5361	8378	

and a negative-prompted one. Ultimately, we obtained 81'850 LLM-written reviews, representing the core contribution of `Gen-Review`. To facilitate downstream usage, each LLM-written review in `Gen-Review` has an identifier that enables to easily discern *(a)* the paper that refers to such a review, and *(b)* the human-submitted reviews available on OpenReview. Table 1 reports overall statistics of `Gen-Review`.

Challenges. We encountered various challenges: First, ChatPDF does not allow interaction with PDF files that are larger than 32MB, which led us to discard 695 papers in total. Moreover, after we collected our data, we inspected it and we found that some reviews were truncated—likely due to network errors (which were not unexpected, given our massive usage of the ChatPDF API). While we tried to sanitize all of these occurrences by reissuing the API query, we acknowledge that some LLM-written reviews in `Gen-Review` may still present some inconsistencies.

4 Analysis and Original Findings

We now analyze our proposed `Gen-Review` dataset by answering our four RQs (see §2.2).

RQ1: Biases of our LLM-written Reviews. To answer RQ1, we compare the scores embedded in each LLM-written review in `Gen-Review` for each of the three prompts we considered. We expect that "negatively-prompted" reviews have scores below the typical acceptance bar ($\leq$5 for ICLR), whereas "positively-prompted" reviews will have scores above the acceptance bar ($\geq$6). However, we do not know what to expect from the "neutral-prompted" reviews. We show the score distribution in Fig. 2; here, a score of 0 indicates that we could not extract any score by employing pattern-matching techniques (the low-level implementation is provided in our code repository), which occurs for 291 LLM-written reviews out of 81850 (0.4%). *There is a substantial bias in LLM-written reviews, which tends to favor a positive outcome.* Particularly, for the neutral-prompted reviews, only 35 AI-generated reviews use the score "5: slightly below the acceptance threshold". All other neutral-prompted reviews deemed the respective paper to be above the acceptance threshold; perhaps surprisingly, the most common rating was that of "8: Top 50% of accepted papers, clear accept". To slightly reinforce the positive bias, we also observe that *(i)* although all negative-prompted reviews do indeed have a reject-class rating, the wide majority has a "4: Ok, but not good enough - rejection"; whereas *(ii)* positive-prompted reviews almost always are rated with an 8 or "9: Top 15% of accepted papers, strong accept" (only two LLM-written reviews rate the paper with a 7). These findings indicate that although the LLM seems to follow our instructions, it does so with an implicit positive bias—a result that echoes recent unpublished work [40].

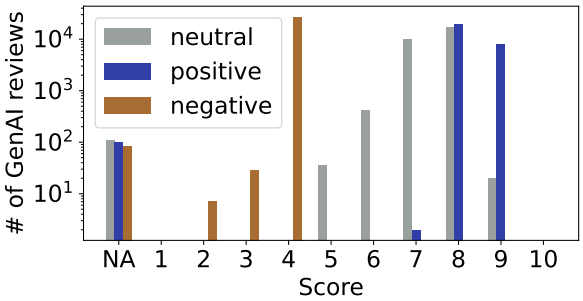
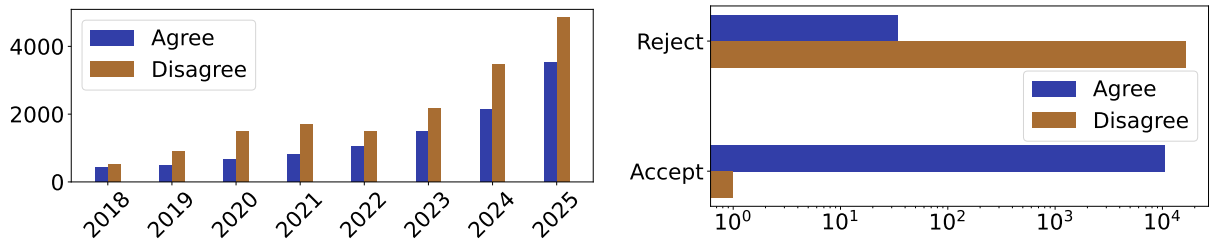
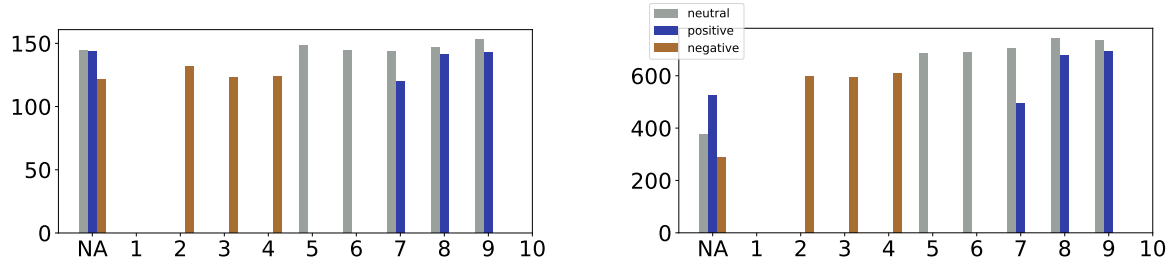


Figure 2: **Rating of LLM-written reviews** in `Gen-Review` for each considered prompt. Ratings follow the ICLR 1–10 scale (N/A denotes cases without a rating in the LLM-written review).



(a) Agreement over the years (y-axis: # of papers). (b) Decision-specific agreement (x-axis: # of papers)

Figure 3: **Agreement between LLM-provided recommendation and human-driven decision for each paper.** We exclude papers that have been "withdrawn" from this analysis.



(a) Length of the "summary" (y-axis: # of words). (b) Length of the "main review" (y-axis: # of words).

Figure 4: **Average length of the LLM-written reviews** for each prompt. The x-axis shows the rating.

RQ2: Alignment of neutral-prompted reviews with human-driven paper's outcome. We investigate the extent to which LLMs can predict the outcome of a given paper. To this end, we take the rating provided by the neutral-prompted reviews in `Gen-Review`, and compare it with the final decision for that paper. Specifically, we consider that the LLM is in agreement if, for a given paper, it recommends a rating $\leq$5 and the paper was rejected; *or* it recommends a rating $\geq$6 and the paper was accepted; we exclude "withdrawn" papers from this analysis. We display the agreement over the years in Fig. 3a, showing that, overall, the LLM's recommendation does not seem to align with the paper's final decision. We further explore this phenomenon in Fig. 3b, showing the decision-specific cases of agreement or disagreement. We can see that the prevalent cases of disagreement entail papers that are ultimately rejected. This finding (which also echoes that of the smaller-scale study in [68]) further reinforces our answer to RQ1: LLMs tend to favor acceptance to a much larger extent than human-driven program committees. We can conclude that *LLMs, being positively biased, cannot reliably predict if a paper will be rejected* (at least to a top-tier venue such as the ICLR).

RQ3: Fulfillment of instructions in the prompt. Our prompts, while simple, embed a variety of constraints and requests. Evidence that LLMs can, to some extent, follow our instructions can already be found in the analysis we did for RQ1: negative-/positive-prompted reviews recommend scores that lean towards rejection/acceptance; however, we were unable to extract the score for 0.35% of reviews—indicating that, in some cases, the LLM either used other words to express a decision, or skipped it entirely. We further analyse the LLM's compliance with our instructions by scrutinizing the length of the "summary" (which should be of 100–300 words, according to our prompt) and of the "main review" (800–1000 words) of the review. To provide a fine-grained analysis, we plot the average length (in words) for each type of prompt and for each rating in Fig. 4a (for the summary) and Fig. 4b (for the main review). While the LLM seem to comply with our requests for the summary (which is typically of 100–130 words), this is not the case for the main body (which hardly goes above 700 words). An ancillary result is that the output length does not vary substantially across ratings. Finally, to explore RQ3 from a different perspective, we study the overall prevalence in the LLM-written reviews of some keywords explicitly mentioned in our prompts (e.g., "strength", "novelty", "clarity"), which the LLM should use to gauge the paper. The results, shown in Table 3 (in Appendix B), reveal that all of our specified terms occur at least once for over 99% of all LLM-written reviews. To conclude, *LLM can generally follow our reviewing instructions, but in some cases they may forget some requests.*

RQ4: Assessment of a AI-generated text detector on `Gen-Review`. Finally, we test how well a state-of-the-art detector of AI-generated text can spot that *(i)* our LLM-written reviews are AI-generated, and we also *(ii)* test its effectiveness on the human-submitted reviews we collected. We consider *Binoculars* [27] due to its popularity (albeit we acknowledge that other tools exist, such as [39]). This detector works by providing a score for the input text, and whether such is above a given threshold (≈ 0.85 that yields 1% false positive rate), the text is deemed as "likely human-generated"; otherwise it is "likely AI-generated". Therefore, we instantiate a local instance of Binoculars and use it to process all of our data—both human-submitted and LLM-written reviews, displaying the results in Fig. 5. We can see that Binoculars works well to pinpoint that our LLM-written reviews are indeed "AI-generated": the recall is 100%. With regard to the human-submitted reviews, we found some instances in which Binoculars predicted the text to be likely AI-generated. We report the occurrence of such "anomalies" across the ICLR editions in Table 2 (in Appendix B). While before 2023 the number of "anomalous" human-submitted reviews is only 1 or 2, this numbers raises to 217 in 2024 and 327 in 2025 (i.e., after the widespread release of LLMs). This result (i.e., the fact that some human-submitted reviews to ICLR may have been AI-generated) echoes the findings of prior work [44, 40]. Unfortunately, due to a lack of ground truth, we cannot claim whether these reviews have been truly AI-generated. Finally, and intriguingly, our analysis showed that Binoculars flagged six human-submitted reviews scattered among the 2019–2022 editions of ICLR: this is surprising, given that no LLMs were publicly available then. Thus, *even though Binoculars is very accurate at identifying genuine AI-generated texts, it may still trigger some false positives.* Therefore, we advise caution in using this tool for detecting LLM-written reviews, as it may lead to false accusations.

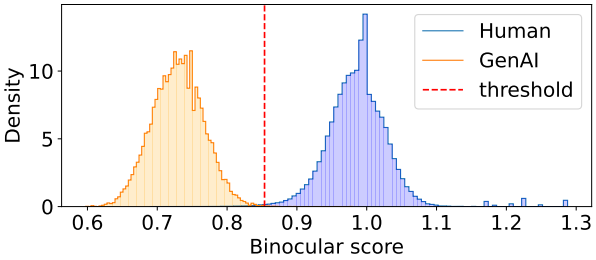


Figure 5: **Assessment of Binoculars** on our AI-generated reviews, and on human-submitted ones.

5 Discussion

We critically examine our research, seeking to provide some explanations, outlining limitations, and identifying how our findings can be leveraged by the scientific community.

5.1 Interpretation of our Findings

In what follows, we make an attempt to interpret the findings stemming from §4 and Appendix C.

Positive bias in LLM-generated reviews. Our results show that ChatPDF and Gemini Flash 2.5 tend to favor "positive" ratings (see §C.1); for ChatPDF, this even happens when prompted to provide a "negative" review (see §4). A potential explanation is that these models have been developed to prioritize providing "helpful" or "encouraging" outputs (see, e.g., [54, 3, 59]), which can be also a byproduct of the so-called Reinforcement Learning with Human Feedback (RLHF) training process of most LLMs during which humans tend to prefer "positive" outputs rather than critical ones [60]. Another possibility is the presence of safety guardrails, which may intrinsically impair generation of outputs of "negative" sentiment (see, e.g., [56]). The fact that Claude Sonnet 4.6 is able to produce some critiques can be due to the increased interest of such form of outputs, which are fundamental to enable the "self-improvement" of LLMs [75].

Non-Compliance with output length. Across our considered LLMs, our results highlight that LLMs do not always follow the instructions of our reviewing prompt in terms of output length. While all LLMs comply with our specified length (100–300 words) in terms of "Summary of the Review", there are stark differences in specification-compliance for the "Main Review" length (which should be 800–1000 words long). Claude Sonnet 4.6 is the most reliable in this regard: the "Main Review" always fall within our specified constraints. In contrast, and intriguingly, the "Main Review" length of ChatPDF is always below our specifications (it is never above 800 words), whereas the length of the "Main Review" for Gemini Flash 2.5 is always above 1000 words in length. A possible explanation for ChatPDF's result is that the LLM interpreted that the

800–1000 words should include both the "summary" and the "main review". Whereas the overly-long reviews by Gemini Flash 2.5 can be explained by the phrasing of our prompt, wherein this request is framed as a "should" (and not a "must"), meaning that the LLM may have preferred complying with other requests rather than following this directive. For Claude Sonnet 4.6, its perfect compliance with our request is indicative of this model being the most recent and powerful among those considered in our analyses.

Detection effectiveness. We found that Binoculars can perfectly detect the reviews of ChatPDF (which relies on GPT-4o models), but struggles with the reviews produced by Gemini Flash 2.5, and does not work against Claude Sonnet 4.6; both of these LLMs, when used to apply some light editing, also allow to bypass Binoculars. A possible explanation for this is that Binoculars [27] was released in early 2024, i.e., when the landscape of available LLMs was not as advanced as the ones used in our additional analyses. For instance, GPT-4o and 4o mini were released in Q1 and Q2 2024, whereas Gemini-Flash 2.5 and Claude Sonnet 4.5 were released in 2025 and 2026: it is hence plausible that Binoculars works well on GPT-4o models but not that well on the others (which have also been used to generate half of our paraphrased reviews). With regards to Pangram, the version we used (3.3.2) correctly classifies all the LLM-generated (or LLM-generated+paraphrased) reviews as being AI-generated. This detector is closed-source, but according to the information provided by Pangram's developers on this specific version,[6] the detector is based on the EditLens architecture discussed in [66] (ICLR'26). According to [66], the underlying detection is carried out by a model trained to detect, among others, text generated by GPT-4.1, Claude 4 Sonnet, and Gemini 2.5 Flash. Hence, it is sensible that Pangram 3.3.2 can recognize the past output of our considered LLMs as being AI-generated text, given that it saw examples of such (or very similar) models during its development phase. Given developments in generated content detection, this may however change in future.

5.2 Limitations

To the best of our knowledge, `Gen-Review` is the largest dataset of LLM-written peer-reviews so far. However, we acknowledge it has some limitations.

First, the reviews in `Gen-Review` only pertain to papers submitted to the ICLR, meaning that our dataset and investigation results may not generalize to other areas outside of computer science—but we never made such a claim. Thisvenue was chosen because it enables a large-scale and fair assessment, since all reviews for its submitted papers (both accepted and rejected ones) are publicly available since its first installment in 2014. In contrast, other venues such as NeurIPS only provide public visibility of review for accepted papers, whereas reviews for rejected papers are available if authors opt-in, which may lead to bias. As a matter of fact, prior work carrying out quantitative analyses of scientific peer reviews also relied on the ICLR for this specific reason (see, e.g., [31, 71, 9, 24, 72]).

Secondly, the reviews in `Gen-Review` have been created by predominantly using ChatPDF, which relied on OpenAI GPT-4o models. We have generated an additional set of 1,600 reviews using LLMs from other vendors (Anthropic and Google), in an attempt to explore if some of our empirical findings hold across different LLMs. While this significantly broadens the diversity of reviews represented in `Gen-Review`, the space of available models remains large and continuously evolving. Therefore, our findings should be interpreted with respect to the set of LLMs included in this study, without drawing conclusions on model families beyond those considered here.

5.3 Broader Impact

In a sense, our findings suggest that our envisioned "honest-but-lazy" reviewer can skew the outcome of the paper selection process if they rely on LLMs that are positively biased (such as those used by ChatPDF, or Gemini Flash 2.5. Further, we have shown that LLMs can be used by a "not-very-honest" (but lazy) reviewer to generate reviews that conform to a desired ("accept" or "reject") outcome with just a single word change to our (very simple) "neutral" prompt. In all such cases, the integrity of the peer-review process is lost, since it is not driven by impartial expert (human) judgment anymore. We note that such "lazy" reviewers are not

[6] https://www.pangram.com/blog/meet-pangram-3-3

uncommon: recent evidence[7] from the ICML'26 shows that ≈400 reviewers instructed an LLM to produce a review and copy-pasted the output with no significant edits.

Fortunately, some existing detectors can (likely with some false positives) flag LLM-generated reviews—particularly, Pangram 3.3.2 seems to be able to do so even after applying fully-automated paraphrasing edits. Nonetheless, from a security standpoint, we endorse taking into account the possibility that some "adversarial reviewers" may attempt to evade the detection process via manual edits, or extensive (automated) editorial changes. For instance, an "adversarial reviewer" who *(i)* is aware of, and *(ii)* has access to, the detector used to flag AI-generated reviews, can iteratively query such a detector with a different version of the LLM-generated review until it is not detected as being "AI generated" anymore. This process, however, has a cost (e.g, Pangram 3.3.2 charges 0.05USD per query of 1000 words, and open-source detectors require computing resources). Ultimately, no detection mechanism is foolproof: the point of security solutions is to discourage adversarial behavior, and not to prevent it entirely [6].

Nonetheless, it is undeniable that LLMs only require seconds to generate a review (e.g., less than 20s were needed by ChatPDF/Gemini/Claude to produce the reviews in `Gen-Review`), whereas human reviewers may take hours [1]. Given the increasing number of papers submitted to scientific outlets, and the increasing difficulty of finding qualified reviewers [63, 23], using LLMs in the peer-review process may become a common practice in some research domains. In our opinion, any such use should be transparent to the affected stakeholders, including the authors—who have the right to be informed of whether an LLM played any role in the peer-review process. In other words, if a reviewer does use an LLM for producing the review, they should clearly state *(i)* that an LLM was used, and *(ii)* the nature and extent of its use (e.g., language editing, drafting, summarization, or review generation). Such transparency would allow the chairs (and other reviewers) to assess the review process and apply appropriate quality-control measures, while ensuring clear accountability for the final review. To maintain the principles of peer review, the reviewer should remain responsible for the accuracy, fairness, and content of the submitted review, irrespective of any LLM assistance.

6 Conclusions and Future Work

Peer review is an essential part of science to ensure the quality of new contributions. It is thus important to understand how new technologies, such as LLMs, may interfere with this process to avoid any harm on science, researchers, or to-be-published works. Our `Gen-Review` can hopefully assist in providing such an understanding. In what follows, we discuss three avenues for future work.

Assessment of additional detectors. Investigating the extent to which LLM-generated reviews can be detected is essential to safeguard the scientific process—especially for those cases in which it is explicitly disallowed to rely on LLMs for peer review (e.g., CVPR'25). Our analyses encompassed Binoculars [27] as a representative open-source detector—but many more detectors of LLM-generated text exist (e.g., [36, 10]). These tools can be tested on the reviews in `Gen-Review` (including human-submitted ones). Particularly, even though we cannot be certain of the "ground truth" of the human-submitted reviews for ICLR 2023–2025, it is safe to assume that reviews submitted for ICLR 2018–2022 (35K in total) are not LLM-written. Hence, our `Gen-Review` can be used as a benchmark to test these detectors. One can also use our dataset to develop ad-hoc detectors for LLM-written reviews (e.g., [39], which we have also tested with a few dozen reviews from `Gen-Review`, and it seem to work very well!).[8] Finally, `Gen-Review` can be used as a basis to create additional variants of "edited" reviews, and gauge how well LLM-text detectors can detect these, too.

Evaluating (and improving) the LLM review quality. We focused on quantitatively analysing, at a high level, the LLM-written reviews in `Gen-Review`, prioritizing investigations reliant on the rating provided by the LLM. Future work can use our data to carry out in-depth analyses to, e.g., scrutinize how accurate the LLM-written review is for each given paper (this is possible given our dataset format), or how much the LLM-written review aligns with the other human-submitted reviews from a content perspective (and not from a rating or decision perspective). For instance, it would be intriguing to explore whether the LLM

[7] https://blog.icml.cc/2026/03/18/on-violations-of-llm-review-policies/

[8] We have also studied (Table 4) the prevalence of the words highlighted by Liang et al. [44] across the LLM-written reviews in `Gen-Review`: many of our reviews include these words, especially "innovative".

provides a factual account of the paper's clarity and significance or if generated reviews contain hallucinations. Answering both of these questions is possible with a paper-by-paper analysis. Finally, developers of LLM can also use our dataset as a baseline to *improve* existing LLMs so that they produce reviews of better quality.

Expanding `Gen-Review`. Despite its large scale, our dataset (and findings) is limited to ICLR and ChatPDF (although we made an effort to test also LLMs from Google and Anthropic). However, to maximize reproducibility and facilitate further research, we have released our prompts. Researchers can thus expand `Gen-Review` in various directions, e.g., using the same prompts by requesting other LLMs to review the same papers; or by using different papers—potentially, from different venues. Put simply, our contributions provide a foundation to investigate the extent to which our findings can map to other disciplines or LLMs.

Ethics statement

As discussed earlier, this paper does not present a position in regards to the "ethical dilemma" of using LLMs for scientific peer-reviewing (see Section 3). Nevertheless, certain aspects of the usage of LLMs can be analyzed (e.g., transparency, utility), and views within the community can be extracted from official policies.

For instance, the ICLR'26 has a dedicated policy on LLMs (see: https://blog.iclr.cc/2025/08/26/policies-on-large-language-model-usage-at-iclr-2026/). From the viewpoint of reviewing, the policy states: "we mandate that reviewers disclose the use of LLMs in their reviews." Such a policy, therefore, does not explicitly disallow using LLMs for peer reviewing. Indeed, in terms of personally observed practices, the authors of this paper *have* received LLM-written reviews in the past (confirmed by an independent investigation of the PC chairs). However, these were not disclosed by the reviewers, and the quality of such reviews was perceived as very limited. Therefore, as we have also stated in Section 2, we urge that every instance of LLM-assisted reviewing (including possibly for this paper submission) is done with proper care, i.e., justified in accordance to policies, and is responsibly disclosed.

Our envisioned "lazy-but-honest" reviewer is a hypothetical (but, we argue, sufficiently realistic) model that we used to create `Gen-Review`. This design choice is not intended to endorse such behavior, but is rather a pragmatic simplification, resulting in a reproducible setting that is still highly informative of broader implications for review quality and integrity. Specifically, a "lazy-but-honest" reviewer is a minimal-effort reviewer who is assumed to issue a one-shot prompt to the LLM. At this stage, the main ethical and integrity issue may be that of a confidentiality violation (if, e.g., the paper is sent to an online service outside the control of the venue, and not to a local model). Yet, subsequent uses of the generated output and their implications may vary. The reviewer can use the output as a starting point to write their own review (thereby being a less "lazy" reviewer), or to compare their own review with that of the LLM, among other options. Based on `Gen-Review`, future research may examine how such varying practices influence review quality and ethical considerations.

We hope that our contributions serve as a foundation to *improve* the overall utility that LLMs can have in the peer-review process, be that benchmarking, methodological reflection, or policy design. Future work can certainly use our dataset to investigate the ethical dimensions of the usage of LLMs for scientific peer-reviewing—but such an objective lies outside the scope of this dataset paper and arguably does not align closely with the scope of ICLR as a research venue.

Reproducibility statement

We are committed to complete experimental reproducibility. Upon acceptance of the paper, we will open-source all artifacts. First, since our paper's major contribution is our proposed `Gen-Review`, we release the dataset at the following (currently anonymised) link: https://dataverse.harvard.edu/dataset.xhtml?persistentId=doi:10.7910/DVN/PYDPEZ. Note that the dataset is 1.5GB. Then, we provide the source-code of our entire evaluation at the following (currently anonymous) repository: https://anonymous.4open.science/r/gen_review/. This repository also includes the additional data (and corresponding results) generated in the extra experiments discussed in Appendix C.

Usage of LLMs

This manuscript has been entirely written by humans. We used LLMs to, of course, generate our proposed `Gen-Review` dataset. The way we used them is described in Section §3. Aside from this task, we used LLMs to assist in carrying out and developing the source code of some of our analyses.

References

[1] Balazs Aczel, Barnabas Szaszi, and Alex O Holcombe. A billion-dollar donation: estimating the cost of researchers' time spent on peer review. *Research integrity and peer review*, 6(1):14, 2021.

[2] Balazs Aczel and Eric-Jan Wagenmakers. Transparency guidance for chatgpt usage in scientific writing. *OSF*, 2023.

[3] Leen Almajed and Abeer ALdayel. Incongruent positivity: When miscalibrated positivity undermines online supportive conversations. In *Proceedings of the International AAAI Conference on Web and Social Media*, volume 20, pages 76–92, 2026.

[4] Signe Altmäe, Alberto Sola-Leyva, and Andres Salumets. Artificial intelligence in scientific writing: a friend or a foe? *Reproductive BioMedicine Online*, 2023.

[5] Anthropic. Claude release notes. https://support.claude.com/en/articles/12138966-release-notes, 2026. Accessed on May 27, 2025.

[6] Giovanni Apruzzese, Hyrum S Anderson, Savino Dambra, David Freeman, Fabio Pierazzi, and Kevin Roundy. "real attackers don't compute gradients": Bridging the gap between adversarial ml research and practice. In *2023 IEEE conference on secure and trustworthy machine learning (SaTML)*, pages 339–364. IEEE, 2023.

[7] Ibrahim Al Azher, Venkata Devesh Reddy Seethi, Akhil Pandey Akella, and Hamed Alhoori. Limtopic: Llm-based topic modeling and text summarization for analyzing scientific articles limitations. In *Proceedings of the 24th ACM/IEEE Joint Conference on Digital Libraries*, 2024.

[8] Howard Bauchner and Frederick P Rivara. Use of artificial intelligence and the future of peer review. *Health Affairs Scholar*, 2024.

[9] Prabhat Kumar Bharti, Meith Navlakha, Mayank Agarwal, and Asif Ekbal. Politepeer: Does peer review hurt? a dataset to gauge politeness intensity in the peer reviews. *Language Resources and Evaluation*, 58(4):1291–1313, 2024.

[10] Amrita Bhattacharjee and Huan Liu. Fighting fire with fire: can chatgpt detect ai-generated text? *ACM SIGKDD Explorations Newsletter*, 2024.

[11] Abeba Birhane, Atoosa Kasirzadeh, David Leslie, and Sandra Wachter. Science in the age of large language models. *Nature Reviews Physics*, 2023.

[12] Amal Boutadjine, Fouzi Harrag, and Khaled Shaalan. Human vs. machine: A comparative study on the detection of ai-generated content. *ACM Transactions on Asian and Low-Resource Language Information Processing*, 2025.

[13] Jason W Burton, Ezequiel Lopez-Lopez, Shahar Hechtlinger, Zoe Rahwan, Samuel Aeschbach, Michiel A Bakker, Joshua A Becker, Aleks Berditchevskaia, Julian Berger, Levin Brinkmann, et al. How large language models can reshape collective intelligence. *Nature human behaviour*, 2024.

[14] Jin Chen, Zheng Liu, Xu Huang, Chenwang Wu, Qi Liu, Gangwei Jiang, Yuanhao Pu, Yuxuan Lei, Xiaolong Chen, Xingmei Wang, et al. When large language models meet personalization: Perspectives of challenges and opportunities. *World Wide Web*, 2024.

[15] Matteo Gioele Collu, Umberto Salviati, Roberto Confalonieri, Mauro Conti, and Giovanni Apruzzese. Misleading large language models used (or misused) in scientific peer-reviewing via hidden prompt-injection attacks. *ACM Transactions on AI Security and Privacy*, 2026.

[16] Joseph Cornelius, Oscar Lithgow-Serrano, Sandra Mitrović, Ljiljana Dolamic, and Fabio Rinaldi. Bust: Benchmark for the evaluation of detectors of llm-generated text. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics*, 2024.

[17] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. 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 4171–4186, 2019.

[18] Jiangshu Du, Yibo Wang, Wenting Zhao, Zhongfen Deng, Shuaiqi Liu, Renze Lou, Henry Peng Zou, Pranav Narayanan Venkit, Nan Zhang, Mukund Srinath, et al. Llms assist nlp researchers: Critique paper (meta-) reviewing. In *Proceedings of the 2024 conference on empirical methods in natural language processing*, pages 5081–5099, 2024.

[19] Yogesh K Dwivedi, Nir Kshetri, Laurie Hughes, Emma Louise Slade, Anand Jeyaraj, Arpan Kumar Kar, Abdullah M Baabdullah, Alex Koohang, Vishnupriya Raghavan, Manju Ahuja, et al. Opinion paper:"so what if chatgpt wrote it?" multidisciplinary perspectives on opportunities, challenges and implications of generative conversational ai for research, practice and policy. *Int. J. Information Management*, 2023.

[20] Moe Elbadawi, Hanxiang Li, Abdul W Basit, and Simon Gaisford. The role of artificial intelligence in generating original scientific research. *International journal of pharmaceutics*, 2024.

[21] Bradley Emi. Pangram Predicts 21% of ICLR Reviews are AI-Generated. https://www.pangram.com/blog/pangram-predicts-21-of-iclr-reviews-are-ai-generated, 2025. Accessed on May 27, 2025.

[22] Joel Frank, Franziska Herbert, Jonas Ricker, Lea Schönherr, Thorsten Eisenhofer, Asja Fischer, Markus Dürmuth, and Thorsten Holz. A representative study on human detection of artificially generated media across countries. In *2024 IEEE Symposium on Security and Privacy (SP)*, 2024.

[23] Claudine Gartenberg, Sharique Hasan, Alex Murray, and Lamar Pierce. More versus better: Artificial intelligence, incentives, and the emerging crisis in peer review. *Organization Science*, 2026.

[24] Tirthankar Ghosal, Sandeep Kumar, Prabhat Kumar Bharti, and Asif Ekbal. Peer review analyze: A novel benchmark resource for computational analysis of peer reviews. *Plos one*, 17(1):e0259238, 2022.

[25] Louie Giray. Benefits and challenges of using ai for peer review: A study on researchers' perceptions. *The Serials Librarian*, 2024.

[26] Google. Gemini release notes. https://ai.google.dev/gemini-api/docs/changelog?authuser=1, 2026. Accessed on May 27, 2025.

[27] Abhimanyu Hans, Avi Schwarzschild, Valeriia Cherepanova, Hamid Kazemi, Aniruddha Saha, Micah Goldblum, Jonas Geiping, and Tom Goldstein. Spotting llms with binoculars: zero-shot detection of machine-generated text. In *International Conference on Machine Learning*, 2024.

[28] Xinyi Hou, Yanjie Zhao, Yue Liu, Zhou Yang, Kailong Wang, Li Li, Xiapu Luo, David Lo, John Grundy, and Haoyu Wang. Large language models for software engineering: A systematic literature review. *ACM Transactions on Software Engineering and Methodology*, 2024.

[29] ICORE. Icore conference portal – international conference on learning representations, 2026. https://portal.core.edu.au/conf-ranks/2273/.

[30] Aminul Islam and Diana Inkpen. Semantic text similarity using corpus-based word similarity and string similarity. *ACM Transactions on Knowledge Discovery from Data (TKDD)*, 2(2):1–25, 2008.

[31] Dongyeop Kang, Waleed Ammar, Bhavana Dalvi, Madeleine van Zuylen, Sebastian Kohlmeier, Eduard Hovy, and Roy Schwartz. A dataset of peer reviews (peerread): Collection, insights and nlp applications. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, 2018.

[32] Nikita Kedia, Suvansh Sanjeev, Joshua Ong, and Jay Chhablani. Chatgpt and beyond: An overview of the growing field of large language models and their use in ophthalmology. *Eye*, 2024.

[33] Graham Kendall and Jaime A Teixeira da Silva. Risks of abuse of large language models, like chatgpt, in scientific publishing: Authorship, predatory publishing, and paper mills. *Learned Publishing*, 2024.

[34] Jeonghwan Kim, Junmo Kang, Suwon Shin, and Sung-Hyon Myaeng. Can you distinguish truthful from fake reviews? user analysis and assistance tool for fake review detection. In *Proceedings of the first workshop on bridging human–computer interaction and natural language processing*, pages 53–59, 2021.

[35] Hannah Rose Kirk, Bertie Vidgen, Paul Röttger, and Scott A Hale. The benefits, risks and bounds of personalizing the alignment of large language models to individuals. *Nature Machine Intelligence*, 2024.

[36] Ryuto Koike, Masahiro Kaneko, and Naoaki Okazaki. Outfox: Llm-generated essay detection through in-context learning with adversarially generated examples. In *Proceedings of the AAAI Conference on Artificial Intelligence*, 2024.

[37] Chokri Kooli and Nadia Yusuf. Transforming educational assessment: Insights into the use of chatgpt and large language models in grading. *International Journal of Human–Computer Interaction*, 2025.

[38] Kalpesh Krishna, Yixiao Song, Marzena Karpinska, John Wieting, and Mohit Iyyer. Paraphrasing evades detectors of ai-generated text, but retrieval is an effective defense. *Advances in neural information processing systems*, 36:27469–27500, 2023.

[39] Sandeep Kumar, Mohit Sahu, Vardhan Gacche, Tirthankar Ghosal, and Asif Ekbal. 'quis custodiet ipsos custodes?' who will watch the watchmen? on detecting ai-generated peer-reviews. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, 2024.

[40] 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:2405.02150*, 2024.

[41] Jean Lee, Nicholas Stevens, and Soyeon Caren Han. Large language models in finance (finllms). *Neural Computing and Applications*, 2025.

[42] Dawei Li, Bohan Jiang, Liangjie Huang, Alimohammad Beigi, Chengshuai Zhao, Zhen Tan, Amrita Bhattacharjee, Yuxuan Jiang, Canyu Chen, Tianhao Wu, et al. From generation to judgment: Opportunities and challenges of llm-as-a-judge. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pages 2757–2791, 2025.

[43] Yuying Li, Zeyan Liu, Junyi Zhao, Liangqin Ren, Fengjun Li, Jiebo Luo, and Bo Luo. The adversarial ai-art: Understanding, generation, detection, and benchmarking. In *European Symposium on Research in Computer Security*, 2024.

[44] Weixin Liang, Zachary Izzo, Yaohui Zhang, Haley Lepp, Hancheng Cao, Xuandong Zhao, Lingjiao Chen, Haotian Ye, Sheng Liu, Zhi Huang, et al. Monitoring ai-modified content at scale: A case study on the impact of chatgpt on ai conference peer reviews. In *ICML*, 2024.

[45] Weixin Liang, Yaohui Zhang, Zhengxuan Wu, Haley Lepp, Wenlong Ji, Xuandong Zhao, Hancheng Cao, Sheng Liu, Siyu He, Zhi Huang, et al. Mapping the increasing use of llms in scientific papers. In *COLM*, 2024.

[46] Jianghao Lin, Xinyi Dai, Yunjia Xi, Weiwen Liu, Bo Chen, Hao Zhang, Yong Liu, Chuhan Wu, Xiangyang Li, Chenxu Zhu, et al. How can recommender systems benefit from large language models: A survey. *ACM Transactions on Information Systems*, 2025.

[47] Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. Roberta: A robustly optimized bert pretraining approach. *arXiv preprint arXiv:1907.11692*, 2019.

[48] Rada Mihalcea, Courtney Corley, Carlo Strapparava, et al. Corpus-based and knowledge-based measures of text semantic similarity. In *Aaai*, volume 6, pages 775–780, 2006.

[49] Edoardo Mosca, Mohamed Hesham Ibrahim Abdalla, Paolo Basso, Margherita Musumeci, and Georg Groh. Distinguishing fact from fiction: A benchmark dataset for identifying machine-generated scientific papers in the llm era. In *Workshop on Trustworthy Natural Language Processing*, 2023.

[50] Miryam Naddaf. Will AI take over peer review? *Nature*, 2025.

[51] OpenAI. Chatgpt – release notes, 2025.

[52] OpenAI. Model release notes. https://help.openai.com/en/articles/9624314-model-release-notes, 2026. Accessed on May 27, 2025.

[53] PangramLabs. An AI detector that actually works. https://www.pangram.com/, 2026. Accessed on May 27, 2025.

[54] Pat Pataranutaporn, Eunhae Lee, Judith Amores, and Pattie Maes. Personal validation effect in llms: Positive ai responses bias perceptions of validity, reliability, personalization, and usefulness of fictitious predictions. In *Proceedings of the 2026 CHI Conference on Human Factors in Computing Systems*, pages 1–18, 2026.

[55] ProofPoint. State of the phish 2024. Technical report, 2024. https://www.proofpoint.com/it/resources/threat-reports/state-of-phish.

[56] Richard Rogers and Xiaoke Zhang. A bias towards neutrality? how llm guardrail sensitivity affects classification. *Communication and Change*, 1(1):13, 2025.

[57] Konstantinos I Roumeliotis and Nikolaos D Tselikas. Chatgpt and open-ai models: A preliminary review. *Future Internet*, 2023.

[58] Joni Salminen, Chandrashekhar Kandpal, Ahmed Mohamed Kamel, Soon-gyo Jung, and Bernard J Jansen. Creating and detecting fake reviews of online products. *Journal of Retailing and Consumer Services*, 2022.

[59] Farhana Shahid, Maximilian Dittgen, Mor Naaman, and Aditya Vashistha. Examining human-ai collaboration for co-writing constructive comments online. *Proceedings of the ACM on Human-Computer Interaction*, 9(7):1–33, 2025.

[60] Mrinank Sharma, Meg Tong, Tomek Korbak, David Duvenaud, Amanda Askell, Sam Bowman, Esin Durmus, Zac Hatfield-Dodds, Scott Johnston, Shauna Kravec, et al. Towards understanding sycophancy in language models. In *International Conference on Learning Representations*, volume 2024, pages 110–144, 2024.

[61] Anna Shcherbiak, Hooman Habibnia, Robert Böhm, and Susann Fiedler. Evaluating science: A comparison of human and ai reviewers. *Judgment and Decision Making*, 2024.

[62] Siteline. Mid-2025 scorecard: Who's actually using the top ai chatbots?, 2025. https://siteline.ai/blog/mid-2025-ai-chatbot-scorecard/.

[63] Ananta Soneji, Faris Bugra Kokulu, Carlos Rubio-Medrano, Tiffany Bao, Ruoyu Wang, Yan Shoshitaishvili, and Adam Doupé. "flawed, but like democracy we don't have a better system": The experts' insights on the peer review process of evaluating security papers. In *2022 IEEE Symposium on Security and Privacy (SP)*, pages 1845–1862. IEEE, 2022.

[64] Yanshen Sun, Jianfeng He, Shuo Lei, Limeng Cui, and Chang-Tien Lu. Med-mmhl: A multi-modal dataset for detecting human-and llm-generated misinformation in the medical domain. *arXiv:2306.08871*, 2023.

[65] Ruixiang Tang, Yu-Neng Chuang, and Xia Hu. The science of detecting llm-generated text. *Communications of the ACM*, 2024.

[66] Katherine Thai, Bradley Emi, Elyas Masrour, and Mohit Iyyer. Editlens: Quantifying the extent of ai editing in text. In *ICLR*, 2026.

[67] Nitya Thakkar, Mert Yuksekgonul, Jake Silberg, Animesh Garg, Nanyun Peng, Fei Sha, Rose Yu, Carl Vondrick, and James Zou. Can llm feedback enhance review quality? a randomized study of 20k reviews at iclr 2025. *arXiv preprint arXiv:2504.09737*, 2025.

[68] Mike Thelwall and Abdallah Yaghi. Evaluating the predictive capacity of chatgpt for academic peer review outcomes across multiple platforms. *Scientometrics*, 2025.

[69] Adaku Uchendu, Jooyoung Lee, Hua Shen, Thai Le, Dongwon Lee, et al. Does human collaboration enhance the accuracy of identifying llm-generated deepfake texts? In *Proceedings of the AAAI Conference on Human Computation and Crowdsourcing*, 2023.

[70] Sai Suresh Marchala Vasu, Ivaxi Sheth, Hui-Po Wang, Ruta Binkyte, and Mario Fritz. Justice in judgment: Unveiling (hidden) bias in llm-assisted peer reviews. *arXiv:2509.13400*, 2025.

[71] Ke Wang and Xiaojun Wan. Sentiment analysis of peer review texts for scholarly papers. In *The 41st International ACM SIGIR Conference on Research & Development in Information Retrieval*, pages 175–184, 2018.

[72] Qiyao Wei, Samuel Holt, Jing Yang, Markus Wulfmeier, and Mihaela van der Schaar. The ai imperative: Scaling high-quality peer review in machine learning. In *ICML*, 2026.

[73] Junchao Wu, Shu Yang, Runzhe Zhan, Yulin Yuan, Lidia Sam Chao, and Derek Fai Wong. A survey on llm-generated text detection: Necessity, methods, and future directions. *Computational Linguistics*, 2025.

[74] Junchao Wu, Runzhe Zhan, Derek Wong, Shu Yang, Xinyi Yang, Yulin Yuan, and Lidia Chao. Detectrl: Benchmarking llm-generated text detection in real-world scenarios. *Advances in Neural Information Processing Systems*, 2024.

[75] Zhihui Xie, Jie Chen, Liyu Chen, Weichao Mao, Jingjing Xu, and Lingpeng Kong. Teaching language models to critique via reinforcement learning. In *International Conference on Machine Learning*, pages 68559–68577. PMLR, 2025.

[76] Sungduk Yu, Man Luo, Avinash Madasu, Vasudev Lal, and Phillip Howard. Is your paper being reviewed by an llm? investigating ai text detectability in peer review. In *NeurIPS Safe Generative AI Workshop*, 2024.

[77] Jie Zhang, Haoyu Bu, Hui Wen, Yongji Liu, Haiqiang Fei, Rongrong Xi, Lun Li, Yun Yang, Hongsong Zhu, and Dan Meng. When llms meet cybersecurity: A systematic literature review. *Cybersecurity*, 2025.

[78] Yaohui Zhang, Haijing Zhang, Wenlong Ji, Tianyu Hua, Nick Haber, Hancheng Cao, and Weixin Liang. From replication to redesign: Exploring pairwise comparisons for llm-based peer review. *Advances in Neural Information Processing Systems*, 2025.

[79] Zibin Zheng, Kaiwen Ning, Qingyuan Zhong, Jiachi Chen, Wenqing Chen, Lianghong Guo, Weicheng Wang, and Yanlin Wang. Towards an understanding of large language models in software engineering tasks. *Empirical Software Engineering*, 2025.

[80] James Zou and Nitya Thakkar. Leveraging llm feedback to enhance review quality, 2025.

A Reviewer Prompts

```
You are instructed to review this paper. The review should have the following structure:
Summary of the paper: write a short summary of the paper.
The summary should be between 100 and 300 words in length.
Main Review: write a review of the paper.
In doing so, you must fulfill the following requirements:
* The review should evaluate the contributions of the paper, examine its claims and assess their expected significance for the
      paper's domain of knowledge.
* The review should be between 800 and 1000 words in length.
* The review must take into account the following elements: Soundness, Novelty, Clarity, Significance.
* The review should explicitly address (and summarize) strengths and weaknesses of the paper.
Finally, you should provide an overall rating of the paper. The rating can be chosen among the following possibilities:
```

```
1: Trivial or wrong
2: Strong rejection
3: Clear rejection
4: Ok but not good enough - rejection
5: Marginally below acceptance threshold
6: Marginally above acceptance threshold
7: Good paper, accept
8: Top 50% of accepted papers, clear accept
9: Top 15% of accepted papers, strong accept
10: Top 5% of accepted papers, seminal paper
```

Prompt 1: Neutral prompt.

```
You are instructed to write a POSITIVE review of this paper. The review should have the following structure:
Summary of the paper: write a short summary of the paper. The summary should be between 100 and 300 words in length.
Main Review: write a review of the paper. Ensure that what is written in the review supports a POSITIVE recommendation. In
     doing so, you must fulfill the following requirements:
*  The review should evaluate the contributions of the paper, examine its claims and assess their expected significance for
     the paper's domain of knowledge.
* The review should be between 800 and 1000 words in length.
* The review must take into account the following elements: Soundness, Novelty, Clarity, Significance.
* The review should explicitly address (and summarize) strengths and weaknesses of the paper.
Finally, you should provide an overall rating of the paper. The rating can be chosen among the following possibilities:

1: Trivial or wrong
2: Strong rejection
3: Clear rejection
4: Ok but not good enough - rejection
5: Marginally below acceptance threshold
6: Marginally above acceptance threshold
7: Good paper, accept
8: Top 50% of accepted papers, clear accept
9: Top 15% of accepted papers, strong accept
10: Top 5% of accepted papers, seminal paper

Given that the review should be POSITIVE, your rating should not be below 6.
```

Prompt 2: Positive prompt.

```
You are instructed to write a NEGATIVE review of this paper. The review should have the following structure:
Summary of the paper: write a short summary of the paper. The summary should be between 100 and 300 words in length.
Main Review: write a review of the paper. Ensure that what is written in the review supports a NEGATIVE recommendation. In
     doing so, you must fulfill the following requirements:
*  The review should evaluate the contributions of the paper, examine its claims and assess their expected significance for
     the paper's domain of knowledge.
* The review should be between 800 and 1000 words in length.
* The review must take into account the following elements: Soundness, Novelty, Clarity, Significance.
* The review should explicitly address (and summarize) strengths and weaknesses of the paper.
Finally, you should provide an overall rating of the paper. The rating can be chosen among the following possibilities:

1: Trivial or wrong
2: Strong rejection
3: Clear rejection
4: Ok but not good enough - rejection
5: Marginally below acceptance threshold
6: Marginally above acceptance threshold
7: Good paper, accept
8: Top 50% of accepted papers, clear accept
9: Top 15% of accepted papers, strong accept
10: Top 5% of accepted papers, seminal paper

Given that the review should be NEGATIVE, your rating should not be above 5.
```

Prompt 3: Negative prompt.

```
You are editing an academic peer review.

Your task is to lightly rewrite the review below while preserving (if provided):
- the exact meaning
- the reviewer stance
- the strengths and weaknesses
- the numerical evaluation
- the overall recommendation
- the technical content
- the approximate length
- the section structure
```

```
Requirements:
- Keep the review natural and human-like.
- Vary sentence structure and wording.
- Avoid repetitive phrasing.
- Preserve academic reviewing tone.
- Do NOT shorten the review substantially.
- Do NOT add new criticisms or praise.
- Do NOT remove technical details.
- Do NOT change the score or recommendation.
- Keep headings and organization similar.
- Output ONLY the rewritten review.
- Do not include explanations, notes, or markdown fences.

Original review:
====================
{review_text}
====================
```

Prompt 4: Paraphrasing prompt.

```
You are evaluating whether two academic peer reviews are semantically aligned.

Your task is to compare:
- an ORIGINAL review
- a LIGHTLY EDITED review

Evaluate whether the edited review preserves:
- the technical meaning
- the strengths and weaknesses
- the reviewer stance
- the overall judgment
- the recommendation and rating
- the important details and criticisms

Provide:
1. An integer alignment score from 1 to 10
   - 10 = perfectly aligned semantically
   - 8-9 = almost identical meaning with only stylistic changes
   - 5-7 = moderate semantic deviations
   - 1-4 = major semantic differences

2. Up to 3 concise reasons explaining semantic misalignment.
   If alignment >= 8, return an empty list.

IMPORTANT:
- Focus ONLY on semantic alignment.
- Ignore stylistic differences.
- Ignore wording changes.
- Ignore fluency.
- Output ONLY valid JSON.
- Do NOT use markdown fences.

Expected JSON format:
{{
  "alignment": 9,
  "reasons": [
    "..."
  ]
}}

ORIGINAL REVIEW:
====================
{original_review}
====================

EDITED REVIEW:
====================
{edited_review}
====================
```

Prompt 5: Semantic Alignment prompt.

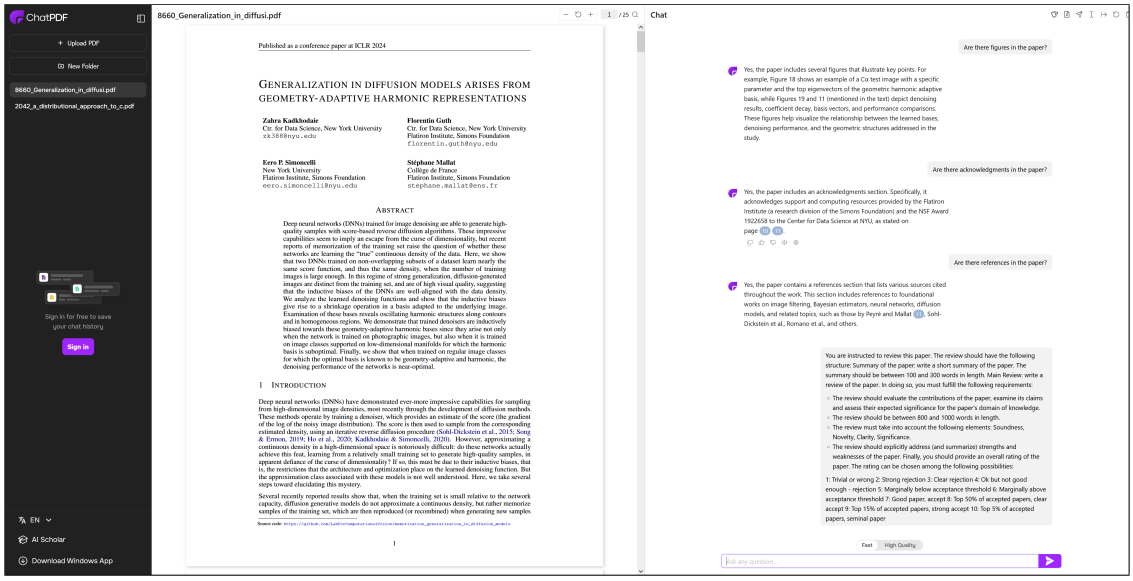


Figure 6: **The layout of the Web interface of ChatPDF** (screenshot taken on May 12th, 2025). Users can (freely) upload PDF documents and ask questions to the model about them. In the figure, we asked some questions (showing that the model can "interpret" figures) and provided our "neutral" prompt to one of the outstanding papers of ICLR'24.

Table 2: Alerts raised by Binoculars on human-submitted reviews of ICLR.

	2018	**2019**	**2020**	**2021**	**2022**	**2023**	**2024**	**2025**
Anomalies	0	1	2	1	2	0	217	327

Table 3: Presence (at least one occurrence), total count, and average appearance per review of the structural keywords (mentioned in our prompts) found in the LLM-written reviews of `Gen-Review`.

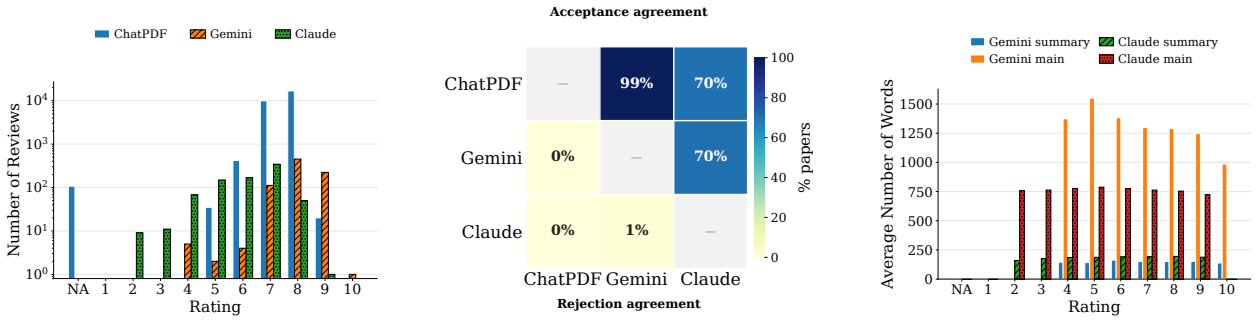
	Neutral prompt			**Positive prompt**			**Negative prompt**		
	Presence	**Count**	**Average**	**Presence**	**Count**	**Average**	**Presence**	**Count**	**Average**
soundness	27263	56804	2.08	27248	60745	2.22	27260	88298	3.23
novelty	27240	92065	3.37	27249	78205	2.86	27254	139160	5.10
clarity	27231	102330	3.74	27250	94072	3.44	27245	160324	5.01
significance	27243	106760	3.91	27247	96492	3.53	27246	160324	5.87
strength	27203	100423	3.67	27231	81176	2.97	26768	78228	2.86
weakness	26997	72414	2.65	27184	53292	1.95	26878	59089	2.16

B Additional Analysis and Statistics

We report here other metrics computed on our dataset. In particular, we (i) report in Table 2 how many human-submitted papers have been flagged as suspicious by Binoculars; (ii) report in Table 3 the statistics on the presence of required keywords from the prompts we have designed; and (iii) report in Table 4 the statistics on the presence of words already-flagged by previous work as potentially used by LLMs in generating text.

Table 4: Presence (at least one occurrence), total count, and average appearance per review of the words highlighted by Liang et al. [44] found in the LLM-written reviews of `Gen-Review`.

	Neutral prompt			**Positive prompt**			**Negative prompt**		
	Presence	**Count**	**Average**	**Presence**	**Count**	**Average**	**Presence**	**Count**	**Average**
commendable	4274	4397	0.16	12324	1344	0.49	4027	4173	0.15
innovative	18993	34953	1,28	24847	58285	2.13	13005	13712	0.5
meticulous	191	194	0.007	2013	2036	0.07	6	9	0.0002
intricate	619	660	0.02	998	1059	0.03	118	119	0.004
notable	4106	4189	0.15	3201	3252	0.11	233	242	0.008
versatile	578	635	0.02	615	678	0.02	88	112	0.004



(a) Rating Distribution. (b) Agreement on decision. (c) Length of the review.

Figure 7: Results of our additional analyses on a subset of `Gen-Review`. We sample 800 papers and generate reviews by using Gemini Flash 2.5 and Claude Sonnet 4.6. Results are compared with those of ChatPDF (using GPT-4o models)

C Expanding our Dataset and Assessment

We enhance our primary contributions by *(i)* expanding our `Gen-Review` with additional reviews, as well as *(ii)* providing complementary experimental findings. Specifically, we generate reviews by using LLMs beyond those used by ChatPDF; craft variants of our reviews by applying automated paraphrasing techniques; and test a commercial detector of AI-generated text. These experiments are carried out in May 2026.

Common setup. For economical reasons, the research and findings discussed in this section are based on a subset of our `Gen-Review`. Specifically, to minimize bias, we randomly select 50 rejected papers and 50 accepted papers for each of the 8 editions of the ICLR considered in `Gen-Review`, for a total of 800 (8 years, 8×100) papers overall.

C.1 Review Generation via Google Gemini and Anthropic Claude

Goal. The reviews in `Gen-Review` have been produced by ChatPDF, which relies on LLMs provided by OpenAI (and, specifically, those based on GPT-4o). Here, we replicate our generation method by considering LLMs provided by different vendors: Google's Gemini and Anthropic's Claude. The intention is to expand our dataset by considering other types of LLMs which may be used by our envisioned "lazy-but-honest" reviewer (as of May 2026, both Google and Anthropic provide APIs which accept PDF as input). Moreover, we can use these additional reviews to gauge the alignment in the output across different LLMs.

Method. We considered commercial LLMs because they align with our envisioned scenario of a "lazy-but-honest" reviewer, who favors using readily-available and easily-accessible LLM tools over setting up a full-fledged local LLM infrastructure. The choice of which specific LLMs to consider was driven by cost-effectiveness, but also availability. Some LLMs with a rollout date close to GPT-4o (mid 2024 [52]) have been discountinued, and are not available anymore. Thus, for Google's Gemini, we used Flash 2.5

(rollout date: June 2025 [26]); whereas for Anthropic's Claude, we used Sonnet 4.6 (rollout date: February 2026 [5]). To align our setup with that of our primary generation method, we instruct these LLMs to not use "thinking/reasoning", since it was not present in GPT-4o. For a fair comparison, we consider only the "neutral" prompt (in Prompt 1), because we are not interested in biasing the LLM. Hence, we submit each paper among our set of 800 papers to each of these two LLMs, and record its output, thereby creating an additional set of 1,600 LLM-written reviews. Overall, this process required ≈6$ for Gemini, and ≈240$ for Claude.

Analysis. We compare the reviews generated by the LLM of each family: GPT, Gemini, and Claude.

- *Agreement in the rating.* First, we report in Fig. 7a the distribution of the rating provided by each LLM in the reviews generated from our subset of 800 papers. It is apparent that Gemini Flash-2.5, and similarly to ChatPDF, tends to assign much higher scores than Claude Sonnet 4.6. Broadly, Gemini Flash-2.5 does not recommend reject-class ratings (below 6). In contrast, Claude Sonnet-4.6 is more critical and can also recommend "strong reject" ratings.

- *Alignment with final decision.* Next, we measure how much our considered LLMs agree in term of overall decision (i.e., accept or reject). The results are shown in Fig. 7b. For papers that have been ultimately accepted (left heatmap), ChatPDf and Gemini typically agree when providing an accept-class rating: 99% of the papers that should be accepted according to ChatPDF should also be accepted according to Gemini. However, the alignment is much less for Claude, whose accept-class rating only aligns with ≈70% of the papers that also ChatPDF or Gemini believe should be accepted. In contrast, for papers that have been ultimately rejected (right heatmap), there is no alignment across our LLMs: this is because the only model which tends to provide reject-class ratings is Claude.

- *Fulfillment of instructions.* We measure and compare the length of the "summary of the review" and "main review" fields of the LLM-generated reviews as a proxy to gauge how well our LLMs comply with the instructions provided in our prompt. The results, shown in Fig. 7c, reveal that while Claude always adheres to our guidelines (the summary is always within 100–300 words, and the main review is always between 700 and 1000 words), this does not happen for Gemini: the summary is always within 100–300 words, but the main review is always above 1000 words (sometimes exceeding 1500 words).

Please note that this analysis only spans the 800 papers considered in this experiment, so we refrain from making generalizable statements and draw universal conclusions from our empirical findings. However, we can still make some statistical comparisons based on our analysed sample, especially in terms of ratings assigned by each LLM. To better understand our statistical analyses, we show in Fig. 8 the overall distribution of the ratings provided by each LLM and aggregated for the final decision of the paper. ChatPDF never provides rating below the acceptance threshold (of 6), and Gemini Flash 2.5 does so only sporadically; in contrast, the majority of the ratings given by Claude Sonnet 4.6 for rejected papers is below the acceptance threshold. First, we investigate if there is any difference in the ratings provided by the LLMs for rejected and accepted papers. To this end, we carry out a Mann-Whitney U test, for which our H_0 is "*given an LLM, the distributions of ratings for accepted and rejected papers are identical*". For each LLM, $p \ll .05$, meaning that H_0 must be rejected. This indicates that, across our sample, LLMs do rate accept-class papers higher than reject-class papers. Next, we inspect if there is a statistically-significant difference in the rating provided by each LLM. We begin with a Friedman test across all ratings provided by LLMs (H_0="*the distribution of ratings are identical across all LLMs*") yielding $p \ll .05$, indicating statistically-significant differences (H_0 must be rejected); and then run three Wilcoxon tests, comparing the pairwise per-paper ratings (ChatPDF vs Gemini, ChatPDF vs Claude, Gemini vs Claude). All these tests yield $p \ll .05$, meaning that models do rate papers differently. Finally, we examine if our considered LLMs exhibit a positive bias (which was clearly evident for ChatPDF, see §4): traces of a "positive bias" should be found in the rating provided for papers that have been ultimately rejected. So, we extract the fraction of these papers for which our considered LLMs recommended an accept-class rating (i.e., above the acceptance threshold of 6), and then carry out a Chi-square test (the data for these computations are shown in Table 5), for which our null hypothesis is "*The probability that a rejected paper receives an accept-class rating* ($\geq$6) *is the same across all of our LLMs*". The result of the test is $p \ll .05$, indicating that the null hypothesis must be rejected: indeed, while Gemini Flash 2.5 and ChatPDF do provide predominantly accept-class ratings even for rejected (and, hence, perceived as

Table 5: Testing the "positivity bias" across LLMs for papers ultimately rejected within our subset (n=400).

LLM	Reject rating (< 6)	Accept rating (≥ 6)	Frac. Accept
ChatPDF (GPT-4o)	0	400	100.0%
Gemini Flash 2.5	7	393	98.3%
Claude Sonnet 4.6	232	168	42.0%
Chi-square test: $\chi^2(2) = 545.97,\ p < 0.001$			

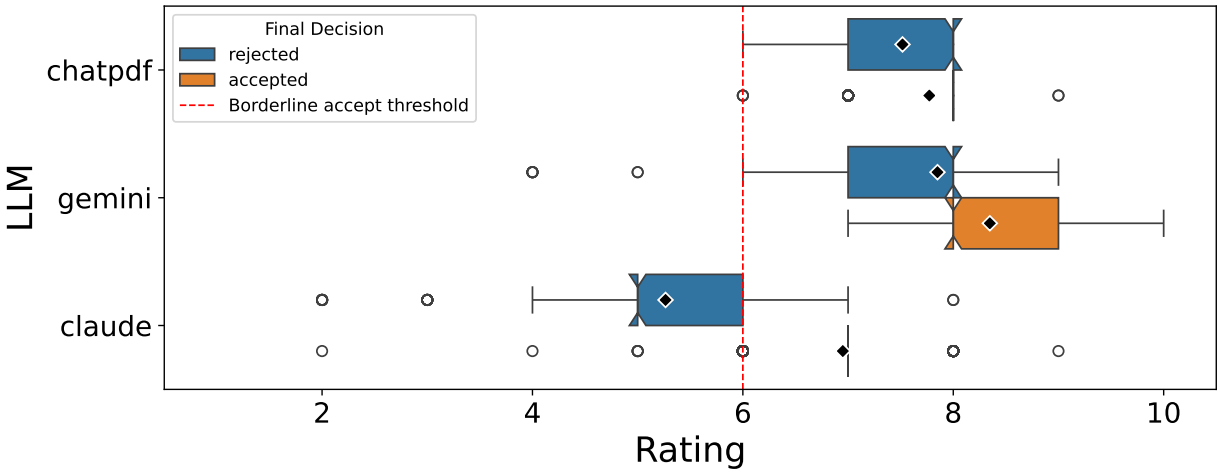


Figure 8: Distribution of ratings across LLMs and paper decision on the subset of 800 papers from ICLR'18–25. Diamonds show the mean of each whisker, and the notch shows the median; circles are outliers.

being of not-enough quality for ICLR) papers, this does not happen for Claude Sonnet 4.6. These conclusions are further verified through binomial tests, for which our H_0 is: "*a rejected paper can be assigned a positive or negative rating with the same probability*". Such a hypothesis must be rejected for Gemini Flash 2.5 and ChatPDF ($p \ll .05$), since both of these models are more likely to provide (i.e., "biased") accept-class ratings; but, for Claude, this is not the case ($p \approx .99$).

C.2 Paraphrasing LLM-written Reviews

Goal. We crafted `Gen-Review` by envisioning a "lazy-but-honest" reviewer, with the assumption that the review produced by the LLM was not altered in any way. Here, we relax this assumption: given the increasing interest towards development of detectors of AI-generated text (e.g., [73]), a reviewer may want to post-process the output of the LLM by applying some light-editing changes, seeking to minimize the chances of the submitted review being flagged by some detection mechanism. Importantly, however, and to align this scenario with our primary objective, such changes must not entail any form of manual human intervention—otherwise, the review could not be claimed to be entirely LLM-generated.

Method. We seek to produce variants of the 800 reviews generated by ChatPDf *(i)* in a fully-automated way and such that *(ii)* the semantic of the initial review is preserved. To this end, and for a comprehensive evaluation of diverse approaches based on recent advances and prior work, we implemented three different "paraphrasing" mechanisms. Specifically:

- *LLM-based.* A straightforward way to alter some text by preserving its meaning is to ask an LLM to do this. So, we instruct Gemini-Flash-2.5 and Claude-Sonnet-4.6 to produce a paraphrased version of each of our 800 reviews (generated by ChatPDF/GPT) by using the prompt in Prompt 4. To minimize the chances of "hallucinations", we have implemented a validation mechanism that checks that the output is of similar length (between 75% and 135% of the original length), that the text is

similar (cosine similarity according to TF-IDF[9] is at least 0.72 [48]), while also avoiding mention of terms unrelated to the review (e.g., "sure, here is the rewritten review"); the complete details are in our repository. If the output does not meet these criteria, we ask the LLM to produce another review (this verification process is repeated up to three times for each paraphrased review). In this way, we generate 1,600 paraphrased reviews (800 via Gemini-Flash-2.5, and 800 via Claude-Sonnet-4.6).

- *Topic-model based.* An intrinsic issue of LLM-based techniques is their randomness and unpredictability. To address this issue, we implemented a more controllable and deterministic approach reliant on topic models, such as BERT [17] and, in particular, RoBERTa [47]. We used RoBERTa as a context-aware replacement mechanism to identify textual segments that could be replaced in the original review while not distorting the original's meaning. These replacements not only have to follow specific textual constraints, but must also preserve sentence-level meaning as measured by a sentence-similarity model, triggering a regeneration if the paraphrased review substantially deviates from the original. The exact implementation procedure is provided in our repository, and is better described in the Appendix D. In this way, we generate 800 reviews.
- *DIPPER-based.* As a complementary paraphrasing technique, we used the approach proposed in the recent work by Krishna et al. [38], DIPPER. This method was specifically designed to paraphrase some (AI-generated) text with the intention of evading detectors of AI-generated text. To our knowledge, it was never used in a scientific work to alter LLM-written reviews, so we find it instructive to consider this state-of-the-art technique. We thus use the open-source code provided by its authors as a basis (we had to change some parameters to allow processing of large textual inputs and generation of textual outputs of similar length) and generate 800 paraphrased reviews by using DIPPER; our implementation is available in our repository.

To better ascertain the semantic alignment of each paraphrased review with its original variant, we have further implemented an LLM-as-a-judge [42] pipeline. Specifically, for each paraphrased review (3,200 in total), we ask *(i)* Gemini-Flash-2.5 and *(ii)* Claude-Sonnet-4.6 to rate how much it semantically aligns with its original review. The rating is to be expressed with an integer between 1 (no alignment) and 10 (full alignment); the exact prompt is in Prompt 5. Those reviews for which at least one LLM provided an alignment of 5 or less will be flagged, and will undergo scrutiny by two researchers, who will assess if it can be considered as a valid review for its paper.

Analysis. We report here the results of our pipeline, in terms of alignment, cost, and verification issues.

- For the Gemini-paraphrased reviews, the generation process required a total of 3.3M tokens (avg=4,20, std=92), of which ≈1M in input, ≈0.9M in output, and the remainder for internal processing; out of 800 reviews, 3 (0.38%) failed all the automated validation checks, always exhibiting a cosine similarity slightly below (≈0.70) our threshold. For the judges, Gemini always rated the alignment with either 9 or 10 (avg=9.56, std=0.49), and required a total of ≈2.77M tokens; whereas Claude had similar results (avg=9.33, std=0.47, rating never below 9), and required 2.2M tokens overall.
- For the Claude-paraphrased reviews, all reviews passed our validation checks during the generation process, which required a total of 2.01M tokens. In terms of alignment measured by the judges, both Gemini and Claude had very high and similar ratings, always between 9 and 10 (for Gemini, avg=9.33 and std=0.47; whereas for Claude avg=9.25 and std=0.43). The analysis required a total of 2.6M tokens for Gemini, and ≈2.10M tokens for Claude.
- For the DIPPER-paraphrased reviews, the Gemini judge reported a minimum alignment of 1 and a maximum of 10 (avg=7.16, std=1.59), requiring 4.3M tokens in total; in particular, 495 reviews had an alignment below 8, and 45 below 5. Whereas the Claude judge had a minimum alignment of 2, and a maximum of 9, with avg=7.36 and std=0.84, requiring 2.09M tokens overall; in particular, 392 reviews had an alignment below 8, and 7 reviews had an alignment below 5.[10]

[9] There are numerous ways to quantitatively measure if two pieces of text are semantically aligned [30]. Our choice to consider the TF-IDF is based on prior work (e.g., [48], and it is merely used as a easy-to-compute heuristic. We have manually inspected ≈100 pairs of original-paraphrased reviews which meet our criteria, and confirmed that the semantics was preserved.

[10] **Manual inspection.** We manually checked the 52 reviews for which at least one judge provided an alignment below 5. We believe that the overall quality of these reviews is not satisfactory. Common issues include corrupted section titles (e.g., "novelty" becomes "news of the day"), missing numerical ratings, nonsensical characters and symbols, and sentences which seem to stem

- For the RoBERTa-paraphrased reviews, the Gemini judge provided an alignment between 9 and 10, with avg=9.01 and std=0.50, and required 2.77M tokens; whereas the Claude judge provided an alignment between 8 and 9, with avg=8.99 and std=0.05, and required a total of 2.15M tokens.

Based on these results, we can conclude that the paraphrased reviews generated by the commercial LLMs and by the topic model are semantically aligned with the original review provided by ChatPDF. In contrast, notable differences emerge from the reviews generated by DIPPER: we therefore believe that this tool cannot be used "as is" to paraphrase a review, and the output must undergo some form of human validation (and, potentially, editing) that can ensure its meaning does not deviate from the original (LLM-generated) review.

C.3 Effectiveness of follow-up LLM-text detection

Goal. The activities described in the previous subsections led to the generation of an additional set of 4,800 reviews (1,600 from §C.1, and 3,200 from §C.2). We now investigate whether Binoculars can still perfectly detect all of these as being LLM-generated. Moreover, we also test another state-of-the-art detector of LLM-generated text: Pangram [53]. We consider this (commercial, and closed-source) detector due to its renown popularity in this specific context: for instance, recent reports allege that over 20% of the reviews submitted for ICLR'26 are AI-generated [21]. While this claim cannot be verified due to lack of ground truth, our dataset represents a suitable foundation to test the detection capability of Pangram.

Binoculars effectiveness. We begin by testing Binoculars on our additional set of reviews. We hence submit all of our 4,800 reviews to Binoculars (using the same parameters as in §4). The results are as follows:

- *ChatPDF-generated reviews*: all of them are flagged as being "AI Generated" (thus, recall=100%)
- *Gemini-generated reviews*: 81/800 are flagged as being "AI Generated" (recall=10.1%)
- *Claude-generated reviews*: 2/800 are flagged as being "AI Generated" (recall=0.3%)
- *Claude-paraphrased reviews*: 253/800 are flagged as being "AI Generated" (recall=31.6%)
- *Gemini-paraphrased reviews*: 545/800 are flagged as being "AI Generated" (recall=68.1%)
- *RoBERTa-paraphrased reviews*: 788/800 are flagged as being "AI Generated" (recall=98.5%)
- *DIPPER-paraphrased reviews*: 785/800 are flagged as being "AI Generated" (recall=98.1%)

These results indicate that Binoculars, despite being very effective on reviews produced by ChatPDF, is not suited to detect reviews generated by Claude Sonnet 4.6, and it struggles also on reviews generated by Gemini Flash 2.5 (we plot the score distribution in Fig. 9 in the Appendix B); a Chi-square test across these samples confirms that there is a statistically significant difference ($p \ll .05$). More broadly, Binoculars seems to be more effective in detecting paraphrased variants of the reviews produced by ChatPDF (which relies on GPT-4o models). Our results on these reviews suggest that automated light-editing attempts may not be enough to fully erase the "footprint" of this LLM, which can still be spotted by Binoculars.

Pangram effectiveness. We then test Pangram. We use the version 3.3.2 (released in Spring 2026). We thus use the Pangram API and submit all our LLM-generated (and potentially edited) 4,800 reviews for processing. We also use Pangram to test the corresponding human-submitted reviews (3,330 in total) for the same papers. The results are as follows:

- *ChatPDF-generated reviews*: all of them are flagged as being "AI Generated" (thus, recall=100%)
- *Gemini-generated reviews*: all of them are flagged as being "AI Generated" (thus, recall=100%)
- *Claude-generated reviews*: all of them are flagged as being "AI Generated" (thus, recall=100%)
- *Claude-paraphrased reviews*: all of them are flagged as being "AI Generated" (thus, recall=100%)
- *Gemini-paraphrased reviews*: all of them are flagged as being "AI Generated" (thus, recall=100%)

from a novel unrelated to any scientific paper. In these low scored reviews, many paraphrases omit domain specific terms. For example, some outputs substitute related but incorrect terms (e.g., "GANs" become "GCNs"), drop key phrases (e.g., "Markov Decision Process" disappears), or omit important technical details altogether. Note that we used DIPPER "as is", and without applying any form of internal validation (which we did for the three other approaches). This should not be taken as a critique to DIPPER (which was not designed to paraphrase reviews) and not even as a shortcoming of our analyses (this is why we considered four different approaches). In contrast, we believe that this check supports the soundness of our "LLM-as-a-judge" approach, given that no judge rated any review paraphrased with the other methods below 8.

- *RoBERTa-paraphrased reviews*: all of them are flagged as being "AI Generated" (thus, recall=100%)
- *DIPPER-paraphrased reviews*: 797/800 are flagged as being "AI Generated"; the remaining three have a "Mixed" output (Pangram found that, for these three, 67% of the segments are AI-generated).
- *Human-submitted reviews*: 3,272/3,330 (98.3%) are flagged as being "Human"; whereas 34 (1%) are flagged as being "AI Generated" (of which 28 submitted for ICLR'25, and 6 for ICLR'24); and 24 (0.7%) as being "Mixed" (of which 17 were submitted for ICLR'25, and 7 for ICLR'24).

We can thus conclude that Pangram v3.3.2 seems to work well in detecting LLM-generated reviews, at least from a detection-rate viewpoint. With regards to the false-positive rate, we cannot be 100% confident that all human-submitted reviews in our sample of 800 papers do not involve any form of AI-generated text. However, the fact that the only reviews flagged as being not fully human-written were submitted for ICLR'24 and ICLR'25 (i.e., after the widespread diffusion of LLMs) suggest that Pangram does not exhibit a high FPR: even if all these detections are erroneous, they represent at most a 1.7% of our analysed samples (3,330).

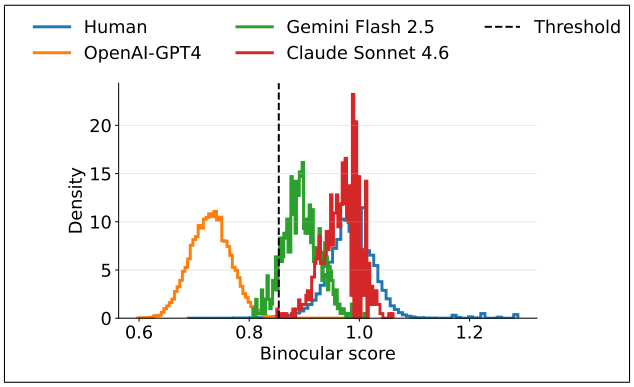


Figure 9: Distribution of the scores provided by Binoculars on the subset of 800 LLM-generated reviews (refer to §C.3)

D Description of our custom topic-model-based paraphrasing mechanism

Rationale. As a more controlled alternative to LLM-based paraphrasing, we implemented a context-aware lexical replacement procedure using the language model RoBERTa [47]. The motivation for this approach was to introduce more static variation into the reviews and avoid the unpredictable paraphrasing. At a high level, the method substitutes selected words and phrases by relying on a lookup table. This makes the transformation more constrained, but also more controllable and more transparent, since each change can be traced back to a specific replacement decision.

Justification. The decision to rely on a topic model is because straightforward application of synonym-replacement strategies are not viable. For instance, one could simply replace words with their synonyms. We tried this: our preliminary attempts showed that this led to edits that lost meaning, intent, or sounded unnatural. Many words are extremely context-dependent. For example "work" may refer to a study, but it is also a synonym for "job". Furthermore, it may also function as a verb, as in "the approach does not work." Likewise "incomplete" is a valid synonym for "insufficient," but if replaced in a sentence such as "The experiments are insufficient", the original's meaning would be substantially altered. Put simply, random replacements introduced grammatical/syntactical errors, altered the viewers stance, or changed the strength of criticism or a claim. For this reason, we employed a more conservative and context-aware approach that has a whole validation pipeline each edit has to go through before it is accepted.

Workflow. Let us explain how our method processes each review. First, the review is split into paragraphs and sentences, preserving the original review's paragraph structure. For each sentence, we identify replacement targets from two lookup tables. The first table contains phrase substitutions, such as replacing "in order

to" with "to" or "for example" with "e.g.". The second table contains word substitutions that are common in academic review writing, such as replacing "paper" with "manuscript", "method" with "approach", or "results" with "findings". We restrict the set of possible substitutions to the vocabulary rather than allowing arbitrary model-generated replacements, since peer reviews contain technical claims whose meaning could be changed by careless paraphrasing. To further reduce unsafe substitutions, we check the grammatical role of each candidate word before allowing it to be replaced. By using NLTK[11], each sentence is first tokenized and annotated with grammatical roles. This allows us to distinguish, whether a word is being used as a noun, verb, adjective, or adverb. A replacement is considered only when the observed grammatical role of the word matches the role expected for that replacement category. This prevents substitutions that are superficially plausible but grammatically or semantically inappropriate. We also maintain a list of technical or domain-specific terms, including model names, metrics, dataset-related terms, and common machine-learning terminology: these words are protected from replacement because altering them could change the technical content of the review. For the replacement candidates, we use distilroberta-base[12] as a masked language model to check whether it plausible in the sentence context. Specifically, the target word or phrase is masked, and the model scores the candidate substitutes according to how well they fit the masked position. Additionally, we apply safety checks that aim to preserve stance and strength of the review. A replacement is rejected if it changes negation, uncertainty, claim strength, or intensity. For example, this rejects changes that remove negation, make a soft statement sound very certain, or turn a minor criticism into a stronger one. We also avoid replacements in citation, formulas, or reference contexts. Finally, the replacement candidates have to pass a semantic-similarity filter: after a substitution is applied, the modified sentence is compared to the original using a all-MiniLM-L6-v2, a widely-used sentence similarity model[13]. The replacement is only accepted if the similarity score is at least 0.90. The number of replacements can also be limited both globally (per review) and locally (per sentence) to prevent excessive changes. Preliminary manual assessments with these thresholds confirmed they yielded edited reviews that preserved the original's meaning.

Remarks. This approach produces paraphrased reviews through a sequence of small, controlled and validated edits. Mimicking a human reviewer who checks the generated review and selectively changes some of the wording while preserving its intended meaning. The RoBERTa-based method is less flexible compared with the LLM-based editing, but more deterministic and easier to verify as each output is accompanied by logs that list the substitutions. Importantly, we used a fixed random seed to apply our method, making our results entirely reproducible. In this way, we generated 800 paraphrased reviews that differ lexically from the originals while preserving the content, stance, and meaning.

[11] https://www.nltk.org/
[12] https://huggingface.co/distilbert/distilroberta-base
[13] https://huggingface.co/sentence-transformers/all-MiniLM-L6-v2