Gen-Review: A Dataset and Large-scale Study of AI-Generated and Human-Authored Peer Reviews

Anonymous Author(s)
Affiliation
Address
email

Abstract

How does the increased adoption of Large Language Models (LLMs) impact the scientific peer review? This multifaceted question is fundamental to the integrity and outcomes of the scientific process. Timely evidence suggests LLMs may have already been used for peer-review, e.g., at the 2024 International Conference of Learning Representations (ICLR), and the LLMs' integration in peer-review was confirmed by various editorial boards (including that of ICLR'25). To seek answers, a comprehensive dataset is needed, but lacking until now. We therefore present Gen-Review, the largest dataset of LLM-written reviews so far. Our dataset includes 81K reviews generated for all submissions to the 2018–2025 editions of the ICLR and by providing the LLM with three independent prompts: a negative, a positive, and a neutral one. Gen-Review also links to the papers and the conference reviews thereby enabling a broad range of investigations. We make a start and use Gen-Review to scrutinize: if LLMs exhibit bias in reviewing (they do); 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 a papers' final outcome (happens only for accepted papers). Link to Gen-Review: https://anonymous.4open.science/r/gen_review/.

1 Introduction

Since the release of ChatGPT in Q4 2022 [35], Large Language Models (LLMs) are revolutionizing many areas of our society [11]. 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 [21, 9, 30, 18, 8, 23, 47, 26, 46]. From a broader perspective, LLMs are also expected to have a profound *impact on science in general*, regardless of their specific fields [6, 29].

LLMs can affect scientific work in various ways. They can be used to revise text [12], summarize prior literature [3], or implement an experimental pipeline or its parts [16]. The use of LLMs for scientific work has initially faced ample criticism [2, 19, 31]. However, LLMs are a valuable asset to researchers [6, 11] 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 [1].

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., [28, 4, 25, 41, 45, 37, 24]. As an almost anecdotal finding, the study of Liang et al. [28] 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 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.

Submitted to 39th Conference on Neural Information Processing Systems (NeurIPS 2025). Do not distribute.

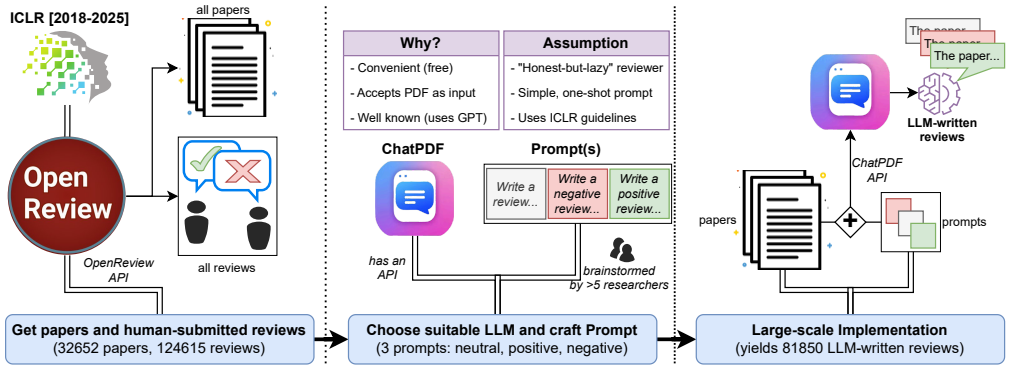


Figure 1: **The workflow to create `Gen-Review`.** We rely on the papers submitted to the [2018–2025] editions of ICLR (we also collect all of their human-submitted reviews). Then, we craft three simple prompts and we leverage the ChatPDF API to generate our large-scale dataset of LLM-written reviews.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

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 [48]. As a result, 27% of reviewers confronted with such feedback updated their reviews [40]. Yet, the overall sentiment towards a large-scale deployment of LLMs for reviewing remains mixed, with opinions ranging from "inevitable" to "a disaster" [32].

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 a state-of-the-art detector of LLM-generated text, *Binoculars* [15], on our collected data.

Contributions. In summary, our paper makes the following 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 describe the creation of `Gen-Review` in Section §3. Exploratory analyses are elucidated in Section §4. We discuss our results and provide avenues for future work in Section §5.

2 Preliminaries, Goals, and Motivation

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 [35]. 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 [33]. 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 [13, 42, 7, 27].

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 [27], or in emails [34]—there has been a growing interest in the development of *automated detectors* of LLM-generated media [39]. Abundant prior works have developed various tools that can estimate whether a given input was generated by an AI (e.g., [22, 5]). For instance, Hans et al. [15] 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 [43], 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 [32, 48], *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 [14]. 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 [14].

2.2 Problem Statement and Research Workflow

At a high-level, our contributions are motivated by two complementary reasons:

- the potentially inescapable integration of LLMs in (parts of) the peer-review process [32], which requires improving our generic understanding of LLM-generated reviews; and
- the necessity of identifying cases of misconduct wherein reviewers relied on LLMs without disclosure (thereby failing to uphold the authors' right to be informed [14]), 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

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

[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.

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 detector (Binoculars [15]) identify the reviews in* `Gen-Review`*?* (and how does it perform on the human-submitted reviews?)

Altogether, answering these RQ helps us better understand some facets of using LLMs for peer-review.

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 the seminal work by Liang et al. [28] 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. [25], 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 directly 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, such as benchmarking the effectiveness of detectors of LLM-written peer reviews.

Small-scale analyses. In their recent work, Thelwall et al. [41] assess ChatGPT's ability to predict the outcome of some papers submitted to ICLR'17 (collected in [17]). Similarly, the authors of [37] carried out a study in which human reviewers' assessments were compared to those of GPT-4 in a total of 325 abstracts, finding alignment only for the best submissions. The analyses of both of these works are preliminary and limited in scale, preventing generalizable conclusions. Our analysis is performed on a much larger scale, aiming to provide more robust empirical evidence and uncover systematic patterns in LLM-assisted reviewing.

Limited-scope datasets of LLM-written reviews. The closest works to our paper are those of Yu et al. [45] and Kumar et al. [24]. Both ultimately seek to propose new methods to detect LLM-written reviews, and such methods were tested also on (genuine) LLM-written reviews based on ICLR submissions. However, the datasets used for such evaluations have a much more limited scope than our proposed `Gen-Review`. For instance, Yu et al. [45] 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 (ours is 32'652). Whereas Kumar et al. [24] 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 [38, 10, 44]; or works that focus on the detection of LLM-written *papers*—and not reviews—such as [31]. Finally, we stress that our work is in no manner related to the detection of "fake reviews" in online platforms (e.g., online marketplaces [20, 36]).

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

We first outline the necessary requirements to reach our goal (see §2.2) and then explain how we collected the backbone of `Gen-Review`, motivating our decisions.

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—which are necessary to facilitate comparison against LLM-written ones.

Collection. We determined that the 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 [32].[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

[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.

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	

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 [24] (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, 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, as well as *(b)* the human-submitted reviews available on OpenReview. The overall statistics of our `Gen-Review` are shown in Table 1.

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

[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 [33])

[5]We issued our queries between February and April 2025: according to the ChatPDF documentation, the queries were routed to either GPT-4o or GPT-4o-mini. We are unfortunately unable to control which specific model was used, but no change was made to ChatPDF during our considered time frame.

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 [25].

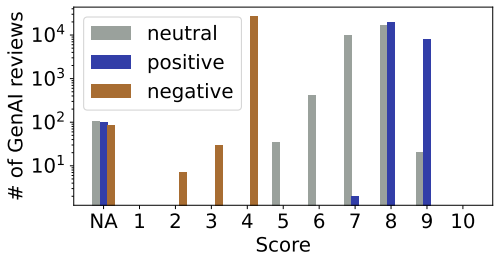
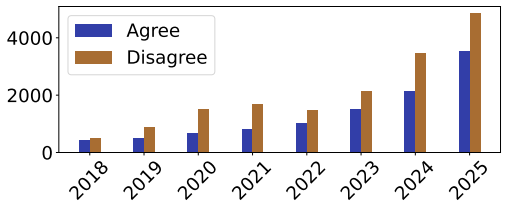
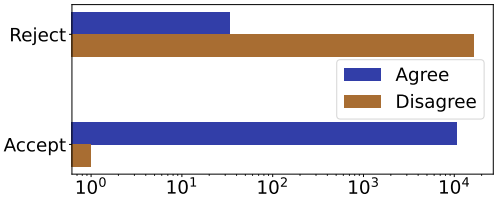


Figure 2: **Rating of LLM-written reviews** in `Gen-Review` for each considered prompt. Ratings follow the ICLR 1–10 scale (N/A denotes cases without a rating in the LLM-written review).

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 [41]) further reinforces our answer to RQ1: LLMs tend to favor acceptance to a much larger extent than human-driven program committees. Ultimately, 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).

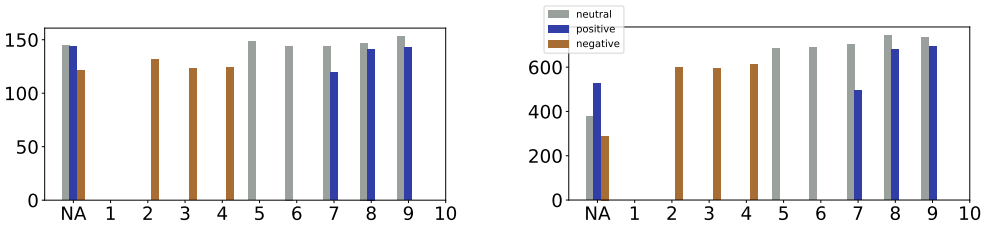


(a) Agreement over the years (y-axis: # of papers).



(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.

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). A potential explanation for this discrepancy is that the LLM interpreted that the 800–1000 words should include both the "summary" and the "main review". Still, even by adding the lengths of the summary and of the main review, we do not always obtain a text within our specified margins. 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* [15] due to its popularity (albeit we acknowledge that other tools exist, such as [24]). This detector works by providing a score for the input text, and whether such is above a given threshold ($\approx$ 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 [28, 25]. 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.

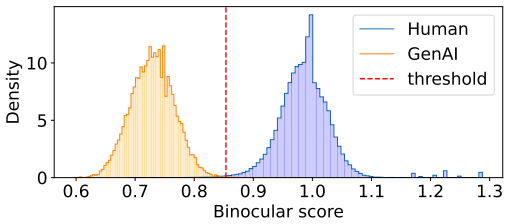


Figure 5: **Assessment of Binoculars** on our AI-generated reviews, and on human-submitted ones.

5 Discussion

5.1 Limitations

`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. Secondly, the reviews in `Gen-Review` have been created by using a single LLM service (i.e., ChatPDF); moreover, we had no control on which model was used to produce each review (ChatPDF would automatically switch between GPT-4o and GPT-4o-mini) meaning that our dataset is not suited to explore the effectiveness of other LLMs (Gemini, Claude, or others).

5.2 Broader Impact

In a sense, our findings suggest that our envisioned "honest-but-lazy" reviewer can skew the outcome of the paper selection process due to an overwhelming positive bias of the underlying LLM. Further, we have further shown that LLMs can be used by a "not-very-honest" 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. Fortunately, some existing detectors can reliably (with some false positives) flag LLM-generated reviews—when no attempt was made to alter the text, or when issued via simple prompts. From a security standpoint, we endorse taking into account the possibility that some "adversarial reviewers" may attempt to evade the detection process.

5.3 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 only considered Binoculars [15], but many more detectors of LLM-generated text exist (e.g., [22, 5]). 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., [24], which we have also tested with a few dozen reviews from `Gen-Review`, and it seem to work very well!).[6]

Evaluating (and improving) the LLM review quality. We mostly focused on quantitatively analysing, at a very high level, the LLM-written reviews in `Gen-Review`, prioritizing the investigation of whether such reviews had some bias. 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 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. However, to maximize reproducibility and facilitate further research, we have released our prompts. Researchers can thus expand our dataset in various directions, e.g., using the same prompts by requesting other LLMs to review the same papers; or by using different papers. It would be intriguing to, e.g., see if our findings can also map to other disciplines, venues, or LLMs.

[6] We have also studied (Table 4) the prevalence of the words highlighted by Liang et al. [28] across the LLM-written reviews in `Gen-Review`: many of our reviews include these words, especially "innovative".

References

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

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

[3] 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.

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

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

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

[7] 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.

[8] 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.

[9] 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.

[10] 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.

[11] 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. *International journal of information management*, 2023.

[12] 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.

[13] 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.

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

[15] 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.

[16] 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.

[17] 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.

[18] 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.

[19] 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.

[20] 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.

[21] 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.

[22] 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.

[23] 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.

[24] 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.

[25] 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.

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

[27] 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.

[28] 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.

[29] 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.

[30] 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.

[31] 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 *Proceedings of the 3rd Workshop on Trustworthy Natural Language Processing (TrustNLP 2023)*, 2023.

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

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

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

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

[36] 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.

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

[38] 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.

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

[40] 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.

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

[42] 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.

[43] 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.

[44] 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.

[45] 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.

[46] 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.

[47] 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.

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

A Reviewer Prompts

We display here the prompts we have used to setup ChatPDF, and generate un- (Prompt 1), positively- (Prompt 2), and negatively-biased (Prompt 3) reviews.

```
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.

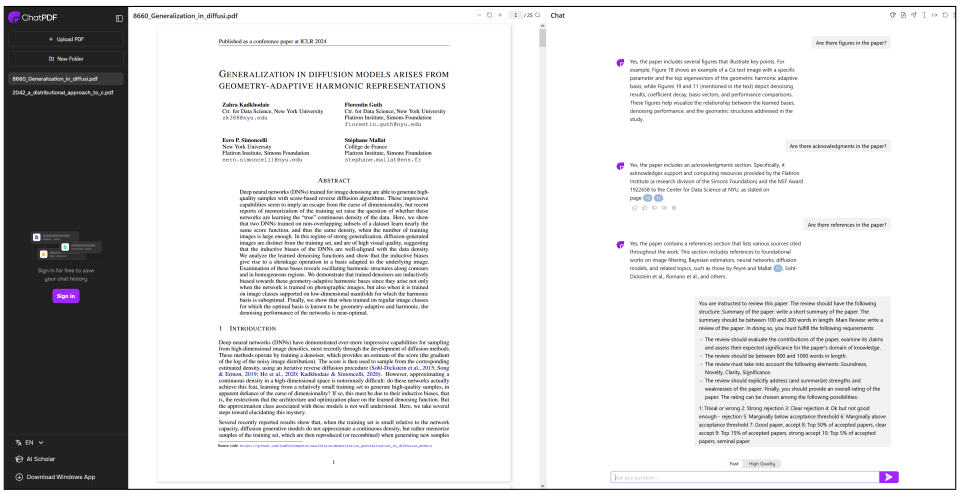


Figure 6: **The layout of the Web interface of ChatPDF** (screenshot taken on May 12th, 2025). Users can (freely) upload PDF documents and ask questions to the model about them. In the figure, we asked some questions (showing that the model can "interpret" figures) and provided our "neutral" prompt to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

Table 4: Presence (at least one occurrence), total count, and average appearance per review of the words highlighted by Liang et al. [28] 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

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.

NeurIPS Paper Checklist

1. **Claims**

   Question: Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope?

   Answer: [Yes]

   Justification: Yes. We have outlined the contributions in the Introduction, and they are described in Section 3 and Section 4 (we discuss the shortcomings of prior work to support our "novelty" in Section 2)

   Guidelines:
   - The answer NA means that the abstract and introduction do not include the claims made in the paper.
   - The abstract and/or introduction should clearly state the claims made, including the contributions made in the paper and important assumptions and limitations. A No or NA answer to this question will not be perceived well by the reviewers.
   - The claims made should match theoretical and experimental results, and reflect how much the results can be expected to generalize to other settings.
   - It is fine to include aspirational goals as motivation as long as it is clear that these goals are not attained by the paper.

2. **Limitations**

   Question: Does the paper discuss the limitations of the work performed by the authors?

   Answer: [Yes]

   Justification: We have a dedicated "Limitations" subsection (Section 5.1) wherein we explain the major limitations of our contribution.

   Guidelines:
   - The answer NA means that the paper has no limitation while the answer No means that the paper has limitations, but those are not discussed in the paper.
   - The authors are encouraged to create a separate "Limitations" section in their paper.
   - The paper should point out any strong assumptions and how robust the results are to violations of these assumptions (e.g., independence assumptions, noiseless settings, model well-specification, asymptotic approximations only holding locally). The authors should reflect on how these assumptions might be violated in practice and what the implications would be.
   - The authors should reflect on the scope of the claims made, e.g., if the approach was only tested on a few datasets or with a few runs. In general, empirical results often depend on implicit assumptions, which should be articulated.
   - The authors should reflect on the factors that influence the performance of the approach. For example, a facial recognition algorithm may perform poorly when image resolution is low or images are taken in low lighting. Or a speech-to-text system might not be used reliably to provide closed captions for online lectures because it fails to handle technical jargon.
   - The authors should discuss the computational efficiency of the proposed algorithms and how they scale with dataset size.
   - If applicable, the authors should discuss possible limitations of their approach to address problems of privacy and fairness.
   - While the authors might fear that complete honesty about limitations might be used by reviewers as grounds for rejection, a worse outcome might be that reviewers discover limitations that aren't acknowledged in the paper. The authors should use their best judgment and recognize that individual actions in favor of transparency play an important role in developing norms that preserve the integrity of the community. Reviewers will be specifically instructed to not penalize honesty concerning limitations.

3. **Theory assumptions and proofs**

   Question: For each theoretical result, does the paper provide the full set of assumptions and a complete (and correct) proof?

   Answer:[NA]

Justification: We do not have "theoretical results", so this does not apply.

Guidelines:

- The answer NA means that the paper does not include theoretical results.
- All the theorems, formulas, and proofs in the paper should be numbered and cross-referenced.
- All assumptions should be clearly stated or referenced in the statement of any theorems.
- The proofs can either appear in the main paper or the supplemental material, but if they appear in the supplemental material, the authors are encouraged to provide a short proof sketch to provide intuition.
- Inversely, any informal proof provided in the core of the paper should be complemented by formal proofs provided in appendix or supplemental material.
- Theorems and Lemmas that the proof relies upon should be properly referenced.

4. **Experimental result reproducibility**

Question: Does the paper fully disclose all the information needed to reproduce the main experimental results of the paper to the extent that it affects the main claims and/or conclusions of the paper (regardless of whether the code and data are provided or not)?

Answer: [Yes]

Justification: We have released the prompts used to generate our dataset, and the other data (i.e., papers and reviews) are publicly available. We note that complete reproducibility is not possible due to the intrinsic randomness of LLMs. The code for the plots is in our repository.

Guidelines:

- The answer NA means that the paper does not include experiments.
- If the paper includes experiments, a No answer to this question will not be perceived well by the reviewers: Making the paper reproducible is important, regardless of whether the code and data are provided or not.
- If the contribution is a dataset and/or model, the authors should describe the steps taken to make their results reproducible or verifiable.
- Depending on the contribution, reproducibility can be accomplished in various ways. For example, if the contribution is a novel architecture, describing the architecture fully might suffice, or if the contribution is a specific model and empirical evaluation, it may be necessary to either make it possible for others to replicate the model with the same dataset, or provide access to the model. In general. releasing code and data is often one good way to accomplish this, but reproducibility can also be provided via detailed instructions for how to replicate the results, access to a hosted model (e.g., in the case of a large language model), releasing of a model checkpoint, or other means that are appropriate to the research performed.
- While NeurIPS does not require releasing code, the conference does require all submissions to provide some reasonable avenue for reproducibility, which may depend on the nature of the contribution. For example
  (a) If the contribution is primarily a new algorithm, the paper should make it clear how to reproduce that algorithm.
  (b) If the contribution is primarily a new model architecture, the paper should describe the architecture clearly and fully.
  (c) If the contribution is a new model (e.g., a large language model), then there should either be a way to access this model for reproducing the results or a way to reproduce the model (e.g., with an open-source dataset or instructions for how to construct the dataset).
  (d) We recognize that reproducibility may be tricky in some cases, in which case authors are welcome to describe the particular way they provide for reproducibility. In the case of closed-source models, it may be that access to the model is limited in some way (e.g., to registered users), but it should be possible for other researchers to have some path to reproducing or verifying the results.

5. **Open access to data and code**

Question: Does the paper provide open access to the data and code, with sufficient instructions to faithfully reproduce the main experimental results, as described in supplemental material?

Answer: [Yes]

Justification: Our dataset is provided at https://doi.org/10.7910/DVN/PYDPEZ, and all the code is available at https://anonymous.4open.science/r/gen_review/. The README of the code also clearly depict how the dataset is shaped. Also, we release the code as zip in the supplementary material.

Guidelines:

- The answer NA means that paper does not include experiments requiring code.
- Please see the NeurIPS code and data submission guidelines (https://nips.cc/public/guides/CodeSubmissionPolicy) for more details.
- While we encourage the release of code and data, we understand that this might not be possible, so "No" is an acceptable answer. Papers cannot be rejected simply for not including code, unless this is central to the contribution (e.g., for a new open-source benchmark).
- The instructions should contain the exact command and environment needed to run to reproduce the results. See the NeurIPS code and data submission guidelines (https://nips.cc/public/guides/CodeSubmissionPolicy) for more details.
- The authors should provide instructions on data access and preparation, including how to access the raw data, preprocessed data, intermediate data, and generated data, etc.
- The authors should provide scripts to reproduce all experimental results for the new proposed method and baselines. If only a subset of experiments are reproducible, they should state which ones are omitted from the script and why.
- At submission time, to preserve anonymity, the authors should release anonymized versions (if applicable).
- Providing as much information as possible in supplemental material (appended to the paper) is recommended, but including URLs to data and code is permitted.

6. **Experimental setting/details**

Question: Does the paper specify all the training and test details (e.g., data splits, hyper-parameters, how they were chosen, type of optimizer, etc.) necessary to understand the results?

Answer: [NA]

Justification: We do not have experiments, just exploratory analyses done via simple SQL queries and pattern-matching scripts that can be found in https://anonymous.4open.science/r/gen_review/. Most of the retrieved content can be fecthed by querying the provied SQLite database.

Guidelines:

- The answer NA means that the paper does not include experiments.
- The experimental setting should be presented in the core of the paper to a level of detail that is necessary to appreciate the results and make sense of them.
- The full details can be provided either with the code, in appendix, or as supplemental material.

7. **Experiment statistical significance**

Question: Does the paper report error bars suitably and correctly defined or other appropriate information about the statistical significance of the experiments?

Answer: [No]

Justification: the experiments we describe in §4 does not require the computation of confidence intervals or other statistical tests. Our analysis focuses on describing relevant metrics of the collected data.

Guidelines:

- The answer NA means that the paper does not include experiments.
- The authors should answer "Yes" if the results are accompanied by error bars, confidence intervals, or statistical significance tests, at least for the experiments that support the main claims of the paper.
- The factors of variability that the error bars are capturing should be clearly stated (for example, train/test split, initialization, random drawing of some parameter, or overall run with given experimental conditions).

- The method for calculating the error bars should be explained (closed form formula, call to a library function, bootstrap, etc.)
- The assumptions made should be given (e.g., Normally distributed errors).
- It should be clear whether the error bar is the standard deviation or the standard error of the mean.
- It is OK to report 1-sigma error bars, but one should state it. The authors should preferably report a 2-sigma error bar than state that they have a 96% CI, if the hypothesis of Normality of errors is not verified.
- For asymmetric distributions, the authors should be careful not to show in tables or figures symmetric error bars that would yield results that are out of range (e.g. negative error rates).
- If error bars are reported in tables or plots, The authors should explain in the text how they were calculated and reference the corresponding figures or tables in the text.

8. **Experiments compute resources**

   Question: For each experiment, does the paper provide sufficient information on the computer resources (type of compute workers, memory, time of execution) needed to reproduce the experiments?

   Answer: [NA]

   Justification: We do not have any experiment, and our analyses are trivial to carry out from a computational perspective.

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - The paper should indicate the type of compute workers CPU or GPU, internal cluster, or cloud provider, including relevant memory and storage.
   - The paper should provide the amount of compute required for each of the individual experimental runs as well as estimate the total compute.
   - The paper should disclose whether the full research project required more compute than the experiments reported in the paper (e.g., preliminary or failed experiments that didn't make it into the paper).

9. **Code of ethics**

   Question: Does the research conducted in the paper conform, in every respect, with the NeurIPS Code of Ethics https://neurips.cc/public/EthicsGuidelines?

   Answer: [Yes]

   Justification: Yes. Our dataset is created by using publicly-available data as a basis, collected in compliance with existing ToS.

   Guidelines:

   - The answer NA means that the authors have not reviewed the NeurIPS Code of Ethics.
   - If the authors answer No, they should explain the special circumstances that require a deviation from the Code of Ethics.
   - The authors should make sure to preserve anonymity (e.g., if there is a special consideration due to laws or regulations in their jurisdiction).

10. **Broader impacts**

    Question: Does the paper discuss both potential positive societal impacts and negative societal impacts of the work performed?

    Answer: [Yes]

    Justification: Our work does enable to improve our understanding of using LLMs for peer-review. It intrinsically has a "broader impact". We discuss the "Broader Impact" in Section 5.2..

    Guidelines:

    - The answer NA means that there is no societal impact of the work performed.
    - If the authors answer NA or No, they should explain why their work has no societal impact or why the paper does not address societal impact.

- Examples of negative societal impacts include potential malicious or unintended uses (e.g., disinformation, generating fake profiles, surveillance), fairness considerations (e.g., deployment of technologies that could make decisions that unfairly impact specific groups), privacy considerations, and security considerations.
- The conference expects that many papers will be foundational research and not tied to particular applications, let alone deployments. However, if there is a direct path to any negative applications, the authors should point it out. For example, it is legitimate to point out that an improvement in the quality of generative models could be used to generate deepfakes for disinformation. On the other hand, it is not needed to point out that a generic algorithm for optimizing neural networks could enable people to train models that generate Deepfakes faster.
- The authors should consider possible harms that could arise when the technology is being used as intended and functioning correctly, harms that could arise when the technology is being used as intended but gives incorrect results, and harms following from (intentional or unintentional) misuse of the technology.
- If there are negative societal impacts, the authors could also discuss possible mitigation strategies (e.g., gated release of models, providing defenses in addition to attacks, mechanisms for monitoring misuse, mechanisms to monitor how a system learns from feedback over time, improving the efficiency and accessibility of ML).

11. **Safeguards**

    Question: Does the paper describe safeguards that have been put in place for responsible release of data or models that have a high risk for misuse (e.g., pretrained language models, image generators, or scraped datasets)?

    Answer: [NA]

    Justification: This does not apply, as we release no models.

    Guidelines:

    - The answer NA means that the paper poses no such risks.
    - Released models that have a high risk for misuse or dual-use should be released with necessary safeguards to allow for controlled use of the model, for example by requiring that users adhere to usage guidelines or restrictions to access the model or implementing safety filters.
    - Datasets that have been scraped from the Internet could pose safety risks. The authors should describe how they avoided releasing unsafe images.
    - We recognize that providing effective safeguards is challenging, and many papers do not require this, but we encourage authors to take this into account and make a best faith effort.

12. **Licenses for existing assets**

    Question: Are the creators or original owners of assets (e.g., code, data, models), used in the paper, properly credited and are the license and terms of use explicitly mentioned and properly respected?

    Answer: [Yes]

    Justification: Yes. We are complying with OpenReview ToS, and all data we used is publicly available already on OpenReview. We are not claiming authorship of the papers in our dataset (whose details are available on OpenReview).

    Guidelines:

    - The answer NA means that the paper does not use existing assets.
    - The authors should cite the original paper that produced the code package or dataset.
    - The authors should state which version of the asset is used and, if possible, include a URL.
    - The name of the license (e.g., CC-BY 4.0) should be included for each asset.
    - For scraped data from a particular source (e.g., website), the copyright and terms of service of that source should be provided.
    - If assets are released, the license, copyright information, and terms of use in the package should be provided. For popular datasets, `paperswithcode.com/datasets` has curated licenses for some datasets. Their licensing guide can help determine the license of a dataset.

- For existing datasets that are re-packaged, both the original license and the license of the derived asset (if it has changed) should be provided.
- If this information is not available online, the authors are encouraged to reach out to the asset's creators.

13. **New assets**

    Question: Are new assets introduced in the paper well documented and is the documentation provided alongside the assets?

    Answer: [Yes]

    Justification: Yes, everything is documented in our repository (at https://anonymous.4open.science/r/gen_review/), and it is also attached as supplementary material.

    Guidelines:

    - The answer NA means that the paper does not release new assets.
    - Researchers should communicate the details of the dataset/code/model as part of their submissions via structured templates. This includes details about training, license, limitations, etc.
    - The paper should discuss whether and how consent was obtained from people whose asset is used.
    - At submission time, remember to anonymize your assets (if applicable). You can either create an anonymized URL or include an anonymized zip file.

14. **Crowdsourcing and research with human subjects**

    Question: For crowdsourcing experiments and research with human subjects, does the paper include the full text of instructions given to participants and screenshots, if applicable, as well as details about compensation (if any)?

    Answer: [NA]

    Justification: We do not do human-subject research.

    Guidelines:

    - The answer NA means that the paper does not involve crowdsourcing nor research with human subjects.
    - Including this information in the supplemental material is fine, but if the main contribution of the paper involves human subjects, then as much detail as possible should be included in the main paper.
    - According to the NeurIPS Code of Ethics, workers involved in data collection, curation, or other labor should be paid at least the minimum wage in the country of the data collector.

15. **Institutional review board (IRB) approvals or equivalent for research with human subjects**

    Question: Does the paper describe potential risks incurred by study participants, whether such risks were disclosed to the subjects, and whether Institutional Review Board (IRB) approvals (or an equivalent approval/review based on the requirements of your country or institution) were obtained?

    Answer: [NA]

    Justification: We do not need an IRB because there is no human-subject research done in our paper.

    Guidelines:

    - The answer NA means that the paper does not involve crowdsourcing nor research with human subjects.
    - Depending on the country in which research is conducted, IRB approval (or equivalent) may be required for any human subjects research. If you obtained IRB approval, you should clearly state this in the paper.
    - We recognize that the procedures for this may vary significantly between institutions and locations, and we expect authors to adhere to the NeurIPS Code of Ethics and the guidelines for their institution.
    - For initial submissions, do not include any information that would break anonymity (if applicable), such as the institution conducting the review.

16. **Declaration of LLM usage**

Question: Does the paper describe the usage of LLMs if it is an important, original, or non-standard component of the core methods in this research? Note that if the LLM is used only for writing, editing, or formatting purposes and does not impact the core methodology, scientific rigorousness, or originality of the research, declaration is not required.

Answer: [Yes]

Justification: We used a LLM to generate our dataset—which is meant for this specific purpose (i.e., providing researchers with LLM-generated data to evaluate the LLM capabilities at generating such data). Aside from this, we did not use a LLM at all.

Guidelines:

- The answer NA means that the core method development in this research does not involve LLMs as any important, original, or non-standard components.
- Please refer to our LLM policy (https://neurips.cc/Conferences/2025/LLM) for what should or should not be described.