Can LLMs Evaluate Research Responsibility? A Benchmark on Conference Submission Checklists

Anonymous ACL submission

Abstract

For a few years, some of the most important conferences have started using checklists as a support for author submissions. The utility of these checklists is twofold. First, they can be used as a self-assessment tool for authors, providing them references on how to correctly address the compliance of their submissions. In addition, reviewers can also use checklists to assist them during the review task. LLMs can be a powerful tool for providing assistance for checklist completion due to their capacity to emulate human-like reasoning. This development has caught the attention of major conferences in the field, such as NeurIPS, which has included an LLM assistant in its submission process for its 2026 edition. However, most of these assistants rely on closed-source models, making them unavailable to a wide range of researchers and conferences. In order to address this issue, this paper presents a benchmark on the performance of several open-source LLMs, comprising a varying range of sizes and complexities, for the checklist completion task. In addition, a tool named *Prêt-à-submit* that encapsulates the full pipeline is provided. The goal of this tool is to serve as an open-source checklist-based assistant for authors. The benchmark results show that, while for some checklist points LLMs can accurately respond and simulate human responses, there is still a significant gap in the responses provided by the authors and LLMs. Moreover, experimentation shows discrepancies between the results provided by the different models, which are especially noticeable in smaller LLMs.

1 Introduction

Peer review is one of the pillars of scientific publication. By subjecting research outputs to independent expert evaluation, peer review not only acts as a quality control mechanism but also promotes the refinement of manuscripts through constructive feedback (Kelly et al., 2014). This collaborative process improves methodological rigor, supports the identification of potential biases, and facilitates dissemination. Although essential, peer review can be very time-consuming, as it is performed manually by experts and requires time and dedication. Moreover, since it is a voluntary task, several researchers decline to participate in this process (Kelly, 2023).

In order to assist reviewers with this tedious process, some conferences have started providing review checklists to authors for them to self-verify the quality and reproducibility of their work. In the context of AI research, some of the most relevant conferences in the area, such as NeurIPS (NeurIPS, 2025) or AAAI (AAAI, 2024), also have their own checklists. The utility of these checklists is twofold. First, they can be helpful for reviewers since they provide guidelines on aspects to assess during the review process. Secondly, they can be useful to authors to self-evaluate their work before making a submission, and subsequently correct and improve their work beforehand.

Additionally, the rapid rise of large language models (LLMs) has significantly impacted the research landscape (Liao et al., 2024). Models such as GPT-3.5 or GPT-4 (OpenAI et al., 2024) and, especially, their chat version, ChatGPT, have drastically changed the way we work and research. In addition to ChatGPT, several LLM-powered research-oriented tools have appeared to assist researchers in tasks such as searching (Srinivas et al., 2022), writing (Paperguide, 2024), and reviewing (Heckel et al., 2023). This has led to the establishment of guidelines on how these powerful tools should be used in research such that they can help researchers without replacing them. These guidelines vary between forums. For example, the well-known research journal *Nature* establishes in its publication policy that safe AI tools can be used to assist reviewers in their process but should not be used to generate entire reviews since the knowledge of re-

searchers is invaluable and irreplaceable (Nature, 2024).

In this context, this paper presents a benchmark of the performance of nine open-source LLMs to review checklist completion, alongside the introduction of *Prêt-à-Submit*, a tool built around the proposed pipeline to provide feedback to authors about their submission process. The results show that, while LLMs can be a powerful tool and can provide reliable answers for certain points, especially those related to content that should be specifically mentioned within the paper, their performance is still far from resembling human criteria. In addition, the results reveal the existence of systematic biases within all models, thus hindering their usage as a fully autonomous agent for checklist completion.

2 Related Works

The PRISMA statement (Page et al., 2021) pioneered the use of review checklists to enhance research paper quality. This practice has now spread to top-tier AI conferences like NeurIPS (NeurIPS, 2025) and AAAI (AAAI, 2024), which have started including their own checklists. Later, Dodge et al. (2019) and Dodge and Smith (2020) introduced checklists focused specifically on AI and NLP reproducibility, which were evaluated by Magnusson et al. (2023). Their findings show that checklists not only improve submission quality and encourage authors to provide more details, but also prompt other conference committees to adopt similar tools.

Recently, LLMs have significantly shifted the research landscape, subsequently affecting the review process. While Evans et al. (2024) argue that expert human knowledge remains irreplaceable for analyzing articles, others explore LLMs for peer review generation. For instance, Liang et al. (2023) found that LLM-generated reviews exhibit biases due to a lack of deep understanding, though they remain useful for preliminary feedback. Regarding checklist verification, Liu and Shah (2023) demonstrated that GPT-4 could verify NeurIPS checklists with 86.6% precision. More recently, Goldberg et al. (2024) automated checklist completion for NeurIPS'24 submissions; however, they evaluated user experience rather than comparing the LLM's outputs with actual author responses, leaving the system's reliability unassessed.

To assist researchers, several LLM-based tools have emerged. ConfReady utilizes Retrieval-Augmented Generation (RAG) to help authors complete checklists (Galarnyk et al., 2025), while the Stanford Agentic Reviewer employs an agent-driven workflow to provide automated feedback enriched with literature search (Jiang and Ng, 2026). However, both tools rely exclusively on commercial LLMs, limiting accessibility and, additionally, ConfReady's evaluation is restricted to a small, historical dataset, thus the achieved results cannot be extrapolated. Despite this issue, due to their positive results, top-tier conferences such as NeurIPS have also started promoting their own LLM-based tools for paper self-review assistance (Deisenroth et al., 2026).

3 Benchmarking LLMs for Automatic Checklist Completion

This work presents a benchmark for the performance of nine different open-source LLMs for review checklist completion. The goal is to assess whether LLMs can accurately reflect the behavior of human researchers for checklist completion, as well as whether they are capable of understanding the content of the papers and answering questions that require a deeper level of comprehension (i.e., whether the authors explore the limitations of the approach, or whether the abstract clearly summarizes the content). We also seek to evaluate if open-source models can be used as the backbone of author assistance utilities.

Human-completed checklists are required to compare whether the LLM achieves the same conclusions on the aspects evaluated. Therefore, NeurIPS'24 accepted papers are selected to conduct the experimentation, since they include, in addition to the paper itself, the checklist filled by the authors as an appendix. Therefore, the checklist submitted by the author can be considered the ground truth, and the responses provided by the LLMs to the same questions are expected to be equal. According to the conference guidelines, the authors must respond to each checklist point with "Yes", "No", and "N/A". Additionally, authors can provide evidence and indicate where in the article a certain criterion is met or not. Although it would have been interesting to consider reviewer-filled checklists as the ground truth for the evaluation, there was no corpus of this nature available, and therefore author-filled checklists are used instead.

The NeurIPS'24 review checklist contains 16 questions, which can be divided into five categories.

```
For all authors...
(a) Do the main claims made in the
    abstract and introduction
    accurately reflect the paper's
    contributions and scope?
(b) Did you describe the limitations
    of your work?
(c) Did you discuss any potential
    negative societal impacts of your
    work?
(d) Have you read the NeurIPS Code of
    Ethics and ensured that your paper
    conforms to it?
(e) Do you have safeguards in place
    for responsible release of models
    or data with a high risk for
    misuse?
(f) Does the paper describe the usage
    of LLMs if they are an important,
    original, or non-standard
    component of the core methods?
```

Listing 1: Author block.

```
If you are including theoretical
    results...
(a) Did you state the full set of
    assumptions of all theoretical
    results and include complete
    proofs of all theoretical results?
```

Listing 2: Theoretical results block.

```
If you ran experiments...
(a) Did you include information on
    the computer resources required to
    reproduce the experiments?
(b) Did you include the code, data,
    and instructions needed to
    reproduce the main experimental
    results?
(c) Did you specify all the training
    details?
(d) Did you report error bars?
(e) Did you include the total amount
    of compute and the type of
    resources used?
```

Listing 3: Experimental block.

```
If you are using existing assets or
    releasing new assets...
(a) If your work uses existing
    assets, did you cite the creators
    and respect the license and terms
    of use?
(b) Did you document the new assets
    and provide these details
    alongside the assets?
```

Listing 4: Assets block.

The first category presents a series of questions aimed at all authors, shown in Listing 1.

Authors Block This first series of points addresses the content aspects of the paper, especially points (a) and (b). Regarding point (d), the answer to this question may not be inferred from the actual content of the paper, and therefore the LLM is expected to fail to answer it correctly. Point (e) is similarly dependent on information external to the paper, whilst point (f) may be answered correctly only if the paper explicitly addresses LLM usage in its methodology. Even though questions (d) and (e) cannot be directly inferred based on the content, they are still kept for the evaluation to detect potential biases within LLMs, since they are expected to answer "N/A" to these points.

Theoretical Results Block The second block refers to theoretical results, as depicted in Listing 2. In this case, the question can be answered based on the content of the paper. However, the answer may not be trivial since it requires the LLM to first discern whether the paper reports theoretical results and which these are, and subsequently assess whether assumptions are stated and proofs are provided. This question may therefore be a good reference point to assess the reasoning capacity of the studied LLMs.

Experimental Block The third block of questions, described in Listing 3, addresses the information required for the experimentation and evaluation process. Although the previous section required a deeper level of understanding, the questions in this section are more concise and targeted, thus requiring a lower level of understanding. Subsequently, LLMs are expected to perform well in answering these points.

Assets Block The fourth block refers to the work's assets, which are outlined in Listing 4. Similarly to the previous block, the answers to these questions should ideally be explicitly declared in the paper and therefore should not require a high level of understanding.

Human Subjects Block The fifth and final block concerns crowdsourcing and research with human subjects, as described in Listing 5. These questions all address aspects that are largely external to the paper content, and therefore the LLM is expected to struggle to answer them accurately from the paper alone, and thus expected to answer "N/A" to these

```
If you used crowdsourcing or
    conducted research with human
    subjects ...
(a) Did you include the full text of
    instructions given to participants
    and details about compensation?
(b) Did you describe potential
    participant risks and obtain IRB
    approval (or equivalent) if
    applicable?
```

Listing 5: Human subjects block.

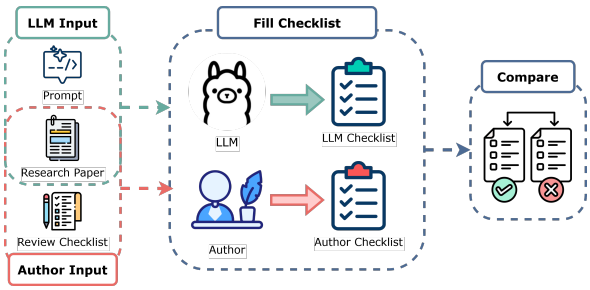


Figure 1: Overview of the conducted workflow.

points.

3.1 Methodology

Figure 1 outlines the workflow of the procedure performed. The research paper and the review checklist act as input. The checklist is then filled by both the author and the LLM, leading to two different versions of the same checklist. In the case of the author, this checklist is filled not only on the basis of the content of the paper but also on personal criteria and contextual knowledge, which the LLM and the reviewer lack. In the case of the LLM, the answers are purely based on the content of the paper, similar to how an external reviewer would do it. Therefore, in those questions where the answer cannot be inferred from the paper, the LLM is expected to answer "N/A", since it does not have enough information. If the LLM provides a different answer, it may evidence the existence of a bias within the model.

According to the conference guidelines, and as previously stated, authors must respond to each checklist point but are not required to justify their answers. Our experiments mirror this approach: the LLM is restricted to responding with "Yes," "No," or "N/A," without providing any explanation. However, in the presented tool, *Prêt-à-Submit*, the LLM is prompted to elaborate on its ratings in order to give researchers meaningful feedback on the checklist validation. These explanations are

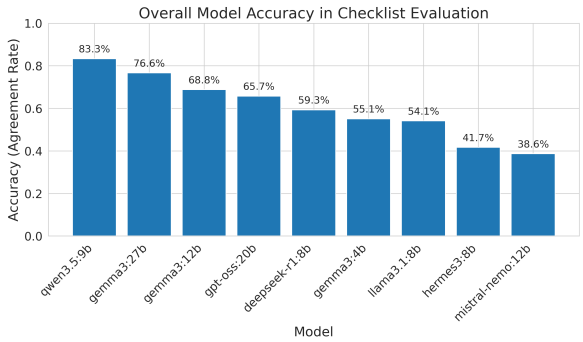


Figure 2: Coincidence w.r.t base checklist per model

returned in JSON format for further processing. The detailed prompt provided to the models is included in Appendix A. To ensure reproducibility and comparability, every model is tested with the same prompt, which contains a description of the task, the full checklist, and the article in Markdown format.

Once the LLM checklist is clean and processed, it can then be compared with the human-filled checklist to assess whether the results provided by the LLM are comparable and resemble human criteria.

3.2 Data Processing

As stated at the beginning of this Section, the corpus of accepted NeurIPS'24 papers is used as a benchmark for the experimentation. The scraping was first performed to retrieve both the metadata of each paper and the file of the article in PDF format. A total of 4,035 papers were first retrieved. The second step comprises the extraction of the author checklist from the PDF file. First, each PDF file is converted to plain text for further processing. Then, using a set of regular expressions, each checklist section along with the authors' response to each point is extracted and stored.

Once the authors' checklist (or reference checklist) is extracted, the LLM is queried to fill the checklist based on the content of the paper. In order to complete this task, each LLM is fed with its corresponding prompt containing the checklist and instructions on how to fill it, along with the content of the paper. Since the paper contains the authors' filled checklist, this content has to be truncated before querying the LLM. In preliminary trials, this content was not truncated from the article, and the results and the evidence provided directly pointed to the results provided by the authors. Therefore, the LLM was not responding to the checklist trying

to reason over the content of the article, but rather by replicating the answers provided by the authors.

After collecting the LLM responses, parsing and post-processing steps are required to extract the clean content from the raw outputs. All responses are enforced to be in JSON format by the *Ollama* API. However, malformations were detected in a subset of the JSON files contained in the responses, thus requiring additional post-processing to obtain clean results for further comparison. The most common sources of malformation include the inclusion of markdown backticks surrounding the JSON structure, as well as natural language preambles such as declarative statements preceding the data. Papers whose author-provided checklists did not conform to the expected format were excluded from the study, as were papers exceeding the context window of the evaluated models. After post-processing and filtering, of the 4,035 initial papers, 3,746 could be fully processed by all LLMs without errors.

3.3 Prêt-à-submit

This workflow is the backbone of Prêt-à-Submit, the open-source checklist assistant presented in this work. The tool provides researchers with a straightforward interface to upload their manuscript and select the target conference checklist, after which it runs the full pipeline locally through the user's own *Ollama* instance. For each checklist item, the model indicates the compliance status alongside a justification grounded in the paper's content. This design enables faster iteration cycles compared to manual self-assessment, and stands as a freely available alternative to privately owned tools that depend on closed-source models and external infrastructure.

4 Evaluation

Nine open-source models were selected for the study in order to analyze both the differences between architectures of similar size and the impact of scaling the number of parameters on the quality of the responses. The models used were:

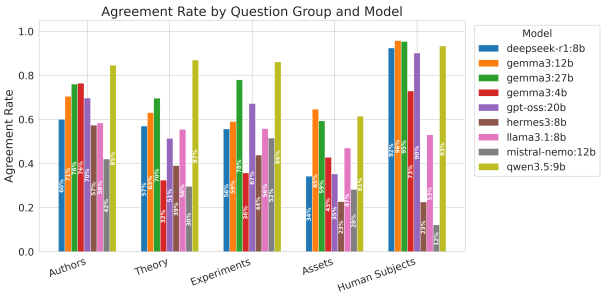
- deepseek-r1:8b (5.2 GB) (DeepSeek-AI et al., 2025)
- hermes3:8b (4.7 GB) (Teknium et al., 2024)
- mistral-nemo:12b (7.1 GB)
- gpt-oss:20b (12 GB) (OpenAI et al., 2025)
- llama3.1:8b (4.9 GB) (Grattafiori et al., 2024)
- gemma3:27b (17 GB) (Team et al., 2025)
- gemma3:12b (8.1 GB)
- gemma3:4b (3.3 GB)
- qwen3.5:9b (6.6 GB) (Yang et al., 2025)

The models were selected based on criteria aligned with the experimental constraints and objectives of this work. First, only open-source models compatible with *Ollama* were considered, ensuring that all experiments can be executed locally in a reproducible and privacy-preserving setting. Second, model size was constrained by the available hardware (40 GB of VRAM), leading to the selection of models in the 4B to 27B parameter range. This allows for a realistic evaluation of performance across different scales while remaining feasible on consumer or moderately provisioned systems. For all models, the *Ollama* Python package was used to query and retrieve responses. The complete code for scraping and preprocessing the data, together with the snippets for the execution and post-processing steps for each model, is available on GitHub.[1]. Since LLMs are trained on a vast set of documents, it is possible that there is some underlying data leakage that may cause biases in the responses. However, this is not expected to happen due to the nature of the selected LLMs, which are general-purpose ones.

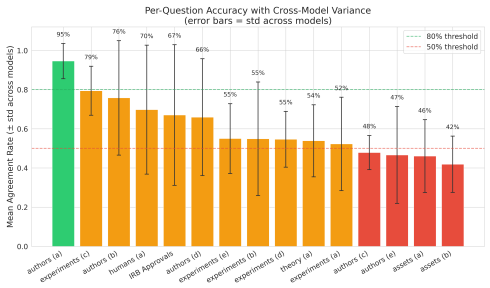
4.1 General Performance Comparison

First, the overall accuracy per model is measured. In this case, accuracy is measured as the coincidence between the response provided by the LLM ("Yes", "No", "N/A"), and the value reported by the authors. Figure 2 provides the results for each model. The results show a disparity in the behavior of each LLM with respect to the checklist. The model with the best performance was *qwen3.5:9b*, with an overall accuracy of 83.3%, followed by *gemma3:27b* at 76.6% and *gemma3:12b* at 68.8%. At the opposite end of the spectrum, *mistral-nemo:12b* barely reached 38.6% and *hermes3:8b* reached only 41.7%. This places all models above the randomness baseline of 33.33%, which would be expected from random selection among three options. It is particularly interesting that models of similar sizes show such marked differences in performance: *gemma3:12b* and *mistral-nemo:12b*, both with 12 billion parameters, differ by more than 30 percentage points in accuracy. This phenomenon evidences that the number of parameters, on its own, is not a reliable measure of

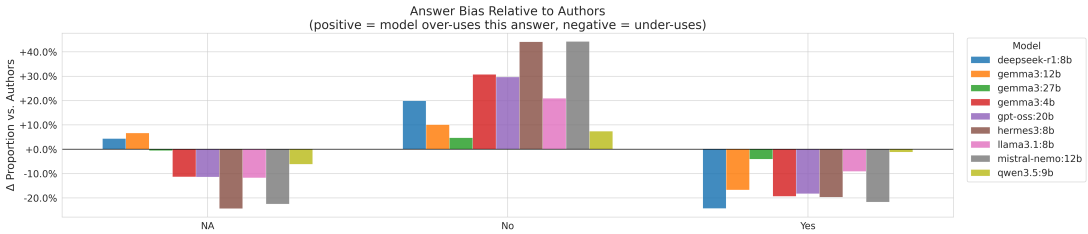
[1] https://anonymous.4open.science/r/pret-a-submit-experiment-4E4D



(a) Coincidence w.r.t base checklist for each block of questions per model.



(b) Coincidence w.r.t base checklist for each question with model variance.



(c) Answer bias per model w.r.t base checklist.

Figure 3: Comparison per LLM on the questions w.r.t the base checklist. Coincidence per block, per question and bias per model.

the quality of the responses. Similarly, *qwen3.5:9b* (83.3%) outperforms *gemma3:27b* (76.6%) with one-third as many parameters. These results suggest that the number of parameters is neither a linear nor a sufficient predictor of the model's ability to tackle this task.

4.2 Response Comparison

The results per question group (Figure 3a) partially confirm the initial hypotheses formulated in Section 3 with respect to the expected performance per question block. According to the formulated hypotheses, models are expected to perform better on the questions from the *authors*, *experimental* and *assets* blocks than on those of the rest of the checklist blocks, since their answers can be inferred straightforwardly from the content of the paper. In the *experiments* section of the checklist, where the questions were anticipated to be more straightforward due to their concise and targeted nature, most models actually achieved higher accuracy, supporting the hypothesis that questions with a lower abstraction level favor LLM performance. This is also the case for the *authors* block, which yielded good performances overall despite containing several questions that rely on knowledge external to the paper. However, several findings deviate from the expected model behavior in certain questions. The *assets* block was anticipated to be relatively accessible since answers should be explicitly stated in the paper and require little abstraction. Yet performance dropped notably across all models, suggesting that even factual, low-abstraction questions can be problematic when their answers depend on specific sections that LLMs may fail to locate or interpret correctly. Similarly, the human subjects block produced deceptively high coincidence values for most models, not necessarily as a sign of strong performance, but rather as a reflection of negative bias in the baseline, where the vast majority of papers respond *No* or *N/A*, rewarding systematic negative responses. The notable exception is *mistral-nemo:12b*, which achieves only 12% coincidence, suggesting that it did not default to negative responses and thus was penalized for attempting to infer answers from the content of the paper. Finally, the *theory* block exhibited the widest performance spread across models, reinforcing its role as a useful proxy for assessing reasoning capacity.

4.3 Question-by-question Analysis

Beyond overall performance, the question-by-question analysis, in which we evaluate each question based on its accuracy at both the model level and the overall level, reveals that the difficulty of the task varies considerably depending on the item being evaluated (Figure 3b). The questions with the

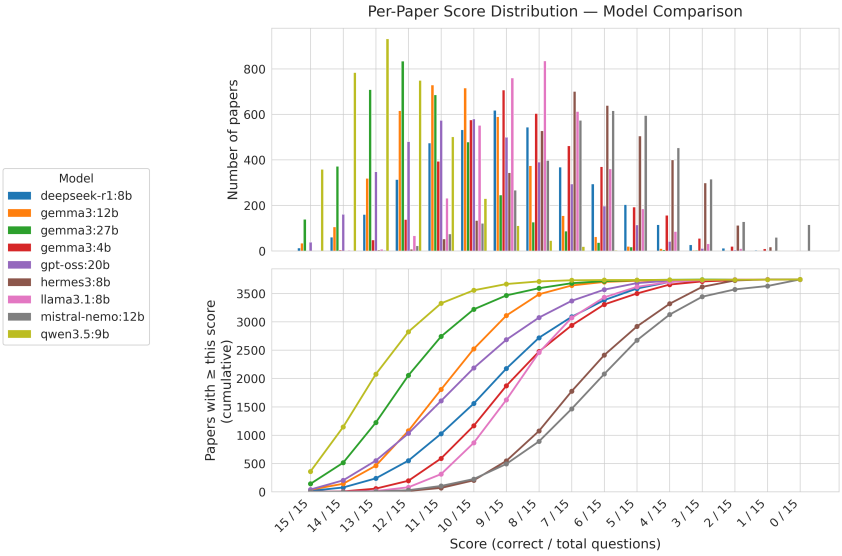


Figure 4: Coincidence per paper w.r.t base checklist and aggregated coincidence per paper.

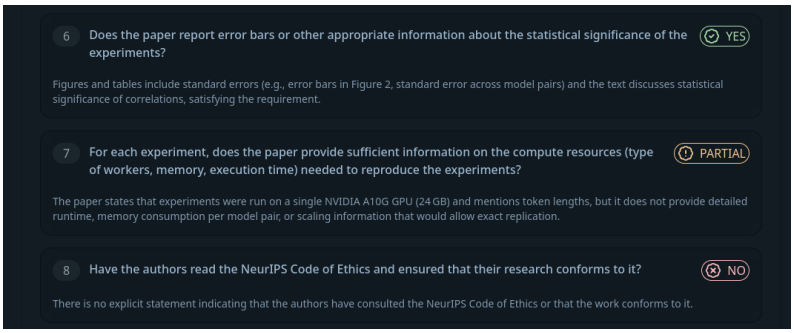


Figure 5: Screenshot of *Prêt-à-submit* Results page.

highest average agreement rate among models are authors(a), or *Claims* (95%); experiments(c), or *Experimental Setting/Details* (79%); and authors(b), or *Limitations* (76%). The first two of these questions also have small error bars, indicating that consensus among models is high and that these are items that LLMs solve consistently regardless of their architecture or training. This suggests that these questions assess aspects that are typically explicitly stated in the article's text. At the other end of the spectrum, questions such as assets(b), or *New Assets* (42%); assets(a), or *Licenses for existing assets* (46%); and authors(e), or *Safeguards* (46%), show both a lower average accuracy and greater variance across models, suggesting that these aspects are more difficult to evaluate automatically. However, it is worth noting that some of the questions with lower average accuracy have reduced variance, potentially indicating that these questions require a broader context than the article itself, such as the question regarding the social impact of the work (*Broader Impacts*). These types of questions require critical thinking or external knowledge about the work to which the model does not have access.

A trend observed in virtually all models is a higher-than-expected bias toward the *No* response and a widespread underestimation of the *N/A* category (Figure 3c). Compared to the actual distribution among authors, the models tend to respond *No* more frequently than expected, while the *N/A* category appears underrepresented in most of them. This bias toward *No* rather than *N/A* offers a plausible explanation for the disparity in question accuracy observed earlier (Figure 3b). Questions whose correct answer is mostly *N/A*, either because they require a context that goes beyond the content of the article itself or because their applicability is restricted to certain types of articles, as is the case with *Licenses for existing assets* (assets(a)) or *Crowdsourcing and Research with Human Subjects* (human subjects block), are particularly disadvantaged: since *N/A* is not considered with sufficient frequency, the model systematically substitutes it

with *No*. Consequently, the low accuracy achieved on these questions may not be indicative of an existing limitation in the model's reading comprehension, but rather of a structural tendency to collapse these two semantically distinct categories into a single one. The explanation for this behavior may lie in the nature of the models; since they are trained to provide definitive answers, identifying a lack of applicability is more complex than choosing a binary option.

4.4 Prêt-à-submit Assessment

These findings collectively inform the design of *Prêt-à-Submit*, the tool presented in this work[2]. Rather than treating the model's output as a fully reliable verdict, the system is conceived as a decision-support aid: a structured, first-pass assessment that flags potential compliance issues and guides the author's attention, while leaving final judgment to the researcher. Crucially, *Prêt-à-Submit* adopts a privacy-first, locally-run architecture. All processing is handled through the user's own Ollama instance, meaning no paper data leaves their infrastructure. This design choice makes it suitable for use with unpublished manuscripts, where confidentiality is paramount. A screenshot depicting the results page of the tool is shown in Figure 5. Further examples of the functioning of the tool are shown in Appendix B.

5 Conclusions and Future Work

This paper presents a first attempt to benchmark the behavior of LLMs for the author checklist completion task. These checklists are a useful tool not only for reviewers, but also for authors, since they enable self-assessment of the quality of their work and also their compliance with respect to the conference guidelines. LLMs can act as assistants to both authors and reviewers in this task, because of their human-like reasoning capacities. This paper provides a study of nine different LLMs for the checklist completion task in the context of the NeurIPS 2024 conference. Specifically, it prompts the checklist questions to the LLMs, and compares the responses provided by the LLMs with respect to human-filled checklists. The study shows that LLMs are still far from accurately emulating human behavior, since the answer to some of the checklist points relies on contextual aspects, and thus cannot be inferred directly from the paper. In addition, the results show that the *qwen3.5:9b* achieved the best performance from a general perspective, with *gemma3:27b* and *gpt-oss:20b* following closely.

Similarly, even lighter models such as *gemma-3:4b*, weighing in at just 3.3 GB, could run on computers with a consumer-grade dedicated graphics card and produce acceptable results. However, the analysis also reveals systematic biases and consistency issues across models, highlighting a lack of maturity in the current state of this technology. These results underscore the need for human oversight and a critical mindset when using these tools.

Future works include extending the study to cover the checklists of different conferences, such as AAAI. This would provide a greater insight on the actual capabilities of LLMs for this task. In the same vein, another potential line of research comprises the application of the developed framework to conference checklists of non-AI areas, in order to assess whether the performance varies. Finally, it would be relevant to study the relation of the prompting strategy, as well as the generation parameters, with respect to the existing bias towards negative responses.

Limitations

The main limitation of this work is the fact that the results of this study are very local and therefore the behavior patterns extracted in this study cannot be extrapolated to any conference. Although this idea could have been explored for different conference checklists, only NeurIPS ones were available.

In the same vein, it cannot be ascertained whether the same conclusions regarding the performance of the studied models would hold for papers of different research domains. Additionally, using the authors' checklist as a ground truth may introduce a possible bias on the responses, since authors may tend to answer favorably to the checklist points. Since there are no validated corpora for this task, these biases may not be verifiable.

Finally, regarding the behavior of the studied models, we observe a bias towards predicting *No* answers in *N/A* points. As a result, this may lead to mistakes in the feedback provided by the tool, encouraging changes in cases in which it may not apply.

[2] https://anonymous.4open.science/r/pret-a-submit-4B44/

References

AAAI. 2024. AAAI reproducibility checklist.

DeepSeek-AI, Daya Guo, Dejian Yang, Haowei Zhang, Junxiao Song, Peiyi Wang, Qihao Zhu, Runxin Xu, Ruoyu Zhang, Shirong Ma, Xiao Bi, Xiaokang Zhang, Xingkai Yu, Yu Wu, Z. F. Wu, Zhibin Gou, Zhihong Shao, Zhuoshu Li, Ziyi Gao, Aixin Liu, Bing Xue, Bingxuan Wang, Bochao Wu, Bei Feng, Chengda Lu, Chenggang Zhao, Chengqi Deng, Chenyu Zhang, Chong Ruan, Damai Dai, Deli Chen, Dongjie Ji, Erhang Li, Fangyun Lin, Fucong Dai, Fuli Luo, Guangbo Hao, Guanting Chen, Guowei Li, H. Zhang, Han Bao, Hanwei Xu, Haocheng Wang, Honghui Ding, Huajian Xin, Huazuo Gao, Hui Qu, Hui Li, Jianzhong Guo, Jiashi Li, Jiawei Wang, Jingchang Chen, Jingyang Yuan, Junjie Qiu, Junlong Li, J. L. Cai, Jiaqi Ni, Jian Liang, Jin Chen, Kai Dong, Kai Hu, Kaige Gao, Kang Guan, Kexin Huang, Kuai Yu, Lean Wang, Lecong Zhang, Liang Zhao, Litong Wang, Liyue Zhang, Lei Xu, Leyi Xia, Mingchuan Zhang, Minghua Zhang, Minghui Tang, Meng Li, Miaojun Wang, Mingming Li, Ning Tian, Panpan Huang, Peng Zhang, Qiancheng Wang, Qinyu Chen, Qiushi Du, Ruiqi Ge, Ruisong Zhang, Ruizhe Pan, Runji Wang, R. J. Chen, R. L. Jin, Ruyi Chen, Shanghao Lu, Shangyan Zhou, Shanhuang Chen, Shengfeng Ye, Shiyu Wang, Shuiping Yu, Shunfeng Zhou, Shuting Pan, S. S. Li, Shuang Zhou, Shaoqing Wu, Shengfeng Ye, Tao Yun, Tian Pei, Tianyu Sun, T. Wang, Wangding Zeng, Wanjia Zhao, Wen Liu, Wenfeng Liang, Wenjun Gao, Wenqin Yu, Wentao Zhang, W. L. Xiao, Wei An, Xiaodong Liu, Xiaohan Wang, Xiaokang Chen, Xiaotao Nie, Xin Cheng, Xin Liu, Xin Xie, Xingchao Liu, Xinyu Yang, Xinyuan Li, Xuecheng Su, Xuheng Lin, X. Q. Li, Xiangyue Jin, Xiaojin Shen, Xiaosha Chen, Xiaowen Sun, Xiaoxiang Wang, Xinnan Song, Xinyi Zhou, Xianzu Wang, Xinxia Shan, Y. K. Li, Y. Q. Wang, Y. X. Wei, Yang Zhang, Yanhong Xu, Yao Li, Yao Zhao, Yaofeng Sun, Yaohui Wang, Yi Yu, Yichao Zhang, Yifan Shi, Yiliang Xiong, Ying He, Yishi Piao, Yisong Wang, Yixuan Tan, Yiyang Ma, Yiyuan Liu, Yongqiang Guo, Yuan Ou, Yuduan Wang, Yue Gong, Yuheng Zou, Yujia He, Yunfan Xiong, Yuxiang Luo, Yuxiang You, Yuxuan Liu, Yuyang Zhou, Y. X. Zhu, Yanhong Xu, Yanping Huang, Yaohui Li, Yi Zheng, Yuchen Zhu, Yunxian Ma, Ying Tang, Yukun Zha, Yuting Yan, Z. Z. Ren, Zehui Ren, Zhangli Sha, Zhe Fu, Zhean Xu, Zhenda Xie, Zhengyan Zhang, Zhewen Hao, Zhicheng Ma, Zhigang Yan, Zhiyu Wu, Zihui Gu, Zijia Zhu, Zijun Liu, Zilin Li, Ziwei Xie, Ziyang Song, Zizheng Pan, Zhen Huang, Zhipeng Xu, Zhongyu Zhang, and Zhen Zhang. 2025. DeepSeek-R1: Incentivizing Reasoning Capability in LLMs via Reinforcement Learning. *Nature*, 645(8081):633–638.

Marc Deisenroth, Finale Doshi-Velez, Nika Haghtalab, David Rolnick, Jenna Wiens, Rajesh Jayaram, Vincent Cohen-Addad, Drew Tyler, David Woodruff, Jinsung Yoon, Mihir Parmar, Palash Goyal, Corinna Cortes, Vahab Mirrokni, Tomas Pfister, Burak Gokturk, and Jean Kossaifi. 2026. NeurIPS Supports Authors with Google's Paper Assistant Tool (PAT) – NeurIPS Blog.

Jesse Dodge, Suchin Gururangan, Dallas Card, Roy Schwartz, and Noah A. Smith. 2019. Show Your Work: Improved Reporting of Experimental Results. ArXiv:1909.03004 [cs].

Jesse Dodge and Noah A. Smith. 2020. Guest post: Reproducibility at emnlp 2020.

Julia Evans, Jennifer D'Souza, and Sören Auer. 2024. Large language models as evaluators for scientific synthesis. In *Proceedings of the 20th Conference on Natural Language Processing (KONVENS 2024)*, pages 1–22, Vienna, Austria. Association for Computational Linguistics.

Michael Galarnyk, Rutwik Routu, Vidhyakshaya Kannan, Kosha Bheda, Prasun Banerjee, Agam Shah, and Sudheer Chava. 2025. ConfReady: A RAG based Assistant and Dataset for Conference Checklist Responses.

Alexander Goldberg, Ihsan Ullah, Thanh Gia Hieu Khuong, Benedictus Kent Rachmat, Zhen Xu, Isabelle Guyon, and Nihar B. Shah. 2024. Usefulness of LLMs as an Author Checklist Assistant for Scientific Papers: NeurIPS'24 Experiment. ArXiv:2411.03417 [cs].

Aaron Grattafiori, Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Alex Vaughan, Amy Yang, Angela Fan, Anirudh Goyal, Anthony Hartshorn, Aobo Yang, Archi Mitra, Archie Sravankumar, Artem Korenev, Arthur Hinsvark, Arun Rao, Aston Zhang, Aurelien Rodriguez, Austen Gregerson, Ava Spataru, Baptiste Roziere, Bethany Biron, Binh Tang, Bobbie Chern, Charlotte Caucheteux, Chaya Nayak, Chloe Bi, Chris Marra, Chris McConnell, Christian Keller, Christophe Touret, Chunyang Wu, Corinne Wong, Cristian Canton Ferrer, Cyrus Nikolaidis, Damien Allonsius, Daniel Song, Danielle Pintz, Danny Livshits, Danny Wyatt, David Esiobu, Dhruv Choudhary, Dhruv Mahajan, Diego Garcia-Olano, Diego Perino, Dieuwke Hupkes, Egor Lakomkin, Ehab AlBadawy, Elina Lobanova, Emily Dinan, Eric Michael Smith, Filip Radenovic, Francisco Guzmán, Frank Zhang, Gabriel Synnaeve, Gabrielle Lee, Georgia Lewis Anderson, Govind Thattai, Graeme Nail, Gregoire Mialon, Guan Pang, Guillem Cucurell, Hailey Nguyen, Hannah Korevaar, Hu Xu, Hugo Touvron, Iliyan Zarov, Imanol Arrieta Ibarra, Isabel Kloumann, Ishan Misra, Ivan Evtimov, Jack Zhang, Jade Copet, Jaewon Lee, Jan Geffert, Jana Vranes, Jason Park, Jay Mahadeokar, Jeet Shah, Jelmer van der Linde, Jennifer Billock, Jenny Hong, Jenya Lee, Jeremy Fu, Jianfeng Chi, Jianyu Huang, Jiawen Liu, Jie Wang, Jiecao Yu, Joanna Bitton, Joe Spisak, Jongsoo Park, Joseph Rocca, Joshua Johnstun, Joshua Saxe, Junteng Jia, Kalyan Vasuden Alwala, Karthik Prasad, Kartikeya Upasani, Kate Plawiak, Ke Li, Kenneth Heafield, Kevin Stone, Khalid El-Arini, Krithika Iyer,

Kshitiz Malik, Kuenley Chiu, Kunal Bhalla, Kushal Lakhotia, Lauren Rantala-Yeary, Laurens van der Maaten, Lawrence Chen, Liang Tan, Liz Jenkins, Louis Martin, Lovish Madaan, Lubo Malo, Lukas Blecher, Lukas Landzaat, Luke de Oliveira, Madeline Muzzi, Mahesh Pasupuleti, Mannat Singh, Manohar Paluri, Marcin Kardas, Maria Tsimpoukelli, Mathew Oldham, Mathieu Rita, Maya Pavlova, Melanie Kambadur, Mike Lewis, Min Si, Mitesh Kumar Singh, Mona Hassan, Naman Goyal, Narjes Torabi, Nikolay Bashlykov, Nikolay Bogoychev, Niladri Chatterji, Ning Zhang, Olivier Duchenne, Onur Çelebi, Patrick Alrassy, Pengchuan Zhang, Pengwei Li, Petar Vasic, Peter Weng, Prajjwal Bhargava, Pratik Dubal, Praveen Krishnan, Punit Singh Koura, Puxin Xu, Qing He, Qingxiao Dong, Ragavan Srinivasan, Raj Ganapathy, Ramon Calderer, Ricardo Silveira Cabral, Robert Stojnic, Roberta Raileanu, Rohan Maheswari, Rohit Girdhar, Rohit Patel, Romain Sauvestre, Ronnie Polidoro, Roshan Sumbaly, Ross Taylor, Ruan Silva, Rui Hou, Rui Wang, Saghar Hosseini, Sahana Chennabasappa, Sanjay Singh, Sean Bell, Seohyun Sonia Kim, Sergey Edunov, Shaoliang Nie, Sharan Narang, Sharath Raparthy, Sheng Shen, Shengye Wan, Shruti Bhosale, Shun Zhang, Simon Vandenhende, Soumya Batra, Spencer Whitman, Sten Sootla, Stephane Collot, Suchin Gururangan, Sydney Borodinsky, Tamar Herman, Tara Fowler, Tarek Sheasha, Thomas Georgiou, Thomas Scialom, Tobias Speckbacher, Todor Mihaylov, Tong Xiao, Ujjwal Karn, Vedanuj Goswami, Vibhor Gupta, Vignesh Ramanathan, Viktor Kerkez, Vincent Gonguet, Virginie Do, Vish Vogeti, Vítor Albiero, Vladan Petrovic, Weiwei Chu, Wenhan Xiong, Wenyin Fu, Whitney Meers, Xavier Martinet, Xiaodong Wang, Xiaofang Wang, Xiaoqing Ellen Tan, Xide Xia, Xinfeng Xie, Xuchao Jia, Xuewei Wang, Yaelle Goldschlag, Yashesh Gaur, Yasmine Babaei, Yi Wen, Yiwen Song, Yuchen Zhang, Yue Li, Yuning Mao, Zacharie Delpierre Coudert, Zheng Yan, Zhengxing Chen, Zoe Papakipos, Aaditya Singh, Aayushi Srivastava, Abha Jain, Adam Kelsey, Adam Shajnfeld, Adithya Gangidi, Adolfo Victoria, Ahuva Goldstand, Ajay Menon, Ajay Sharma, Alex Boesenberg, Alexei Baevski, Allie Feinstein, Amanda Kallet, Amit Sangani, Amos Teo, Anam Yunus, Andrei Lupu, Andres Alvarado, Andrew Caples, Andrew Gu, Andrew Ho, Andrew Poulton, Andrew Ryan, Ankit Ramchandani, Annie Dong, Annie Franco, Anuj Goyal, Aparajita Saraf, Arkabandhu Chowdhury, Ashley Gabriel, Ashwin Bharambe, Assaf Eisenman, Azadeh Yazdan, Beau James, Ben Maurer, Benjamin Leonhardi, Bernie Huang, Beth Loyd, Beto De Paola, Bhargavi Paranjape, Bing Liu, Bo Wu, Boyu Ni, Braden Hancock, Bram Wasti, Brandon Spence, Brani Stojkovic, Brian Gamido, Britt Montalvo, Carl Parker, Carly Burton, Catalina Mejia, Ce Liu, Changhan Wang, Changkyu Kim, Chao Zhou, Chester Hu, Ching-Hsiang Chu, Chris Cai, Chris Tindal, Christoph Feichtenhofer, Cynthia Gao, Damon Civin, Dana Beaty, Daniel Kreymer, Daniel Li, David Adkins, David Xu, Davide Testuggine, Delia David, Devi Parikh, Diana Liskovich, Didem Foss, Dingkang Wang, Duc Le, Dustin Holland, Edward Dowling, Eissa Jamil, Elaine Montgomery, Eleonora Presani, Emily Hahn, Emily Wood, Eric-Tuan Le, Erik Brinkman, Esteban Arcaute, Evan Dunbar, Evan Smothers, Fei Sun, Felix Kreuk, Feng Tian, Filippos Kokkinos, Firat Ozgenel, Francesco Caggioni, Frank Kanayet, Frank Seide, Gabriela Medina Florez, Gabriella Schwarz, Gada Badeer, Georgia Swee, Gil Halpern, Grant Herman, Grigory Sizov, Guangyi, Zhang, Guna Lakshminarayanan, Hakan Inan, Hamid Shojanazeri, Han Zou, Hannah Wang, Hanwen Zha, Haroun Habeeb, Harrison Rudolph, Helen Suk, Henry Aspegren, Hunter Goldman, Hongyuan Zhan, Ibrahim Damlaj, Igor Molybog, Igor Tufanov, Ilias Leontiadis, Irina-Elena Veliche, Itai Gat, Jake Weissman, James Geboski, James Kohli, Janice Lam, Japhet Asher, Jean-Baptiste Gaya, Jeff Marcus, Jeff Tang, Jennifer Chan, Jenny Zhen, Jeremy Reizenstein, Jeremy Teboul, Jessica Zhong, Jian Jin, Jingyi Yang, Joe Cummings, Jon Carvill, Jon Shepard, Jonathan McPhie, Jonathan Torres, Josh Ginsburg, Junjie Wang, Kai Wu, Kam Hou U, Karan Saxena, Kartikay Khandelwal, Katayoun Zand, Kathy Matosich, Kaushik Veeraraghavan, Kelly Michelena, Keqian Li, Kiran Jagadeesh, Kun Huang, Kunal Chawla, Kyle Huang, Lailin Chen, Lakshya Garg, Lavender A, Leandro Silva, Lee Bell, Lei Zhang, Liangpeng Guo, Licheng Yu, Liron Moshkovich, Luca Wehrstedt, Madian Khabsa, Manav Avalani, Manish Bhatt, Martynas Mankus, Matan Hasson, Matthew Lennie, Matthias Reso, Maxim Groshev, Maxim Naumov, Maya Lathi, Meghan Keneally, Miao Liu, Michael L. Seltzer, Michal Valko, Michelle Restrepo, Mihir Patel, Mik Vyatskov, Mikayel Samvelyan, Mike Clark, Mike Macey, Mike Wang, Miquel Jubert Hermoso, Mo Metanat, Mohammad Rastegari, Munish Bansal, Nandhini Santhanam, Natascha Parks, Natasha White, Navyata Bawa, Nayan Singhal, Nick Egebo, Nicolas Usunier, Nikhil Mehta, Nikolay Pavlovich Laptev, Ning Dong, Norman Cheng, Oleg Chernoguz, Olivia Hart, Omkar Salpekar, Ozlem Kalinli, Parkin Kent, Parth Parekh, Paul Saab, Pavan Balaji, Pedro Rittner, Philip Bontrager, Pierre Roux, Piotr Dollar, Polina Zvyagina, Prashant Ratanchandani, Pritish Yuvraj, Qian Liang, Rachad Alao, Rachel Rodriguez, Rafi Ayub, Raghotham Murthy, Raghu Nayani, Rahul Mitra, Rangaprabhu Parthasarathy, Raymond Li, Rebekkah Hogan, Robin Battey, Rocky Wang, Russ Howes, Ruty Rinott, Sachin Mehta, Sachin Siby, Sai Jayesh Bondu, Samyak Datta, Sara Chugh, Sara Hunt, Sargun Dhillon, Sasha Sidorov, Satadru Pan, Saurabh Mahajan, Saurabh Verma, Seiji Yamamoto, Sharadh Ramaswamy, Shaun Lindsay, Shaun Lindsay, Sheng Feng, Shenghao Lin, Shengxin Cindy Zha, Shishir Patil, Shiva Shankar, Shuqiang Zhang, Shuqiang Zhang, Sinong Wang, Sneha Agarwal, Soji Sajuyigbe, Soumith Chintala, Stephanie Max, Stephen Chen, Steve Kehoe, Steve Satterfield, Sudarshan Govindaprasad, Sumit Gupta, Summer Deng, Sungmin Cho, Sunny Virk, Suraj Subramanian, Sy Choudhury, Sydney Goldman, Tal Remez, Tamar Glaser, Tamara Best, Thilo Koehler, Thomas Robinson, Tianhe Li, Tianjun Zhang, Tim Matthews, Timothy Chou, Tzook Shaked, Varun Vontimitta, Victoria Ajayi, Victoria Montanez, Vijai

Mohan, Vinay Satish Kumar, Vishal Mangla, Vlad Ionescu, Vlad Poenaru, Vlad Tiberiu Mihailescu, Vladimir Ivanov, Wei Li, Wenchen Wang, Wenwen Jiang, Wes Bouaziz, Will Constable, Xiaocheng Tang, Xiaojian Wu, Xiaolan Wang, Xilun Wu, Xinbo Gao, Yaniv Kleinman, Yanjun Chen, Ye Hu, Ye Jia, Ye Qi, Yenda Li, Yilin Zhang, Ying Zhang, Yossi Adi, Youngjin Nam, Yu, Wang, Yu Zhao, Yuchen Hao, Yundi Qian, Yunlu Li, Yuzi He, Zach Rait, Zachary DeVito, Zef Rosnbrick, Zhaoduo Wen, Zhenyu Yang, Zhiwei Zhao, and Zhiyu Ma. 2024. The llama 3 herd of models.

Max Heckel, Keith Hermann, Erica Price, and Ryan Montalvo. 2023. Scisummary: Use ai to summarize scientific articles.

Yixing Jiang and Andrew Ng. 2026. Stanford Agentic Reviewer. https://paperreview.ai/.

Diana Kelly. 2023. Peer review: Problematic or promising? *Econ. Labour Relat. Rev.*, 34(2):193–198.

J Kelly, T Sadeghieh, and K Adeli. 2014. Peer review in scientific publications: Benefits, critiques, & a survival guide. *EJIFCC*, 25(3):227–243.

Weixin Liang, Yuhui Zhang, Hancheng Cao, Binglu Wang, Daisy Ding, Xinyu Yang, Kailas Vodrahalli, Siyu He, Daniel Smith, Yian Yin, Daniel McFarland, and James Zou. 2023. Can large language models provide useful feedback on research papers? A large-scale empirical analysis. ArXiv:2310.01783 [cs].

Zhehui Liao, Maria Antoniak, Inyoung Cheong, Evie Yu-Yen Cheng, Ai-Heng Lee, Kyle Lo, Joseph Chee Chang, and Amy X. Zhang. 2024. Llms as research tools: A large scale survey of researchers' usage and perceptions.

Ryan Liu and Nihar B. Shah. 2023. ReviewerGPT? An Exploratory Study on Using Large Language Models for Paper Reviewing. ArXiv:2306.00622 [cs].

Ian Magnusson, Noah A. Smith, and Jesse Dodge. 2023. Reproducibility in NLP: What Have We Learned from the Checklist? In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 12789–12811, Toronto, Canada. Association for Computational Linguistics.

Nature. 2024. Artificial intelligence (ai) - editorial policies.

NeurIPS. 2025. Neurips reproducibility checklist.

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

Preston Tuggle, Nick Turley, Jerry Tworek, Juan Felipe Cerón Uribe, Andrea Vallone, Arun Vijayvergiya, Chelsea Voss, Carroll Wainwright, Justin Jay Wang, Alvin Wang, Ben Wang, Jonathan Ward, Jason Wei, CJ Weinmann, Akila Welihinda, Peter Welinder, Jiayi Weng, Lilian Weng, Matt Wiethoff, Dave Willner, Clemens Winter, Samuel Wolrich, Hannah Wong, Lauren Workman, Sherwin Wu, Jeff Wu, Michael Wu, Kai Xiao, Tao Xu, Sarah Yoo, Kevin Yu, Qiming Yuan, Wojciech Zaremba, Rowan Zellers, Chong Zhang, Marvin Zhang, Shengjia Zhao, Tianhao Zheng, Juntang Zhuang, William Zhuk, and Barret Zoph. 2024. Gpt-4 technical report.

OpenAI, Sandhini Agarwal, Lama Ahmad, Jason Ai, Sam Altman, Andy Applebaum, Edwin Arbus, Rahul K. Arora, Yu Bai, Bowen Baker, Haiming Bao, Boaz Barak, Ally Bennett, Tyler Bertao, Nivedita Brett, Eugene Brevdo, Greg Brockman, Sebastien Bubeck, Che Chang, Kai Chen, Mark Chen, Enoch Cheung, Aidan Clark, Dan Cook, Marat Dukhan, Casey Dvorak, Kevin Fives, Vlad Fomenko, Timur Garipov, Kristian Georgiev, Mia Glaese, Tarun Gogineni, Adam Goucher, Lukas Gross, Katia Gil Guzman, John Hallman, Jackie Hehir, Johannes Heidecke, Alec Helyar, Haitang Hu, Romain Huet, Jacob Huh, Saachi Jain, Zach Johnson, Chris Koch, Irina Kofman, Dominik Kundel, Jason Kwon, Volodymyr Kyrylov, Elaine Ya Le, Guillaume Leclerc, James Park Lennon, Scott Lessans, Mario Lezcano-Casado, Yuanzhi Li, Zhuohan Li, Ji Lin, Jordan Liss, Lily, Liu, Jiancheng Liu, Kevin Lu, Chris Lu, Zoran Martinovic, Lindsay McCallum, Josh McGrath, Scott McKinney, Aidan McLaughlin, Song Mei, Steve Mostovoy, Tong Mu, Gideon Myles, Alexander Neitz, Alex Nichol, Jakub Pachocki, Alex Paino, Dana Palmie, Ashley Pantuliano, Giambattista Parascandolo, Jongsoo Park, Leher Pathak, Carolina Paz, Ludovic Peran, Dmitry Pimenov, Michelle Pokrass, Elizabeth Proehl, Huida Qiu, Gaby Raila, Filippo Raso, Hongyu Ren, Kimmy Richardson, David Robinson, Bob Rotsted, Hadi Salman, Suvansh Sanjeev, Max Schwarzer, D. Sculley, Harshit Sikchi, Kendal Simon, Karan Singhal, Yang Song, Dane Stuckey, Zhiqing Sun, Philippe Tillet, Sam Toizer, Foivos Tsimpourlas, Nikhil Vyas, Eric Wallace, Xin Wang, Miles Wang, Olivia Watkins, Kevin Weil, Amy Wendling, Kevin Whinnery, Cedric Whitney, Hannah Wong, Lin Yang, Yu Yang, Michihiro Yasunaga, Kristen Ying, Wojciech Zaremba, Wenting Zhan, Cyril Zhang, Brian Zhang, Eddie Zhang, and Shengjia Zhao. 2025. Gpt-oss-120b & gpt-oss-20b Model Card.

Matthew J Page, Joanne E McKenzie, Patrick M Bossuyt, Isabelle Boutron, Tammy C Hoffmann, Cynthia D Mulrow, Larissa Shamseer, Jennifer M Tetzlaff, Elie A Akl, Sue E Brennan, Roger Chou, Julie Glanville, Jeremy M Grimshaw, Asbjørn Hróbjartsson, Manoj M Lalu, Tianjing Li, Elizabeth W Loder, Evan Mayo-Wilson, Steve McDonald, Luke A McGuinness, Lesley A Stewart, James Thomas, Andrea C Tricco, Vivian A Welch, Penny Whiting, and David Moher. 2021. The PRISMA 2020 statement: an updated guideline for reporting systematic reviews. *BMJ*, 372. Publisher: BMJ Publishing Group Ltd _eprint: https://www.bmj.com/content/372/bmj.n71.full.pdf.

Paperguide. 2024. Paperguide: The ai research assistant.

Aravind Srinivas, Johnny Ho, Andy Konwinsky, and Denis Yarats. 2022. Perplexity ai.

Gemma Team, Aishwarya Kamath, Johan Ferret, Shreya Pathak, Nino Vieillard, Ramona Merhej, Sarah Perrin, Tatiana Matejovicova, Alexandre Ramé, Morgane Rivière, Louis Rouillard, Thomas Mesnard, Geoffrey Cideron, Jean-bastien Grill, Sabela Ramos, Edouard Yvinec, Michelle Casbon, Etienne Pot, Ivo Penchev, Gaël Liu, Francesco Visin, Kathleen Kenealy, Lucas Beyer, Xiaohai Zhai, Anton Tsitsulin, Robert Busa-Fekete, Alex Feng, Noveen Sachdeva, Benjamin Coleman, Yi Gao, Basil Mustafa, Iain Barr, Emilio Parisotto, David Tian, Matan Eyal, Colin Cherry, Jan-Thorsten Peter, Danila Sinopalnikov, Surya Bhupatiraju, Rishabh Agarwal, Mehran Kazemi, Dan Malkin, Ravin Kumar, David Vilar, Idan Brusilovsky, Jiaming Luo, Andreas Steiner, Abe Friesen, Abhanshu Sharma, Abheesht Sharma, Adi Mayrav Gilady, Adrian Goedeckemeyer, Alaa Saade, Alex Feng, Alexander Kolesnikov, Alexei Bendebury, Alvin Abdagic, Amit Vadi, András György, André Susano Pinto, Anil Das, Ankur Bapna, Antoine Miech, Antoine Yang, Antonia Paterson, Ashish Shenoy, Ayan Chakrabarti, Bilal Piot, Bo Wu, Bobak Shahriari, Bryce Petrini, Charlie Chen, Charline Le Lan, Christopher A. Choquette-Choo, C. J. Carey, Cormac Brick, Daniel Deutsch, Danielle Eisenbud, Dee Cattle, Derek Cheng, Dimitris Paparas, Divyashree Shivakumar Sreepathihalli, Doug Reid, Dustin Tran, Dustin Zelle, Eric Noland, Erwin Huizenga, Eugene Kharitonov, Frederick Liu, Gagik Amirkhanyan, Glenn Cameron, Hadi Hashemi, Hanna Klimczak-Plucińska, Harman Singh, Harsh Mehta, Harshal Tushar Lehri, Hussein Hazimeh, Ian Ballantyne, Idan Szpektor, Ivan Nardini, Jean Pouget-Abadie, Jetha Chan, Joe Stanton, John Wieting, Jonathan Lai, Jordi Orbay, Joseph Fernandez, Josh Newlan, Ju-yeong Ji, Jyotinder Singh, Kat Black, Kathy Yu, Kevin Hui, Kiran Vodrahalli, Klaus Greff, Linhai Qiu, Marcella Valentine, Marina Coelho, Marvin Ritter, Matt Hoffman, Matthew Watson, Mayank Chaturvedi, Michael Moynihan, Min Ma, Nabila Babar, Natasha Noy, Nathan Byrd, Nick Roy, Nikola Momchev, Nilay Chauhan, Noveen Sachdeva, Oskar Bunyan, Pankil Botarda, Paul Caron, Paul Kishan Rubenstein, Phil Culliton, Philipp Schmid, Pier Giuseppe Sessa, Pingmei Xu, Piotr Stanczyk, Pouya Tafti, Rakesh Shivanna, Renjie Wu, Renke Pan, Reza Rokni, Rob Willoughby, Rohith Vallu, Ryan Mullins, Sammy Jerome, Sara Smoot, Sertan Girgin, Shariq Iqbal, Shashir Reddy, Shruti Sheth, Siim Põder, Sijal Bhatnagar, Sindhu Raghuram Panyam, Sivan Eiger, Susan Zhang, Tianqi Liu, Trevor Yacovone, Tyler Liechty, Uday Kalra, Utku Evci, Vedant Misra, Vincent Roseberry, Vlad Feinberg, Vlad Kolesnikov, Woohyun

Han, Woosuk Kwon, Xi Chen, Yinlam Chow, Yuvein Zhu, Zichuan Wei, Zoltan Egyed, Victor Cotruta, Minh Giang, Phoebe Kirk, Anand Rao, Kat Black, Nabila Babar, Jessica Lo, Erica Moreira, Luiz Gustavo Martins, Omar Sanseviero, Lucas Gonzalez, Zach Gleicher, Tris Warkentin, Vahab Mirrokni, Evan Senter, Eli Collins, Joelle Barral, Zoubin Ghahramani, Raia Hadsell, Yossi Matias, D. Sculley, Slav Petrov, Noah Fiedel, Noam Shazeer, Oriol Vinyals, Jeff Dean, Demis Hassabis, Koray Kavukcuoglu, Clement Farabet, Elena Buchatskaya, Jean-Baptiste Alayrac, Rohan Anil, Dmitry, Lepikhin, Sebastian Borgeaud, Olivier Bachem, Armand Joulin, Alek Andreev, Cassidy Hardin, Robert Dadashi, and Léonard Hussenot. 2025. Gemma 3 Technical Report.

Ryan Teknium, Jeffrey Quesnelle, and Chen Guang. 2024. Hermes 3 Technical Report.

An Yang, Anfeng Li, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Gao, Chengen Huang, Chenxu Lv, Chujie Zheng, Dayiheng Liu, Fan Zhou, Fei Huang, Feng Hu, Hao Ge, Haoran Wei, Huan Lin, Jialong Tang, Jian Yang, Jianhong Tu, Jianwei Zhang, Jianxin Yang, Jiaxi Yang, Jing Zhou, Jingren Zhou, Junyang Lin, Kai Dang, Keqin Bao, Kexin Yang, Le Yu, Lianghao Deng, Mei Li, Mingfeng Xue, Mingze Li, Pei Zhang, Peng Wang, Qin Zhu, Rui Men, Ruize Gao, Shixuan Liu, Shuang Luo, Tianhao Li, Tianyi Tang, Wenbiao Yin, Xingzhang Ren, Xinyu Wang, Xinyu Zhang, Xuancheng Ren, Yang Fan, Yang Su, Yichang Zhang, Yinger Zhang, Yu Wan, Yuqiong Liu, Zekun Wang, Zeyu Cui, Zhenru Zhang, Zhipeng Zhou, and Zihan Qiu. 2025. Qwen3 Technical Report.

A LLM Prompt

This Appendix presents the prompt used for all model inferences. The paper is inserted in markdown format at the `{paper_content}` tag and the JSON template is inserted at the `{str(ChecklistResponses.model_json_schema())}` tag.

> You are a paper rater for the conference NeurIPS 2024, you have been given the following questions that you must answer in order to evaluate the paper: Evaluate the paper and answer these questions:
>
> 1. Claims - Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope?
>
> 2. Limitations - Does the paper discuss the limitations of the work performed by the authors?
>
> 3. Theory Assumptions and Proofs - For each theoretical result, does the paper provide the full set of assumptions and a complete (and correct) proof?
>
> 4. Experimental Result Reproducibility - 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)?
>
> 5. Open access to data and code - 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?
>
> 6. Experimental Setting/Details - Does the paper specify all the training and test details (e.g., data splits, hyperparameters, how they were chosen, type of optimizer, etc.) necessary to understand the results?
>
> 7. Experiment Statistical Significance - Does the paper report error bars suitably and correctly defined or other appropriate information about the statistical significance of the experiments?
>
> 8. Experiments Compute Resources - 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?
>
> 9. Code Of Ethics - Does the research conducted in the paper conform, in every respect, with the NeurIPS Code of Ethics https://neurips.cc/public/EthicsGuidelines ?
>
> 10. Broader Impacts - Does the paper discuss both potential positive societal impacts and negative societal impacts of the work performed?
>
> 11. Safeguards - 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)?
>
> 12. Licenses for existing assets - 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?
>
> 13. New Assets - Are new assets introduced in the paper well documented and is the documentation provided alongside the assets?
>
> 14. Crowdsourcing and Research with Human Subjects - 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)?
>
> 15. Institutional Review Board (IRB) Approvals - 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?
>
> Following these guidelines, evaluate each question for the following paper,

provide your answer for each question. For each question, answer with ONLY one of: Yes, No, or NA for Not Applicable. You must answer in json format, following the provided template: {str(ChecklistResponses.model_json_schema())}. Make sure your output is a valid json and do not use any special characters that may break the json.

Here is the paper you must rate in markdown format: {paper_content}

B Prêt-à-Submit Screenshots

This appendix illustrates how the application operates through screenshots.

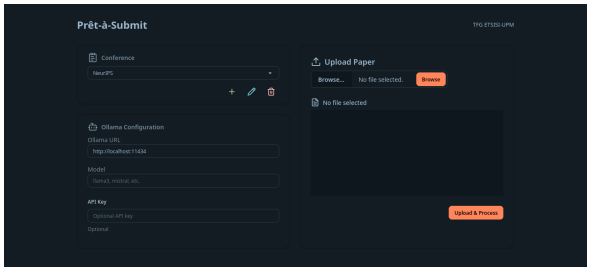


Figure 6: Application main screen

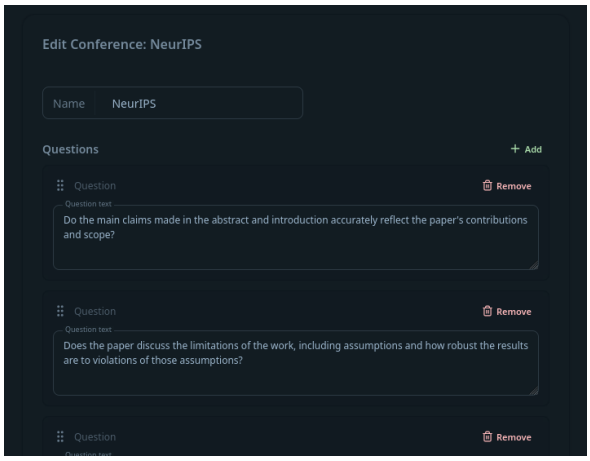


Figure 7: Conference modification screen

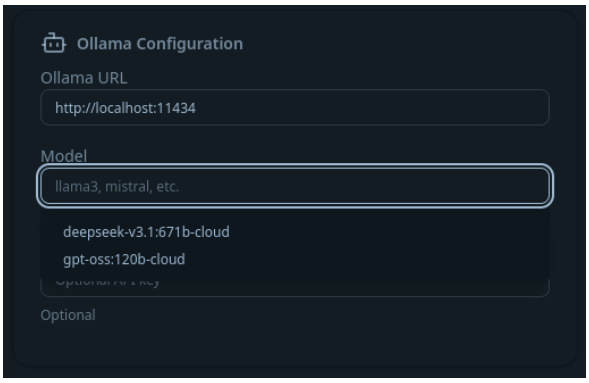


Figure 8: Ollama URL and Model Selector

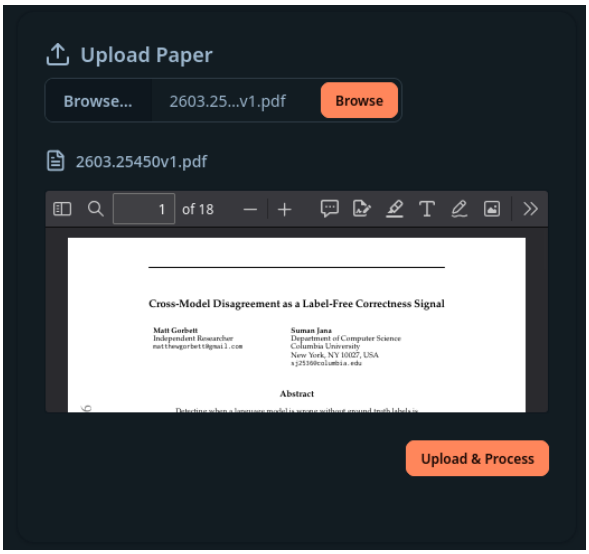


Figure 9: File uploader

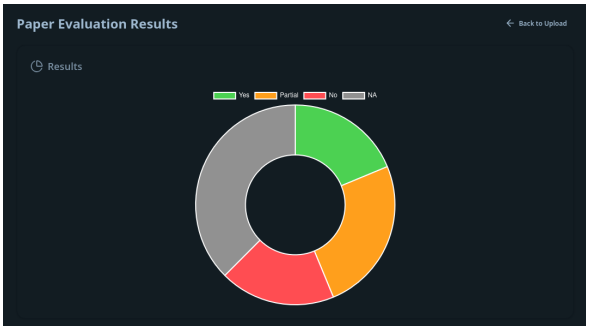


Figure 10: Result graph of the checklist results