Benchmarking LLMs for Automatic Responsible Checklist Generation

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, it can be used as a self-assessment tool for authors, providing them references on how to improve the quality of their submissions. In addition, reviewers can also use checklists to assist them during the review task. Although useful, filling out the checklist is usually a time-consuming task, as it is done manually. LLMs can be a powerful tool for providing assistance for this task due to their capacity to emulate human-like reasoning. This paper presents a study of three different LLMs for the author checklist completion task: GPT-3.5-turbo, DeepSeek-R1, and Llama-3. The 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, the 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) have started providing authors with reproducibility 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.

Additionally, the surge of large language models (LLMs) has significantly impacted the research landscape (Liao et al., 2024). Models such as GPT-3.5 or GPT-4 (Radford et al., 2018) 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 researchers' knowledge is invaluable and irreplaceable (Nature, 2024).

In this context, this paper presents a benchmarking on the performance of different LLMs for the completion of reviewing checklists. Section 2 provides an overview of the related works, and sets the building foundation of our work. Section 3 presents the methodology followed, describing the data and the LLMs that were used, as well as the followed procedure. The results of the study are presented in Section 4, while conclusions and future lines of research are described in Section 5.

2 Related Works

The PRISMA statement (Page et al., 2021), first addressed in 2009, was one of the first attempts to develop review checklists specifically designed for research papers. The idea of using checklists to ensure better quality research papers quickly spread within the research community, leading top-tier conferences such as NeurIPS(NeurIPS, 2025) and AAAI (AAAI, 2024) to design their own review checklists. More recently, Dodge et al. (2019) focused on the development of checklists specifically geared toward AI research and, more specifically, to the reproducibility of the work presented in these papers. In 2020, Dodge and Smith (2020) published the "NLP Reproducibility Checklist", focused on addressing reproducibility aspects of machine learning and, more specifically, Natural Language Processing (NLP) models. The impact of this checklist was then evaluated in Magnusson et al. (2023). According to the authors, the inclusion of checklists not only improved the overall quality of the submissions, but also encouraged the authors to provide more details on their work. In addition, it also encouraged other conference committees to develop their own checklists.

More recently, LLMs have disrupted the research scene, subsequently affecting the review process. Evans et al. (2024) focused on comparing the use of LLMs with respect to human performance to analyze and summarize research articles. According to the authors, the results show a poor correlation, thus supporting the idea that expert knowledge is indispensable and irreplaceable. Other works, such as Liang et al. (2023) explore the use of LLMs to automatically generate peer reviews. The authors compare the overlap between human and LLM-generated revisions, leading to the discovery of significant biases in the reviews generated by the LLM due to their issues regarding deeper understanding. However, the goal of this work is not to replace the role of human reviewers but to provide useful feedback to authors for further improvement. Liu and Shah (2023) also focused on the application of LLMs in the review process, conducting an exploratory study on three different aspects: identifying errors, verifying checklists, and choosing the "better" paper. Regarding checklist verification, the authors conducted an evaluation using GPT-4 in 15 NeurIPS articles, achieving 86.6% precision. More recently, Goldberg et al. (2024) aimed to replace the author's role for checklist completion in NeurIPS'24 submissions. Although the core objective of this work is similar to ours, they focus on evaluating the user experience after receiving feedback from the LLM on the checklist points. Subsequently, the authors do not compare the responses provided by the LLM (in this case, GPT-4) with the actual responses of the authors. Moreover, they rely exclusively on GPT-4, not considering other free-to-use LLMs.

3 Benchmarking LLMs for Automatic Checklist Completion

This work presents a benchmark on the performance of different LLMs on completion of the review checklist. The goal is to assess whether LLMs can accurately reflect the behavior of human reviewers and 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). Three different LLMs were considered for benchmarking: GPT-3.5-turbo, Llama-3 (Grattafiori et al., 2024), and DeepSeek-R1 (DeepSeek-AI et al., 2025). These three models present different sizes and features, with GPT-3.5-turbo being the largest one (175B parameters) and DeepSeek-R1 having the least parameters (7B). This model selection serves twofold purpose. First, it sets a basis to assess whether smaller models can also be used for research-related tasks, since GPT-4 is usually the preferred model as evidenced in the literature. Secondly, it also opens the opportunity for the development of free-to-use research tools based on open-source LLMs.

Human-completed checklists are required to compare whether the LLM achieves the same conclusions on the aspects evaluated. Therefore, NeurIPS'22 accepted papers are selected to conduct the experimentation, since they include, in addition to the paper itself, the checklist filed 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 as similar as possible. According to the conference guidelines, the authors must respond to each

```
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 ethics review
    guidelines and ensured that your
    paper conforms to them?
```

Listing 1: Author checklist section.

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

Listing 2: Theoretical results checklist section.

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. The NeurIPS'22 review checklist contains several questions, which are divided into five categories. The first category presents a series of questions aimed at the authors that are shown in Listing 1. These first series of points address the content aspects of the paper, especially points (a) to (c). 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 correctly to this point.

The second block refers to theoretical results, as depicted in Listing 2. In this case, both questions can be answered based on the content of the paper. However, the answer to these questions may not be trivial since it requires a deep level of understanding. First, it requires the LLM to discern whether the paper reports theoretical results and which are. Then, on the basis of the theoretical results extracted, the LLMs must answer the proposed questions. Subsequently, these questions may be a good reference point to assess the reasoning capacity of the studied LLMs.

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

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

Listing 3: Experiments checklist section.

```
If you are using existing assets, or
    curating/releasing new assets...
(a) If your work uses existing assets,
     did you cite the creators?
(b) Did you mention the license of the
     assets?
(c) Did you include any new assets
    either in the supplemental material
     or as a URL?
(d) Did you discuss whether and how
    consent was obtained from people
    whose data you're using/curating?
(e) Did you discuss whether the data
    you are using/curating contains
    personally identifiable information
     or offensive content?
```

Listing 4: Assets checklist section.

concise and targeted, thus requiring a lower level of understanding. Subsequently, LLMs are expected to perform well in answering these points. The fourth and final block of the checklist refers to the work's assets, which are outlined in Listing 4. Similarly to the previous block, the answer to these questions should ideally be explicitly declared in the paper and therefore should not require a high level of understanding. It should be noted that the original checklist comprises five blocks of questions, the last relating to crowd-sourcing or research conducted with human subjects. These questions all address aspects external to the paper and therefore it would be impossible to accurately answer them just from the content itself. Therefore, they are not considered in our work, since a human reviewer would not be capable of answering them either.

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 filed 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 filed not only on the basis

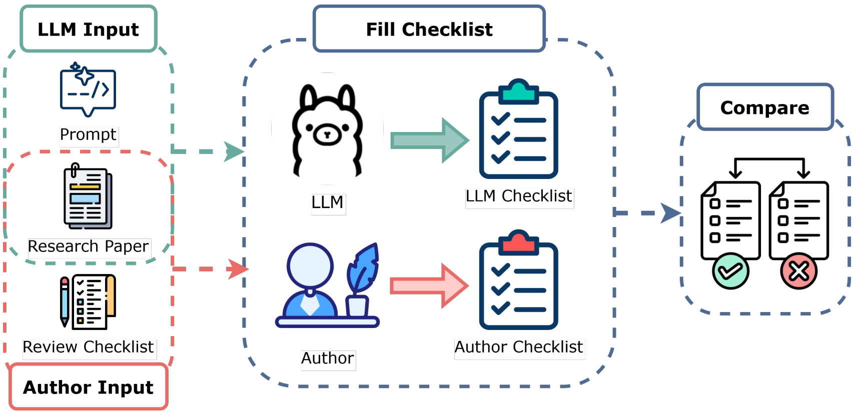


Figure 1: Overview of the conducted workflow.

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.

According to the conference guidelines and as previously stated, authors are required to respond to each checklist point, but are not required to provide evidence supporting their response. In our experiments, since we wanted to assess whether the answers provided by the LLM are actually based on the content of the paper and not a product of hallucination, we explicitly ask in the prompt to indicate the evidence in the content in which the answer is based. Moreover, we introduce an additional layer of granularity to the responses, distinguishing between whether a point is "fully covered" or "partially covered" in the paper. The response is returned in JSON format for further processing. For easier processing, each part of the checklist is asked individually, leading to four different prompts: one for the authors checklist, one for the theoretical results, one for the experiments, and one for the assets. Appendix A provides the detailed prompts per checklist section and LLM.

Once the LLM checklist is clean and processed, it can then be compared with the human-filed 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'22 papers is used as a benchmark for the experimentation. The scraping was first performed to retrieve both the metadata of each paper, along with the file files of the article in PDF format. A total of 2,671 papers were first retrieved. The second step comprises the extraction of the author checklist from the PDF file. First, using the PyPDF library, 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 stored.

Once the authors' checklist (or base 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 results, parsing and post-processing steps are required to extract the clean content from the LLM responses. Despite being explicitly declared in the prompt, which is the expected response format, GPT-3.5-turbo is the only model that provides the response as a JSON file with the specified fields. In the case of DeepSeek-R1 and Llama-3, some of the specified fields suffered mutations in the response process (for example, the field "point" was replaced by "question"), and variations in the format of the responses. For example, the responses for the authors' checklist are returned as a list of entries, but

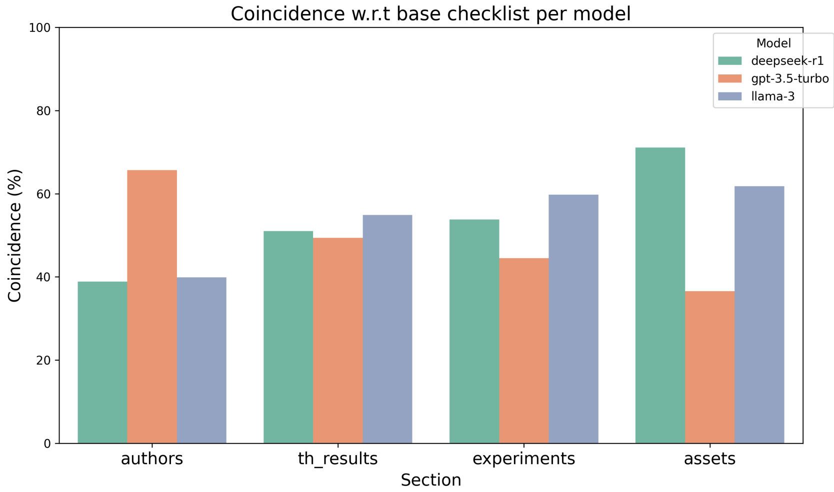


Figure 2: Comparison per LLM on the entire checklist per section w.r.t. to the base checklist.

for the assets checklist, the responses are provided as a set of independent entries. In some cases, some malformations were detected in the JSON files contained in the responses. Therefore, extensive and detailed post-processing was required to obtain the clean results for further comparison. After post-processing, of the 2,671 initial papers, only 575 papers could be fully processed by the three LLMs without errors.

4 Evaluation

As stated in Section 3, three different LLMs are considered for the benchmarking: GPT-3.5-turbo, Llama-3, and DeepSeek-R1. For GPT-3.5-turbo, the *llm*[1] Python package was used to query and retrieve responses. For Llama-3 and DeepSeek-R1, *Ollama* was used. The complete code for scrapping and preprocessing the data, together with the snippets for the execution and post-processing steps for each model, is available on GitHub [2].

After post-processing, an additional homogenization step was required to allow a direct comparison between the baseline checklist (filled by the authors) and the LLM-filled checklist. Since the authors could reply with "N/A", but this answer is unreachable by LLMs without external context, "N/A" has been mapped to the "No" answer. Additionally, the granular answers "partially" and "fully" have been mapped to the "Yes" answer, since the authors are not required to make this distinction.

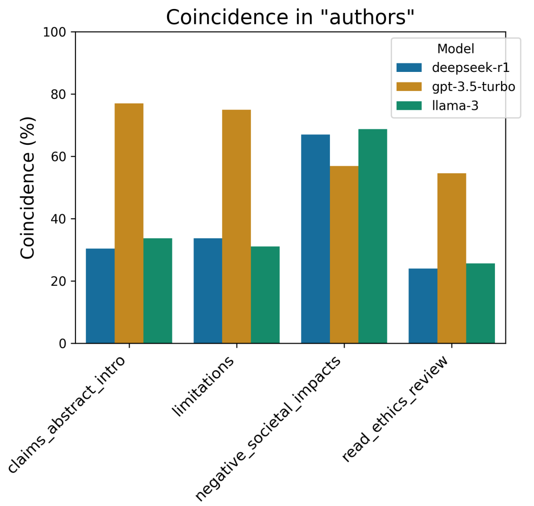
[1]https://pypi.org/project/llm/

[2]https://github.com/eamadord/LLMReproducibilityChecklist

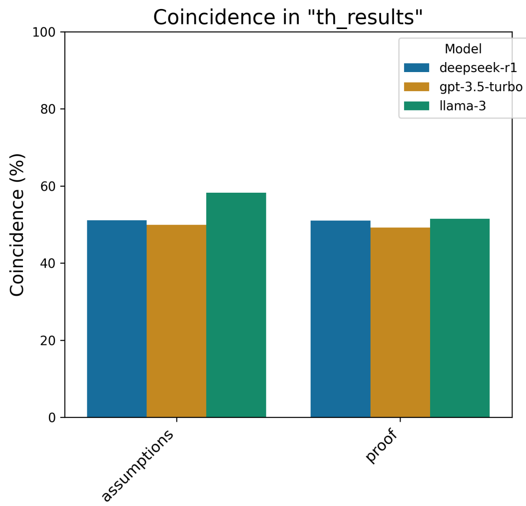
4.1 General comparison

Figure 2 provides the results per checklist section for each LLM. The results show a disparity in the behavior of each LLM with respect to the block in question. As expected, the first block of questions, targeted towards the authors, was the most difficult to answer by the LLM, with only GPT-3.5-turbo reaching over 50% of coincidence in the answers. In the case of the second block (theoretical results) and the third block (experiments), the coincidence proportion by LLM is almost identical in both cases. Regarding the questions related to theoretical results, which were expected to be a bit challenging due to the level of comprehension they require, almost all LLMs behaved equally. This may also be due to the fact that this is the smallest block comprising only a couple of questions, opposite to the five that comprise the third block. Moreover, with regard to the third block, the results are fairly similar in the three LLMs, with Llama-3 achieving the best results. Finally, with respect to the fourth block (assets), there is a significant disparity between the performance of the three LLMs. Although DeepSeek-R1 achieved a coincidence value of more than 70%, the coincidence achieved by GPT-3.5-turbo is lower than 40%. Considering that GPT-3.5-turbo is the biggest model and therefore was expected to have the best performance of all three, this is a very remarkable finding. Moreover, the GPT-3.5-turbo did not show any remarkable improvement in performance with respect to the other two smaller models.

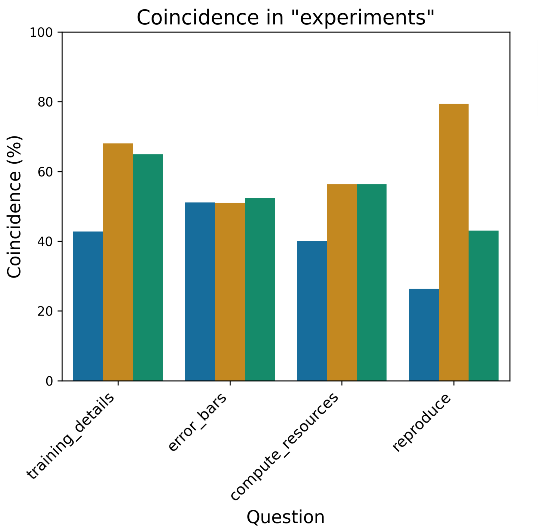
Figure 3 provides a closer look at the coincidence achieved by LLM with respect to each of



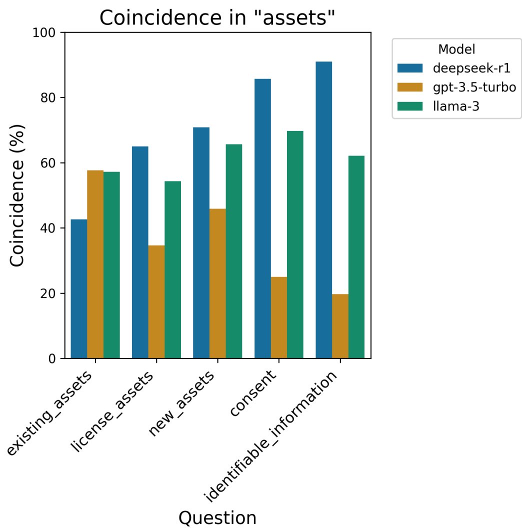
(a) Coincidence w.r.t base checklist for the author block of questions.



(b) Coincidence w.r.t base checklist for the theoretical results block of questions.



(c) Coincidence w.r.t base checklist for the experiments block of questions.



(d) Coincidence w.r.t base checklist for the assets block of questions.

Figure 3: Comparison per LLM on the questions per section w.r.t to the base checklist. In blue, the results achieved by DeepSeek-R1, in yellow the results achieved by GPT-3.5-turbo and, in green, the results achieved by llama-3.

the questions in each section. Regarding the first block of questions, which addressed author-related aspects, GPT-3.5-turbo performed significantly better. This is especially noticeable in questions that require a higher level of abstraction and understanding. Although the coincidence with respect to the baseline in questions such as *Do the main claims made in the abstract and introduction accurately reflect the paper's contribution and scope?* or *Did you describe the limitations of your work?* is close to 80% for GPT-3.5-turbo, it barely exceeds 30% for the other two. This behavior demonstrates that, for more complex and human-like questions that require a higher level of abstraction, small LLMs may be insufficient. In the third question, all three LLMs achieved similar results, but this is directly related to a bias in the data, since around 70% of the responses in the baseline checklist for this question are either "No" or "N/A", which both map to "No". Therefore, both DeepSeek-R1 and Llama-3 systematically replied "No" to this question, which directly relates to the coincidence in the results.

In the second block of questions, there is a slightly higher homogeneity among the responses provided by the LLMs. However, it should be noted that the coincidence is close to 50% for all models. Although the answer to both questions can be inferred from the content of the paper and is not subject to contextual aspects, it is worth noting that both questions require a high level of abstraction, making them difficult to reply.

Regarding the experimental results block (Figure 3c), there is a notable disparity among the responses provided for the different questions. From a general perspective, GPT-3.5-turbo has the highest coincidence with respect to the results provided by the authors. This is especially noticeable in the reproducibility question, in which the coincidence achieved by GPT-3.5-turbo to the question *Did you include the code, data, and instructions needed to reproduce the main experimental results?* is around 80%, while for Llama-3 and DeepSeek-R1 does not even reach the 40%. For the remaining questions, all coincidences are close to 50%. In the case of the question *Did you include error bars?*, this low value may be due to the negative responses provided for all models to this question, since it relates to the content of graphical elements within the paper, which may not be properly rendered when converted to plain text.

Finally, with respect to the asset-related block of questions, there is a striking difference between the

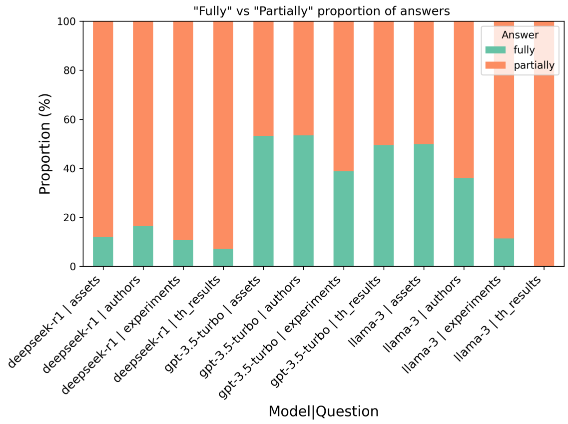


Figure 4: Proportion of "fully" vs "partially" answers per LLM.

performance of the different models per question. In the last two questions, regarding consent and the removal of identifiable information, the high coincidence in the responses provided by DeepSeek-R1 is actually the result of a bias in the data, since more than 90% of the samples received a negative response (either "No" or "N/A") from the authors, while DeepSeek-R1 answer negatively for all samples, thus leading to a significant overleap. Since these questions are related to contextual aspects and may not be easily inferred from the content of the article, the low coincidence by GPT-3.5-turbo, which answered positively for more than 85% of the samples, may be due to hallucinations in the inference process.

4.2 Response comparison

In addition to answering positively or negatively to each point in the checklist, LLMs were also queried to provide an additional level of granularity, answering "fully" or "partially". The goal of this additional level of granularity is to serve as an indicator to authors on whether a point in the checklist is fully met or certain changes are required to fully address the point. Figure 4 depicts the proportion of "fully met" vs. "partially met" answers per LLM and section. It can be observed that for DeepSeek-R1 and Llama-3, the "fully" response is less featured, with the exception of the questions related to assets, to which Llama-3 used the "fully" answer in around half of the samples. GPT-3.5-turbo is the only LLM that consistently uses both answers in all fields. Therefore, the responses provided by both Llama-3 and DeepSeek-R1 tend to be more negative in terms of point completion. This may be due to these LLMs not having the same ca-

Model	Response	Evidence
GPT-3.5-turbo	"fully"	*The paper clearly states and discusses all assumptions made in deriving theoretical results*
Llama-3	"partially"	*The paper assumes that the dataset is random and independent, but does not explicitly state this assumption*
DeepSeek-R1	"not at all"	*The text does not mention any theoretical results or their assumptions*

Table 1: Example of the responses of the different LLMs for the same sample to the question *Does the paper state the full set of assumptions of all theoretical results?*

pacity to handle big contexts, and therefore missing important information within the paper that may lead to correctly and accurately answering each point.

An example of this mismatch in the responses is provided in Table 1, where the response for the same question on the same sample is provided in the different LLMs. Although GPT-3.5-turbo and Llama-3 both responded positively to this point, their responses regarding the level of completion of the given point are different. GPT-3.5-turbo states that all assumptions are clearly discussed, while Llama-3 argues that the paper assumes that the dataset is random and independent and points to Section 2.1 within the paper for further evidence. However, this argument does not actually apply to the content of the paper and may be actually a product of hallucination due to the repeated use of terms "independent variables" and "random features" within the article, which incorrectly leads to the provided response. This phenomenon occurs in multiple samples, which may show a pattern of potential hallucinations in the responses provided by Llama-3. After reviewing the evidence and the results achieved by the different LLMs, the GPT-3.5 turbo clearly showed the most reliable performance.

5 Conclusions and Future Work

This paper shows 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 three different LLMs: GPT-3.5-turbo, Llama-3, and DeepSeek-R1 for the checklist completion task in the context of the NeurIPS 2022 conference. This study prompts the checklists into 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 can not be inferred directly from the paper. In addition, the results show that the GPT-3.5-turbo achieved the best performance from a general perspective, with Llama-3 and DeepSeek-R1 matching performance for some of the points.

Furthermore, in addition to answering "Yes" or "No" at each point, LLMs were asked to indicate whether each point was met "fully" or "partially" and to provide evidence supporting this point. GPT-3.5-turbo provided both answers in a relatively similar proportion, while Llama-3 and DeepSeek-R1 tended to answer "partially" to most points. Looking at the evidence provided by the models, Llama-3 suffered hallucinations, providing supporting evidence that related to aspects that were not described in the paper. This phenomenon may be related to a smaller size in terms of both the parameters and context window.

Future works include extending the study to other LLMs, to extract more complete and clearer behavior patterns on their suitability for the task. Moreover, it would be interesting to assess whether there is a difference in the performance of general-use LLMs, such as those studied in this paper, with respect to LLMs trained specifically for research, such as Nous-Hermes2[3]. In the same vein, the benchmark could also be extended to cover the checklists of different conferences, such as AAAI. This would provide a wider vision on the actual capacities of LLMs for this task.

Finally, another line of research would be the development of LLM-built assistance frameworks to assist authors, as well as reviewers, in author completion tasks. In addition, the results of these checklists can also be returned as feedback to the authors, helping them detect potential gaps in their work.

[3]https://ollama.com/library/nous-hermes2

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. Additionally, the post-processing of the LLM outputs has been one of the most challenging points of the process, since despite the expected output format being explicitly declared in the prompt, only GPT-3.5-turbo actually produced valid outputs. For Llama-3 and DeepSeek-R1, hand-crafted rules needed to be devised to extract the information from the provided output into a structured format that could subsequently be used for comparison. This issue resulted in a significant loss of valid samplings, leaving only 1/5 of the total amount of samples for comparison. Finally, since only three LLMs and one checklist example were considered for the study, the extracted patterns are bounded to just the context of the study and, therefore, may not be an indicator of the behavior of similar models. For example, even though they are the same size, it cannot be ascertained whether Qwen3 will behave like Llama-3 or will have a different behavior.

References

AAAI. 2024. Aaai reproducibility checklist.

DeepSeek-AI, Aixin Liu, Bei Feng, Bing Xue, Bingxuan Wang, Bochao Wu, Chengda Lu, Chenggang Zhao, Chengqi Deng, Chenyu Zhang, Chong Ruan, Damai Dai, Daya Guo, Dejian Yang, 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, Haowei Zhang, Honghui Ding, Huajian Xin, Huazuo Gao, Hui Li, Hui Qu, J. L. Cai, Jian Liang, Jianzhong Guo, Jiaqi Ni, Jiashi Li, Jiawei Wang, Jin Chen, Jingchang Chen, Jingyang Yuan, Junjie Qiu, Junlong Li, Junxiao Song, Kai Dong, Kai Hu, Kaige Gao, Kang Guan, Kexin Huang, Kuai Yu, Lean Wang, Lecong Zhang, Lei Xu, Leyi Xia, Liang Zhao, Litong Wang, Liyue Zhang, Meng Li, Miaojun Wang, Mingchuan Zhang, Minghua Zhang, Minghui Tang, Mingming Li, Ning Tian, Panpan Huang, Peiyi Wang, Peng Zhang, Qiancheng Wang, Qihao Zhu, Qinyu Chen, Qiushi Du, R. J. Chen, R. L. Jin, Ruiqi Ge, Ruisong Zhang, Ruizhe Pan, Runji Wang, Runxin Xu, Ruoyu Zhang, Ruyi Chen, S. S. Li, Shanghao Lu, Shangyan Zhou, Shanhuang Chen, Shaoqing Wu, Shengfeng Ye, Shengfeng Ye, Shirong Ma, Shiyu Wang, Shuang Zhou, Shuiping Yu, Shunfeng Zhou, Shuting Pan, T. Wang, Tao Yun, Tian Pei, Tianyu Sun, W. L. Xiao, Wangding Zeng, Wanjia Zhao, Wei An, Wen Liu, Wenfeng Liang, Wenjun Gao, Wenqin Yu, Wentao Zhang, X. Q. Li, Xiangyue Jin, Xianzu Wang, Xiao Bi, Xiaodong Liu, Xiaohan Wang, Xiaojin Shen, Xiaokang Chen, Xiaokang Zhang, Xiaosha Chen, Xiaotao Nie, Xiaowen Sun, Xiaoxiang Wang, Xin Cheng, Xin Liu, Xin Xie, Xingchao Liu, Xingkai Yu, Xinnan Song, Xinxia Shan, Xinyi Zhou, Xinyu Yang, Xinyuan Li, Xuecheng Su, Xuheng Lin, Y. K. Li, Y. Q. Wang, Y. X. Wei, Y. X. Zhu, Yang Zhang, Yanhong Xu, Yanhong Xu, Yanping Huang, Yao Li, Yao Zhao, Yaofeng Sun, Yaohui Li, Yaohui Wang, Yi Yu, Yi Zheng, Yichao Zhang, Yifan Shi, Yiliang Xiong, Ying He, Ying Tang, Yishi Piao, Yisong Wang, Yixuan Tan, Yiyang Ma, Yiyuan Liu, Yongqiang Guo, Yu Wu, Yuan Ou, Yuchen Zhu, Yuduan Wang, Yue Gong, Yuheng Zou, Yujia He, Yukun Zha, Yunfan Xiong, Yunxian Ma, Yuting Yan, Yuxiang Luo, Yuxiang You, Yuxuan Liu, Yuyang Zhou, Z. F. Wu, Z. Z. Ren, Zehui Ren, Zhangli Sha, Zhe Fu, Zhean Xu, Zhen Huang, Zhen Zhang, Zhenda Xie, Zhengyan Zhang, Zhewen Hao, Zhibin Gou, Zhicheng Ma, Zhigang Yan, Zhihong Shao, Zhipeng Xu, Zhiyu Wu, Zhongyu Zhang, Zhuoshu Li, Zihui Gu, Zijia Zhu, Zijun Liu, Zilin Li, Ziwei Xie, Ziyang Song, Ziyi Gao, and Zizheng Pan. 2025. Deepseek-v3 technical report.

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.

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.

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.

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.

Alec Radford, Karthik Narasimhan, Tim Salimans, and Ilya Sutskever. 2018. Improving language understanding by generative pre-training.

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

A Prompts per LLM

This Appendix presents the

A.1 GPT-3.5-Turbo Prompts

Author Checklist *"Check whether the attached file meets the points on this checklist. Return the results in JSON format, where for each point in the checklist report whether it is met fully, partially or not at all in a field called 'score'. Provide evidence for each point as well in another JSON field called 'evidence'. Checklist: -Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope? -Does the paper describe the limitations of the work? -Does the paper discuss any potential negative societal impacts of your work? -Does the paper address the ethics review guidelines?"*

Theoretical Results Checklist *"Check whether the attached file meets the points on this checklist. Return the results in JSON format, where for each point in the checklist report whether it is met fully, partially or not at all in a field called 'score'. Provide evidence for each point as well in another JSON field called 'evidence'. Checklist: -Does the paper state the full set of assumptions of all theoretical results? -Does the paper include complete proofs of all theoretical results"*

Experiments Checklist *Check whether the attached file meets the points on this checklist. Return the results in JSON format, where for each point in the checklist report whether it is met fully, partially or not at all in a field called 'score'. Provide evidence for each point as well in another JSON field named 'evidence'. Checklist: -Does the paper include the code, data, and instructions needed to reproduce the main experimental results? -Are all the training details specified? -Are error bars reported? -Is the total amount of compute and the type of resources used included in the paper*

Assets Checklist *Check whether the attached file meets the points on this checklist. Return the results in JSON format, where for each point in the checklist report whether it is met fully, partially or not at all in a field called 'score'. Provide evidence for each point as well in another JSON field named 'evidence'. Checklist: -If the work references existing assets, are these assets properly cited? -Is the license of the assets mentioned? -Are new assets included either in the supplemental material or in the URL? -Does the paper discuss whether and how the consent was obtained from people whose data is used/curated? -Does the paper discuss whether the data used/curated contains personally identifiable information or offensive content?*

A.2 Llama-3 Prompts

Author Checklist *From this paper text, tell me whether it meets the points of the following checklist. Return the results in JSON format, where for each point in the checklist report whether it is met fully, partially or not at all in a field called 'score'. Provide evidence for each point as well in another JSON field called 'evidence'. Checklist: -Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope? -Does the paper describe the limitations of the work? -Does the paper discuss any potential negative societal impacts of your work? -Does the paper address the ethics review guidelines?*

Theoretical Results Checklist *From this paper text, tell me whether it meets the points of the following checklist. Return the results in JSON format, where for each point in the checklist report whether it is met fully, partially or not at all in a field called 'score'. Provide evidence for each point as well in another JSON field called 'evidence'. Checklist: -Does the paper state the full set of assumptions of all theoretical results? -Does the paper include complete proofs of all theoretical results*

Experiments Checklist *From this paper text, tell me whether it meets the points of the following checklist. Return the results in JSON format, where for each point in the checklist report whether it is met fully, partially or not at all in a field called 'score'. Provide evidence for each point as well in another JSON field named 'evidence'. Checklist: -Does the paper include the code, data, and instructions needed to reproduce the main experimental results? -Are all the training details specified? -Are error bars reported? -Is the total amount of compute and the type of resources used included in the paper*

Assets Checklist *From this paper text, tell me whether it meets the points of the following checklist. Return the results in JSON format, where for each point in the checklist report whether it is met fully, partially or not at all in a field called 'score'. Provide evidence for each point as well in another JSON field named 'evidence'. Checklist: -If the work references existing assets, are these* *assets properly cited? -Is the license of the assets mentioned? -Are new assets included either in the supplemental material or in the URL? -Does the paper discuss whether and how the consent was obtained from people whose data is used/curated? -Does the paper discuss whether the data used/curated contains personally identifiable information or offensive content?*

A.3 DeepSeek-R1 Prompts

Author Checklist *The following text is part of a research article: text. From the previous text, tell me whether it meets the points of the following checklist. Return the results in JSON format, where for each point in the checklist, copy its content in a field called "point", and then report whether the point is met fully, partially or not at all in a field called 'score'. Provide evidence for each point as well in another JSON field called 'evidence'. Checklist: -Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope? -Does the paper describe the limitations of the work? -Does the paper discuss any potential negative societal impacts of your work? -Does the paper address the ethics review guidelines?*

Theoretical Results Checklist *The following text is part of a research article: text. From the previous text, tell me whether it meets the points of the following checklist. Return the results in JSON format, where for each point in the checklist, copy its content in a field called "point", and then report whether the point is met fully, partially or not at all in a field called 'score'. Provide evidence for each point as well in another JSON field called 'evidence'. Checklist: -Does the paper state the full set of assumptions of all theoretical results? -Does the paper include complete proofs of all theoretical results*

Experiments Checklist *The following text is part of a research article: text. From the previous text, tell me whether it meets the points of the following checklist. Return the results in JSON format, where for each point in the checklist, copy its content in a field called "point", and then report whether the point is met fully, partially or not at all in a field called 'score'. Provide evidence for each point as well in another JSON field named 'evidence'. Checklist: -Does the paper include the code, data, and instructions needed to reproduce the main experimental results? -Are all the training details specified? -Are error bars reported? -Is the total amount of compute and the type of resources used included in the paper*

Assets Checklist *The following text is part of a research article: text. From the previous text, tell me whether it meets the points of the following checklist. Return the results in JSON format, where for each point in the checklist, copy its content in a field called "point", and then report whether the point is met fully, partially or not at all in a field called 'score'. Provide evidence for each point as well in another JSON field named 'evidence'. Checklist: -If the work references existing assets, are these assets properly cited? -Is the license of the assets mentioned? -Are new assets included either in the supplemental material or in the URL? -Does the paper discuss whether and how the consent was obtained from people whose data is used/curated? -Does the paper discuss whether the data used/curated contains personally identifiable information or offensive content?*