Large Language Models for Reviewing Research Papers

Evaluating Claim-Level Completeness in Machine Learning Research

Simona Ivanova Simeonova[1]
Supervisor(s): David M.J. Tax[1]**, Chenxu Hao**[1]
[1]EEMCS, Delft University of Technology, The Netherlands

A Thesis Submitted to EEMCS Faculty Delft University of Technology,
In Partial Fulfilment of the Requirements
For the Bachelor of Computer Science and Engineering
June 19, 2026

Name of the student: Simona Ivanova Simeonova
Final project course: CSE3000 Research Project
Thesis committee: David M.J. Tax, Chenxu Hao, Hayley Hung, Nergis Tömen, Klaus Hildebrandt

Abstract

Scientific peer review is an important part of the scientific process. However, the growing number of submissions has sparked interest in automated review tools. Recent work has shown that Large Language Models (LLMs) can generate reviews and evaluate author-provided checklists, yet it is unclear to what extent they can independently identify the scientific claims that are made in papers and perform structured reviews. This thesis investigates whether an LLM can automatically extract scientific claims from research papers in the machine learning field and then complete the NeurIPS Checklist without relying on author-written justifications. The evaluation focuses on claim extraction accuracy, preserving the semantic meaning of claims, and agreement between LLM-generated checklist annotations and human judgment. Gemini 3 Flash's claim extraction and checklist annotations are compared against human ground-truth annotations on NeurIPS 2024 papers. The results show that the model successfully identifies primary claims of papers, with a recall of 0.99 and precision of 0.75. Most errors are caused by over-segmentation or incorrect classification. For checklist annotation, the system achieves a mean accuracy of 0.85 and a mean Cohen's Kappa of 0.58 compared to human annotations. Agreement is strongest for objective checklist criteria. These findings indicate that LLMs can effectively support claim-based scientific review, but are not advanced enough to fully replace expert reviewers.

1 Introduction

In recent years, there has been rapid growth in scientific publications. This has placed an increasing strain on the peer-review process. Conferences such as NeurIPS and ICLR receive thousands of submissions, yet the number of capable reviewers has not significantly increased, making it difficult to ensure timely and consistent reviews. This is referred to as "reviewer fatigue" and raises concerns about the efficiency and reliability of the evaluation [12]. Advances in Large Language Models (LLMs) have sparked interest in their potential to assist with or automate the scientific reviewing process.

While prior studies [16, 10] investigate error detection, scoring, and claim-evidence reasoning, there is still limited understanding of how well LLMs can extract the central scientific claims of a paper. Scientific claims were specifically chosen as the main focus of this research, because understanding and assessing a paper's claims is the foundation of peer-review reasoning. Peer review has been categorized as a claim-centric process and requires critique and thorough verification of the claims made by authors [18]. Therefore, this paper addresses the following main research question: *To what extent can a Large Language Model (LLM) extract and summarize the scientific claims made in NeurIPS papers as defined by the official NeurIPS Checklist?* The NeurIPS Checklist is a standardized set of questions that authors must answer, covering topics such as the paper's claims, limitations, theory assumptions and proofs, experimental result reproducibility, and open access to data and code [5].

To answer this question, we consider three sub-questions:

1. Can the LLM accurately extract the primary claims as stated in the abstract and introduction?
2. To what extent does the LLM preserve the original semantic meaning of claims without introducing hallucinations?
3. Does the LLM's binary (Yes/No) response to the NeurIPS Checklist align with human judgment?

To address these questions, we construct an empirical evaluation on a dataset of NeurIPS 2024 Main Conference Track papers. Human annotations are used as the ground truth against which the performance of the Gemini 3.0 Flash model [8] is evaluated. This setup allows to systematically test both the accuracy and reliability of LLM-generated outputs in a more realistic reviewing context.

The main contribution of this work is focused on evaluation of LLM capabilities in peer review, specifically extracting scientific claims, bridging the gap between surface-level assistance and deeper understanding. By using the NeurIPS Checklist for part of our analysis, we provide a reproducible method for assessing LLM performance in peer-reviewing. First, we develop an automated pipeline for scientific paper analysis. It preprocesses NeurIPS submissions, extracts and enumerates scientific claims using the chosen model, and independently completes the NeurIPS checklist with specific justifications. We evaluate the effectiveness of the model for peer review, showing strong claim extraction performance with a low hallucination rate, but a tendency toward redundant claims. We further demonstrate that although the model is able to accurately answer many checklist items, it is not fully reliable enough to act as an autonomous reviewer.

The remainder of this paper is organized as follows. First, related work is introduced in detail to identify a research gap. Then, in section 3, the problem definition and research methodology provide justifications for some design choices that were made. Next, in section 4, the approach and design is defined, including dataset construction, annotation procedure, prompting strategy, and evaluation. This is followed by a description of the experimental setup and the results. Finally, ethical research considerations and limitations of the work are discussed.

2 Related Work

This chapter reviews prior work relevant to automated scientific evaluation using LLMs. It covers research on LLM-based peer review and the use of LLMs as evaluators, automated checklist annotation, and scientific claim verification. These studies provide the foundation for the present work and highlight some limitations of current approaches. This motivates the need for a claim-grounded method for NeurIPS paper evaluation.

2.1 LLMs for Scientific Evaluation

Recent advances in LLMs have led to growing interest in their use for peer review and evaluation. Earlier studies show that models such as GPT-4 can perform specific reviewing tasks including error detection and checklist verification with a satisfactory accuracy [16]. Subsequently these ideas were extended to automated review generation through reinforcement learning or multi-agent systems. This showed that LLM-generated reviews could exceed current automated baselines [21, 2], however they can be more biased than human reviewers [13].

Beyond review generation, researchers have increasingly studied LLMs as evaluators. The LLM-as-a-judge paradigm uses models to assign scores, make binary decisions, compare alternatives, or select options [9]. These capabilities are what make LLMs useful for scientific evaluation tasks. However, several studies have identified important limitations, including possible evaluation biases, calibration issues, and judgment inconsistencies [9, 18]. The growing use of AI-assisted review has also raised concerns about the influence on outcomes and decision-making processes [14, 12]. Because this research evaluates claim extraction and

checklist classification using an LLM, it may be susceptible to similar biases. This motivates human comparison and error analysis in our evaluation.

2.2 Checklist Annotation and Claim Verification

Other works have explored the use of LLMs for more structured assessment. The NeurIPS Checklist Assistant demonstrated that LLM-generated feedback can be useful to authors but since it relies on author-provided justifications it could be more vulnerable to inaccuracies and manipulation [7]. Moreover, claim verification research has investigated whether claims are supported by the available evidence in the paper. Recent approaches combine techniques to verify claim-evidence relationships and improve reliability of research [20, 10]. LLMs have also been shown to identify, classify, and align claims across papers with promising accuracy [17, 19]. However, these approaches focus on claim extraction and verification as separate tasks rather than integrating them for a more complete review.

2.3 Research Gap

Existing research shows that LLMs can generate peer reviews and perform structured evaluations by extracting and verifying claims. However, most approaches verify on author-written justifications, evidence in later sections of the paper, or separation of extraction and verification stages. Our work addresses that gap by investigating a claim-grounded approach to reviewing, while also assessing performance of an easily-available model on NeurIPS Checklist evaluation. Rather than relying on authors' explanations, the proposed pipeline extracts claims directly from the paper and answers the checklist questions independently. This enables an evaluation of whether LLMs can perform peer reviews in a single workflow.

3 Problem Definition and Methodology

This chapter introduces the problem definition and methodology of this study. The main question addressed is whether a currently accessible LLM can accurately extract scientific claims and correctly interpret checklist-related requirements at a level consistent with human judgments in peer review settings. To answer this question, the chapter first formalizes the problem setting by providing some necessary definitions. Then, it presents the proposed methodology, explaining the design choices made.

3.1 Problem Formulation for LLM Peer Review

The NeurIPS conference is a multi-track venue for machine learning research, featuring peer-reviewed paper presentations and an exposition on machine learning applications [4]. A key component of papers submitted to NeurIPS is the NeurIPS Paper Checklist, which is designed to promote responsible research practices by ensuring transparency, reproducibility, and ethical awareness. Authors are required to complete a set of binary (Yes/No/NA) questions addressing various aspects of the research. The checklist is used in the review process as part of paper evaluation criteria and author answers can support their choices [5].

In this work, we define a scientific claim as a statement appearing in the abstract or introduction of a paper that asserts a contribution, result, property, or comparison presented by the authors as an outcome or achievement of the work. These claims are central to peer

evaluation, as reviewers have to assess whether such claims are supported by the evidence provided in the paper.

3.2 Methodology

The overall pipeline consists of four main stages that form an end-to-end framework for evaluating LLM performance on scientific understanding tasks, as illustrated in Figure 1. First, at the "Parser" stage, NeurIPS PDF papers were preprocessed to remove only the author-completed checklist while preserving the remaining scientific content. Second, the processed papers were provided as input to the LLM, which performs both scientific claim extraction and binary checklist evaluation using structured prompts. In parallel, human annotations were generated to establish ground truth for both tasks. Third, model outputs were compared against the corresponding human annotations to assess semantic correctness and agreement. Finally, quantitative and qualitative evaluation was performed through task-specific metrics, error categorization, and pattern analysis, allowing for a detailed assessment of the model's reasoning capabilities over scientific papers.

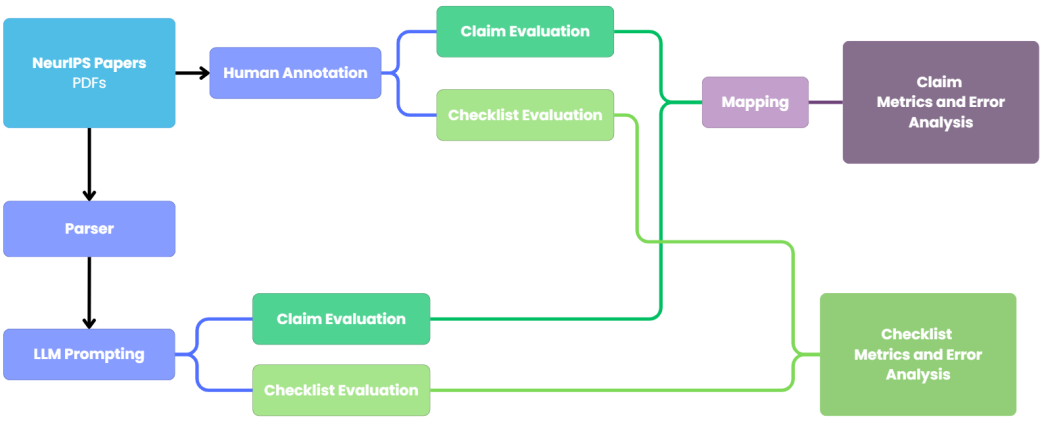


Figure 1: Overview of the Evaluation Pipeline

We evaluate the dataset using *gemini-3-flash-preview*. This model was selected due to its strong balance between performance and accessibility among free available models. Alternative models were considered during model selection at the beginning of the project. However, *gemini-3-flash-preview* provides a larger context window, multimodal input support, and stronger overall benchmark performance [1]. These characteristics are particularly relevant for scientific peer-review tasks, as research papers are long documents that often contain visual elements such as figures, graphs, or diagrams. The larger context window would further enable processing of complete paper submissions without extensive segmentation, making the model more suitable for evaluating scientific content in a realistic review setting.

The model is evaluated using zero-shot structured prompting combined with role-play instructions and with explicitly defined constraints. The prompts ensure that model responses follow a predefined structure suitable for automated evaluation later on. In addition, rule-based filtering is applied through explicit inclusion and exclusion criteria to reduce ambiguity

in generation. The full prompts can be found in Appendix A. Overall, structured prompting and calibration techniques have been shown to improve the reliability and consistency of large language model outputs, particularly in information extraction and reasoning tasks by reducing prompt sensitivity and improving output controllability [15]. Zero-shot prompting was combined with role-playing to improve task immersion and output consistency. Prior work has shown that role-play can enhance reasoning performance without requiring task-specific examples [11].

4 Approach and Design

This chapter presents the design of the approach used for evaluating LLMs in the context of scientific peer review. It describes the dataset construction, annotation procedure, model used, prompting strategy, and evaluation methodology to assess the results. The goal of this chapter is to provide a clear description of the setup.

4.1 Dataset and Ground Truth Construction

This study uses a dataset of 80 NeurIPS 2024 main conference track papers, selected as the first 80 papers listed on the official NeurIPS proceedings website at the time of data collection. The ordering criterion used by the proceedings website is not documented. These papers are publicly available and represent recent peer-reviewed machine learning research. Supplementary material was excluded and only the main paper PDFs were used as input to the pipeline. From this collection, two evaluation subsets were derived.

First, a subset of 20 papers was used for claim-level evaluation, where a human annotator manually extracted scientific claims from the abstract and introduction and constructed a corresponding ground truth set. The claims were extracted following a strict interpretation of the authors' statements, excluding descriptive or definitive text to ensure focus on explicit scientific claims.

Second, all 80 papers were used for checklist evaluation, where five human annotators independently completed the NeurIPS Paper Checklist, excluding the author-provided responses used for comparison. Each annotator evaluated 20 distinct papers, with an overlap of 5 papers between annotators to provide some validation and increase reliability. They provided Yes/No/NA labels and a short justification for each answer was given, if possible stating the exact part of the paper that supported their answer.

4.2 Data Preprocessing

All papers were processed from PDF format using a custom Python-based pipeline. The pipeline removes only the author-completed checklist section to prevent leakage of answers that may match the ground truth while preserving the full scientific content of each paper. This design ensures that the LLM operates under realistic conditions similar to an automated peer-review setting, where full submissions are available rather than just isolated sections.

4.3 Model Configuration and Prompting Strategy

The evaluation used a single-model setup based on *gemini-3-flash-preview*. A structured prompting framework was designed to ensure consistency of outputs. Each prompt includes a clearly defined role and context, instructing the model to act as a peer reviewer for NeurIPS

2024 papers. The prompts specify a main task, either scientific claim extraction, as shown in Appendix A.1, or checklist evaluation, as shown in Appendix A.2, along with explicit data selection rules that define what parts of the input should be considered or ignored. In addition, the prompts include formal definitions relevant to each task, such as the definition of a scientific claim for the claim task or the interpretation of checklist responses for the checklist task. The output is constrained to a specific JSON format, ensuring structured and machine-readable results. For the claim extraction task, each extracted claim includes its statement, location in the paper, and type classification as either a contribution, comparison, property, or result. For the checklist task, the model produces Yes/No/NA answers for each question, along with justifications that include supporting evidence from the paper. The model's responses are then saved to separate files for the claim and the checklist tasks. These design choices are intended to enable detailed post-hoc analysis, error categorization, and quantitative evaluation.

4.4 Evaluation

The evaluation was conducted separately for claim extraction and checklist reasoning.

For the claim extraction task, model outputs were matched against human-annotated claims. A human claim is considered covered if at least one LLM-generated claim corresponds to it. An LLM claim is considered correct if it maps to at least one human-identified claim. Based on these mappings, several metrics were computed, including precision showing the amount of correct LLM-extracted claims, recall showing the number of covered human claims, total number of extracted claims, redundancy showing multiple LLM claims mapping to a single human claim, and the average number of LLM-extracted claims per human-extracted claim. In addition, the hallucination rate was assessed through manual inspection of extracted claims. In this research, a hallucination is defined as any claim generated by the LLM that is not explicitly supported by the evidence in the scientific paper.

For the checklist evaluation task, performance was measured using accuracy of LLM checklist answers and Cohen's Kappa to assess agreement between the LLM and human annotators. Kappa was included because it accounts for agreement due to chance [3].

Furthermore, qualitative error analysis was conducted for both tasks to identify recurring patterns in incorrect or inconsistent responses, providing additional insight beyond aggregate performance metrics. Together, these evaluation procedures enable a comprehensive assessment of the factual extraction capability of LLMs in structured scientific understanding tasks.

5 Experimental Setup and Results

This chapter describes the experimental configuration used to evaluate the research questions using the proposed pipeline. It aims to ensure reproducibility and present the system's performance across the three established sub-questions. It provides quantitative and qualitative evidence for the model's performance as a peer-reviewer.

5.1 Experimental Setup

The pipeline was implemented in Python 3.13 (64-bit) and executed on a laptop equipped with a 12th Generation Intel Core i9-12900H processor, 32 GB of RAM, and an NVIDIA

GeForce RTX 3070 Ti Laptop GPU with 8 GB of dedicated memory. Although a dedicated GPU was available, the language model inference was performed through the Google GenAI API and therefore executed on a remote infrastructure.

The implementation consisted of a checklist trimming script for PDF processing, a claim extraction script, a checklist annotation script, and evaluation scripts for both the claim and the checklist tasks. The processing was performed using the *fitz* library, while evaluation and analysis relied on *pandas*, *numpy*, *matplotlib*, and *scikit-learn*. The complete source code is available online [1].

All claim extraction and checklist annotation experiments used the Gemini 3 Flash Preview model through the Google GenAI API. To maximize determinism and reduce output variability as much as possible, a temperature of 0.0 and a fixed random seed of 42 were used throughout the experiments. For each of the tasks, each paper was processed independently using a single prompt. The system iterated over all papers in the dataset and submitted one API request per paper. Runtime varied depending on API latency and model availability. A representative measurement showed that processing three consecutive papers required 274.832 seconds, corresponding to an average runtime of about 92 seconds per paper.

5.2 Claim Extraction Accuracy

To evaluate whether the primary scientific claims can be identified from the abstract and introduction, claims extracted by the model were compared against manually-identified claims. Table 1 summarizes the overall results across the 20 evaluated papers.

Table 1: Overall Claim Extraction Metrics

Metric	**Value**
Total human claims	98
Total LLM claims	168
Covered human claims	97
Human claim coverage (Recall)	0.99
Correct LLM claims (Precision)	0.75
Redundancy	1.71
Average LLM claims per human claim	1.31

The results show that the model was successful at identifying the main contributions of the papers. Out of the 98 claims identified by the human annotator, 97 were covered by at least one LLM-extracted claim. This corresponds to the high recall of 0.99, which suggests that the model rarely missed a major contribution when it was explicitly stated by the authors. However, the precision was lower at 0.75, meaning that approximately a quarter of the LLM-extracted claims did not correspond to human-annotated claims. Upon closer inspection of the outputs, it was made clear that these additional claims were often not fully incorrect. The model decomposed a single high-level contribution into multiple smaller statements or extracted certain implementation details, properties, and definitions that a human annotator would not classify as a primary claim. This is reflected in the redundancy score and average LLM claims per human claim. Rather than missing claims, the model tended to produce more claims than necessary and the resulting mapping of the claims was many-to-many.

[1] `https://gitlab.tudelft.nl/cse3000/analysisMLresults/-/tree/simona-project?ref_type=heads`

Figure 2 shows the average coverage depth for each paper. Coverage depth measures how many LLM-extracted claims were mapped to each human claim. Higher values indicate greater redundancy.

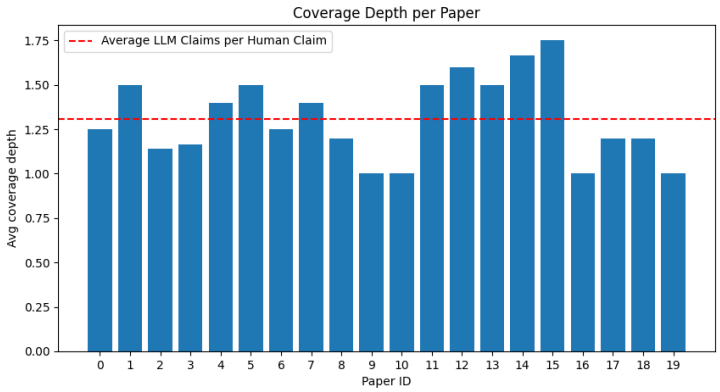


Figure 2: Average Claim Coverage per Paper

The qualitative analysis revealed a difference between how humans and the LLM interpreted scientific contributions. Human annotations tended to focus on *what* a paper contributed, combining the proposed method with its effect in a single claim. On the other hand, the model focused more on *how* the contribution was achieved. It often separated implementation details, method components, and descriptive properties into distinct claims. As a result, the LLM-extracted claim sets were more granular than the human-annotated ones.

Another important pattern was that papers with more clearly structured contribution lists near the end of the introduction produced lower redundancy. When contributions were written into the introduction text instead of as a bullet list at the end, the model was more likely to split them into overlapping claims. This suggests that explicit contribution statements improve the consistency of automated claim extraction.

Overall, the results show that the LLM can very reliably identify the primary claims presented in research papers. It achieved almost full coverage of human-identified claims. The main limitation was over-segmentation rather than claim omission. However, for the purpose of automated peer review, this may be preferable, even though it would introduce additional processing requirements.

5.3 Hallucination Analysis

The second sub-question examines whether the LLM preserves the original semantic meaning of claims without introducing hallucinations. To evaluate this, each extracted claim was manually inspected and assigned to one of three categories: correct extraction, wrong location, or misclassification. Figure 3 summarizes the results.

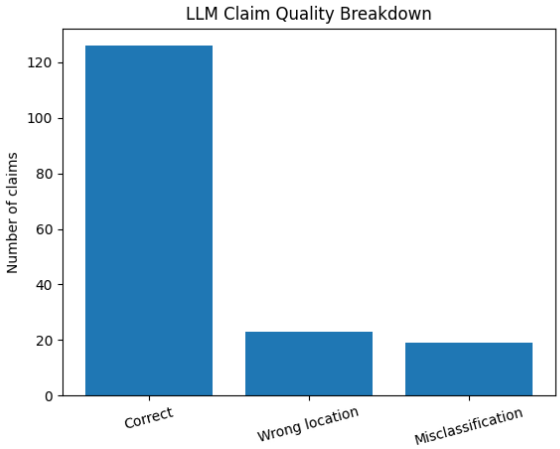


Figure 3: Claim Error Classification Analysis

Of the 168 LLM-extracted claims, 126 (75.0%) were classified as "correct". These claims accurately reflected the statements made by authors in the abstract or introduction and preserved their original meaning. Commonly, the model followed the original text quite closely, even sometimes quoting some of the original wording. Those claims were also labeled correctly as either a contribution, property, result, or comparison. Another 23 claims (13.7%) were classified as a "wrong location" error. These were not hallucinations as per the definition, as the extracted information was present in other later sections of the paper. However, the model retrieved statements from sections such as Methods, Experiments, or Results, thus not following the provided definition of a claim in the prompt. One of these cases can be considered a partial hallucination, as the model explicitly stated that the extracted sentence was from the Introduction, when in reality it appeared in the Results section. The remaining 19 claims (11.3%) were classified as misclassifications. The extracted text was correctly from the location the model claimed, but they also did not satisfy the definition of a scientific claim. Instead, the model extracted background information, descriptions or definitions, despite those being in the exclusion criteria of the prompt. These errors correspond to failures in claim identification and not hallucinated content.

Overall, the results indicate that hallucination was not a significant issue during the automated claim extraction. The model consistently used the paper for its output. Most errors were due to violations of exclusion criteria or incorrect application of the definition of a claim. This suggests that the main challenge is not factual accuracy but accurate prompt adherence.

5.4 Checklist Alignment with Human Judgment

The third sub-question investigates whether the LLM's checklist annotations align with the human annotators' judgment. To evaluate this, the checklist responses generated by the model were compared to the manually-assigned labels. Agreement was measured both with accuracy and Cohen's Kappa. While accuracy shows the proportion of matching labels, Cohen's Kappa accounts for agreement that may occur by chance. Table 2 presents the per-question results and Figure 4 visualizes them.

Table 2: Per-question Checklist Alignment Results. Entries marked with * indicate cases where Cohen's Kappa is affected by the Kappa paradox [6] due to highly imbalanced label distributions.

Question	**Accuracy**	**Cohen's Kappa**
Claims	0.99	*0.00
Limitations	0.92	0.21
Theory Assumptions and Proofs	0.95	0.90
Experimental Result Reproducibility	0.95	0.64
Open Access to Data and Code	0.81	0.57
Experimental Setting Details	0.99	0.93
Experiment Statistical Significance	0.79	0.64
Experiments Compute Resources	0.85	0.60
Code of Ethics	0.94	*0.00
Broader Impacts	0.60	0.35
Safeguards	0.85	0.38
Licenses for Existing Assets	0.40	0.12
New Assets	0.68	0.42
Crowdsourcing and Human Research	0.96	0.71
IRB Approval	0.98	0.74
Declaration of LLM Usage	0.94	0.83
Mean	**0.85**	**0.58**

Across all questions, the model achieved a mean accuracy of 0.85 and a mean Cohen's Kappa of 0.58. Several questions achieved stronger performance, including *Theory Assumptions and Proofs*, *Experimental Setting Details*, and *IRB Approval*. This is due to the fact that items such as those tend to require explicit statements in the paper, making them easier for the LLM to detect.

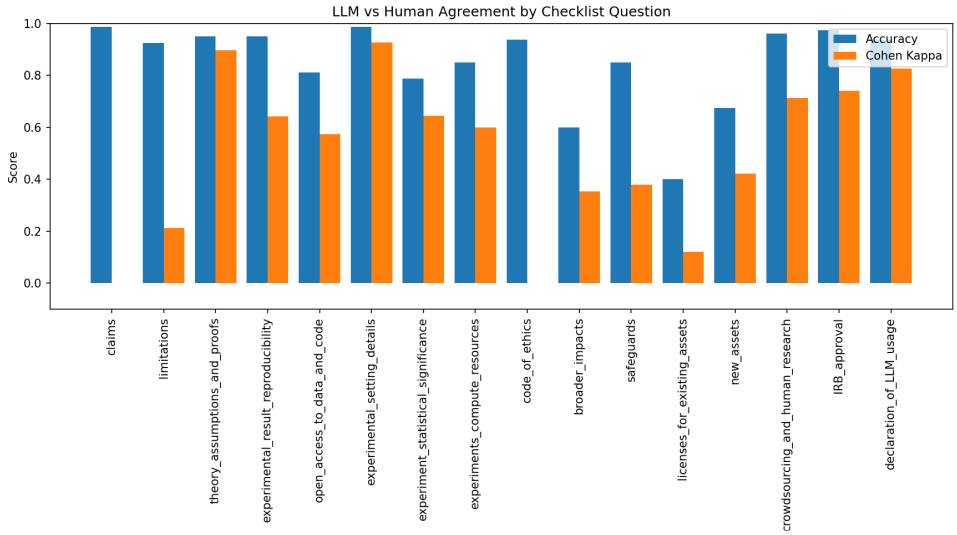


Figure 4: LLM vs Human Agreement on Checklist Question Items

Two checklist questions, *Claims* and *Code of Ethics*, produced Kappa values of 0.00 despite having high accuracies. These values were excluded from the mean Kappa calculation because they did not exceed the threshold $|\kappa| \leq 0.05$. This occurred due to the Kappa Paradox, in which agreement appears low despite nearly perfect classification performance. Cohen's Kappa is defined as: $\kappa = \frac{P_o - P_e}{1 - P_e}$ where P_o is the observed agreement and P_e is the agreement expected due to chance. When almost all observations are of the same category, P_o and P_e both approach 1. As a result, Kappa tends to zero even if the accuracy remains high. This is especially common in highly imbalanced datasets [6]. In our study, the dataset consists only of accepted and published NeurIPS papers, meaning that nearly all papers should satisfy the checklist requirements relating to supported claims and adherence to the code of ethics. Therefore, nearly all labels for these questions should be assigned as "Yes" by both the human and LLM annotators. This explains the occurrence of the paradox.

To better see all disagreements, Figure 5 shows agreement on the paper level. Each cell indicates whether the LLM and human annotator agreed on a question for a given paper. An agreement of 1.0 means that both the human and the LLM evaluator gave the same answer to the checklist item, while an agreement of 0.0 indicates that their answers differed.

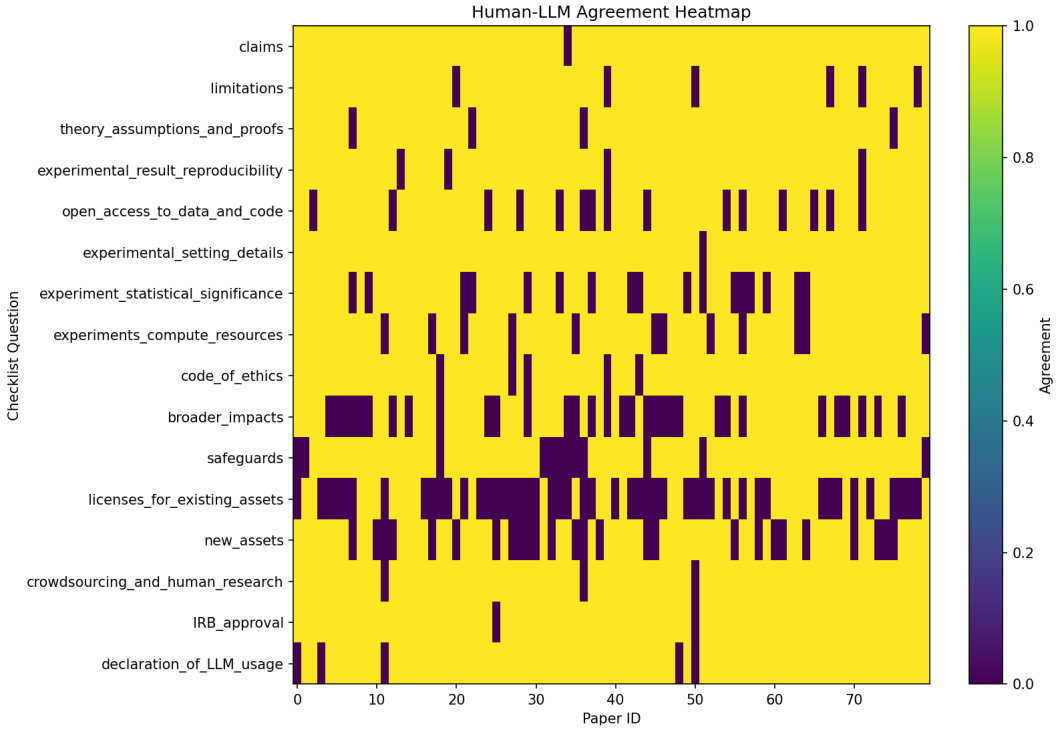
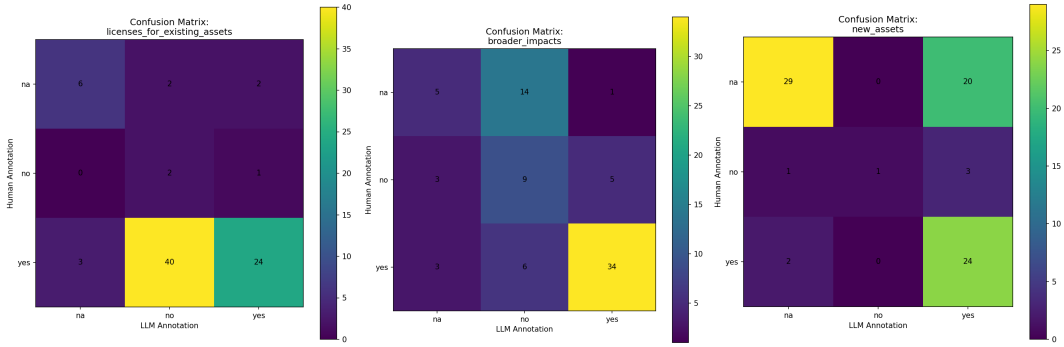


Figure 5: Human-LLM Agreement Heatmap

The heatmap shows that weakest agreement between human and LLM annotators was observed for *Licenses for Existing Assets*, *Broader Impacts*, and *New Assets*. In contrast to some other questions, these require more interpretation of the content or the authors' intent. To investigate these questions further, confusion matrices were generated for each of them in Figure 6. Across the three questions, the confusion matrices reveal specific disagreement patterns rather than random classification errors, indicating differences in

question interpretation between human and LLM annotators.

After a manual qualitative evaluation, some causes for these disagreements were identified. For the *Licenses for Existing Assets* question, the largest source of disagreement stems from differing interpretations of what sufficient licensing information is. The LLM generally required an explicit declaration, such as a Creative Commons license. Human annotators were more lenient and accepted a citation to the original paper introducing the asset. As a result, the LLM assigned the label "No" more frequently. Disagreements for the *Broader Impacts* were based on different expectations on whether such a section is necessary and how detailed it should be. The LLM expected the section to be included for most papers but accepted relatively brief discussions. On the other hand, humans followed the guidelines more closely and expected consideration of both positive and negative impacts, while requiring the section for overall less papers. For the question relating to *New Assets*, the LLM marked the item as a "Yes" if any repository link at all was present within the paper. Human annotations separated whether an asset is released or even necessary better. They were able to distinguish if a repository contained actual released assets. Since the evaluation used only the main paper PDF and not the supplementary materials, the model was unable to verify the contents of linked repositories directly.



(a) Licenses for Existing Assets (b) Broader Impacts (c) New Assets

Figure 6: Confusion Matrices for the Three Checklist Questions

Overall, the results show that the LLM achieves moderate agreement with human judgment but has a high accuracy compared to the established ground truth. Disagreements mainly occur in a small number of criteria that require further contextual interpretation or verification. This suggests that the proposed approach is able to reproduce most objective human checklist decisions, but context-dependent questions are less easily automated.

6 Responsible Research

This study performs an automated analysis of scientific papers using a large language model. While the methodology is fully automated, several ethical and responsible research considerations must be addressed, particularly regarding data usage, reproducibility, and the limitations of LLM-based evaluation.

All papers used in this study originate from the publicly available NeurIPS proceedings. No private or sensitive data was used. The dataset consists solely of published research papers, meaning no human subjects or personally identifiable information were involved in

the experiments. From a reproducibility perspective, all components of the pipeline are specified, including dataset selection criteria, model configuration, prompting strategy, and evaluation metrics. The implementation is deterministic where possible. All experiments rely on the Gemini 3 Flash Preview model, however full reproducibility may still be limited by potential future changes to the underlying model or API behavior. A key ethical consideration is the use of LLMs as evaluators of scientific content. While the system demonstrates strong agreement with human judgments in many cases, it cannot yet substitute expert peer review. The model may encode biases related to training data or prompt interpretation, and its decisions should therefore be interpreted as approximate assessments rather than more authoritative evaluations.

Beyond the direct contribution, this work has broader implications for the scientific publication process. Automated peer review systems could help reduce reviewer workload and fatigue by performing repetitive assessment tasks and increasing evaluation consistency. At the same time, excessive reliance on LLM-based reviews could lead to incorrect outcomes if model errors are not identified by human experts first. Furthermore, biases in the training data or model behavior may reinforce existing inequalities in academia by systematically favoring writing styles, research topics, or certain publication conventions for example. Therefore, such systems should be viewed as tools to support reviewers rather than replacements for human accountability.

Generative AI tools were used during the research and writing phases of this thesis. Gemini 3.0 Flash was used as part of the experimental methodology to perform claim extraction and NeurIPS Checklist annotation. Moreover, ChatGPT (OpenAI) was used to support the writing process by improving readability and academic style of the paper draft. It was not used to generate the research contributions. Any generated content was reviewed and verified or revised. We retain responsibility for the methodology, results, interpretations, and final content of this work.

7 Discussion

While the results indicate strong performance of the proposed annotation pipeline, several limitations must be considered when interpreting the findings.

Human-annotated claims were used as reference labels for the claim extraction evaluation, and checklist annotations made by the research team were used as ground truth for the checklist alignment task. However, these annotations are not independently verified and may themselves contain inconsistencies. In particular, the overlap between human annotations was limited to only five papers, restricting the cross-validation. A second limitation is the relatively small scale of the experiments. As 20 papers were used for claim extraction and 80 for checklist evaluation, these datasets remain limited compared to the diversity of NeurIPS submissions. As a result, the reported metrics do not capture performance across different research domains or paper structures. The pipeline is also constrained by incomplete access to submissions. The LLM uses only the main paper PDF and does not have access to supplementary materials. This may lead to incorrect or partial interpretations, particularly for checklist items related to datasets or released assets. Similarly, there is no accurate mechanism to determine whether authors used LLM assistance in writing but did not disclose it, which introduces potential ambiguity in the evaluation of LLM usage disclosure. From a technical perspective, the experiments relied on the free tier of the Gemini API. This introduced practical constraints, including daily rate limits and occasional unavailability. As a result, the full experimental pipeline had to be executed over multiple

days, increasing runtime variability and reducing efficiency. Although model settings were kept deterministic where possible, external API variability may still have influenced the outputs. Moreover, all experiments were conducted using a single LLM. Therefore, it remains unclear to what extent the observed performance generalizes to other models that may have different training data or architecture.

Despite these limitations, the study provides a controlled and reproducible evaluation of LLM-based scientific claim extraction and checklist annotation.

8 Conclusions and Future Work

This study investigated whether large language models can be used to support scientific peer review by extracting claims from research papers and completing NeurIPS checklist evaluations. Three research questions were addressed: (i) whether an LLM can accurately extract primary scientific claims from abstracts and introductions, (ii) whether these extracted claims remain semantically faithful without hallucination, and (iii) whether LLM-generated checklist answers align with human judgment.

The results show that LLMs are highly effective at identifying relevant scientific claims, achieving almost complete coverage of human-identified claims. At the same time, the LLM-extracted claims were often more granular than the human annotations, decomposing broader claims into more specific statements. Errors primarily arise not from hallucination, but from prompt misinterpretation or misclassification of non-claim text. For checklist evaluation, the model achieves on average moderate agreement with human annotators, with strong performance on objective checklist items. However, performance decreases for more subjective criteria that require interpretation or external context. Disagreement analysis shows that these differences are often caused by differing interpretation standards or more ambiguous guidelines of the NeurIPS Checklist.

Several directions for future research arise from this work. First, scaling the evaluation to larger and more diverse datasets, as well as multiple LLMs, would improve the robustness and reproducibility of the findings across different scientific domains. Second, incorporating expert human evaluators when establishing the ground-truth would help improve the accuracy of the experiments. Third, separating the ground truth claims into truly atomic statements could improve the accuracy of the results. Future work could also include a more structured evaluation of prompt design, as preliminary tests showed that small prompt variations can negatively affect the extraction performance. Finally, extending the system to include access to supplementary materials could improve performance on certain checklist items. This would allow for a more complete evaluation and further bridge the gap between automated and human-level peer review.

Overall, the findings suggest that LLMs have the potential to serve as useful tools for automated scientific review, particularly for objective and well-defined checklist criteria. However, they are not yet reliable substitutes for human reviewers.

A Prompts

A.1 Claim Extraction

```
You are an expert peer-reviewer for the NeurIPS conference. You are
provided with a PDF of a research paper from the 2024 Main Conference
Track. Please state the distinct semantic primary claims of the
research paper.

Please include: statements of what the paper proposes, statements about
the properties or capabilities of any proposed models, statements about
experimental results or performance, or statements comparing the
proposed method against any prior work. Merge semantically identical or
overlapping statements into a single claim (e.g.: if the same idea
appears in the Abstract and Introduction, include it only once).
Please do not include: background or motivation statements regarding
the research field, problem formulations, definitions, direct quotes,
references or citations, unnecessary details, or future work
suggestions.

A claim is defined as a statement in the abstract or introduction of a
paper that asserts a contribution, property, result, or comparison that
authors present as a finding or achievement of the work.
For each claim, you must identify its ID (a sequential integer starting
from 0), the textual statement it makes, its type (one of contribution,
property, result, comparison), and its location in the paper (the title
of the section you found it in).

Return the output in the following format:
    {
        "paper_title": "",
        "claims": [
            {
                "claim_id": ,
                "claim_text": "",
                "claim_type": "",
                "claim_location": ""
            }
        ]
    }
You MUST return only the valid JSON object with no additional text,
explanations, or formatting.
```

A.2 Checklist Evaluation

```
You are an expert peer-reviewer for the NeurIPS conference. You are
provided with a PDF of a research paper from the 2024 Main Conference
Track. Based only on information explicitly in the provided PDF, please
fill in the NeurIPS Paper Checklist.
```

Please read the checklist guidelines carefully for information on how to answer these questions. For each question in the checklist:

- You should answer "yes", "no", or "na" in lowercase.
- "na" means either that the question is Not Applicable for that particular paper or the relevant information is Not Available.
- Please provide a short (1-2 sentence) justification right after your answer (even for "na"). The justification must reference a specific section, figure, table, or direct quote from the paper as evidence. All supporting evidence can appear either in the main paper or the supplemental material, provided in the appendix.

1. Claims

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

   Guidelines:

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

2. Limitations

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

   Guidelines:

   - The answer "na" means that the paper has no limitation while the answer "no" means that the paper has limitations, but those are not discussed in the paper.
   - The authors are encouraged to create a separate "Limitations" section in their paper.
   - The paper should point out any strong assumptions and how robust the results are to violations of these assumptions (e.g., independence assumptions, noiseless settings, model well-specification, asymptotic approximations only holding locally). The authors should reflect on how these assumptions might be violated in practice and what the implications would be.
   - The authors should reflect on the scope of the claims made, e.g., if the approach was only tested on a few datasets or with a few runs. In general, empirical results often depend on implicit assumptions, which should be articulated.

- The authors should reflect on the factors that influence the performance of the approach. For example, a facial recognition algorithm may perform poorly when image resolution is low or images are taken in low lighting. Or a speech-to-text system might not be used reliably to provide closed captions for online lectures because it fails to handle technical jargon.
- The authors should discuss the computational efficiency of the proposed algorithms and how they scale with dataset size.
- If applicable, the authors should discuss possible limitations of their approach to address problems of privacy and fairness.
- While the authors might fear that complete honesty about limitations might be used by reviewers as grounds for rejection, a worse outcome might be that reviewers discover limitations that aren't acknowledged in the paper. The authors should use their best judgment and recognize that individual actions in favor of transparency play an important role in developing norms that preserve the integrity of the community. Reviewers will be specifically instructed to not penalize honesty concerning limitations.

3. Theory Assumptions and Proofs
   Question: For each theoretical result, does the paper provide the full set of assumptions and a complete (and correct) proof?
   Guidelines:
   - The answer "na" means that the paper does not include theoretical results.
   - All the theorems, formulas, and proofs in the paper should be numbered and crossreferenced.
   - All assumptions should be clearly stated or referenced in the statement of any theorems.
   - The proofs can either appear in the main paper or the supplemental material, but if they appear in the supplemental material, the authors are encouraged to provide a short proof sketch to provide intuition.
   - Inversely, any informal proof provided in the core of the paper should be complemented by formal proofs provided in appendix or supplemental material.
   - Theorems and Lemmas that the proof relies upon should be properly referenced.

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

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

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

Guidelines:

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

6. Experimental Setting/Details

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

   Guidelines:

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

7. Experiment Statistical Significance

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

   Guidelines:

   - The answer "na" means that the paper does not include experiments.

- The authors should answer "Yes" if the results are accompanied by error bars, confidence intervals, or statistical significance tests, at least for the experiments that support the main claims of the paper.
- The factors of variability that the error bars are capturing should be clearly stated (for example, train/test split, initialization, random drawing of some parameter, or overall run with given experimental conditions).
- The method for calculating the error bars should be explained (closed form formula, call to a library function, bootstrap, etc.)
- The assumptions made should be given (e.g., Normally distributed errors).
- It should be clear whether the error bar is the standard deviation or the standard error of the mean.
- It is OK to report 1-sigma error bars, but one should state it. The authors should preferably report a 2-sigma error bar than state that they have a 96% CI, if the hypothesis of Normality of errors is not verified.
- For asymmetric distributions, the authors should be careful not to show in tables or figures symmetric error bars that would yield results that are out of range (e.g. negative error rates).
- If error bars are reported in tables or plots, The authors should explain in the text how they were calculated and reference the corresponding figures or tables in the text.

8. Experiments Compute Resources

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

   Guidelines:

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

9. Code Of Ethics

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

Guidelines:

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

10. Broader Impacts

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

Guidelines:

- The answer "na" means that there is no societal impact of the work performed.
- If the authors answer "na" or "no", they should explain why their work has no societal impact or why the paper does not address societal impact.
- Examples of negative societal impacts include potential malicious or unintended uses (e.g., disinformation, generating fake profiles, surveillance), fairness considerations (e.g., deployment of technologies that could make decisions that unfairly impact specific groups), privacy considerations, and security considerations.
- The conference expects that many papers will be foundational research and not tied to particular applications, let alone deployments. However, if there is a direct path to any negative applications, the authors should point it out. For example, it is legitimate to point out that an improvement in the quality of generative models could be used to generate deepfakes for disinformation. On the other hand, it is not needed to point out that a generic algorithm for optimizing neural networks could enable people to train models that generate Deepfakes faster.
- The authors should consider possible harms that could arise when the technology is being used as intended and functioning correctly, harms that could arise when the technology is being used as intended but gives incorrect results, and harms following from (intentional or unintentional) misuse of the technology.
- If there are negative societal impacts, the authors could also discuss possible mitigation strategies (e.g., gated release of models, providing defenses in addition to attacks, mechanisms for monitoring misuse, mechanisms to monitor how a system learns from feedback over time, improving the efficiency and accessibility of ML).

11. Safeguards

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

Guidelines:

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

12. Licenses for existing assets

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

Guidelines:

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

13. New Assets

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

14. Crowdsourcing and Research with Human Subjects
Question: For crowdsourcing experiments and research with human subjects, does the paper include the full text of instructions given to participants and screenshots, if applicable, as well as details about compensation (if any)?
Guidelines:
- The answer "na" means that the paper does not involve crowdsourcing nor research with human subjects.
- Including this information in the supplemental material is fine, but if the main contribution of the paper involves human subjects, then as much detail as possible should be included in the main paper.
- According to the NeurIPS Code of Ethics, workers involved in data collection, curation, or other labor should be paid at least the minimum wage in the country of the data collector.

15. Institutional Review Board (IRB) Approvals or Equivalent for Research with Human Subjects
Question: Does the paper describe potential risks incurred by study participants, whether such risks were disclosed to the subjects, and whether Institutional Review Board (IRB) approvals (or an equivalent approval/review based on the requirements of your country or institution) were obtained?
Guidelines:
- The answer "na" means that the paper does not involve crowdsourcing nor research with human subjects.
- Depending on the country in which research is conducted, IRB approval (or equivalent) may be required for any human subjects research. If you obtained IRB approval, you should clearly state this in the paper.
- We recognize that the procedures for this may vary significantly between institutions and locations, and we expect authors to adhere to the NeurIPS Code of Ethics and the guidelines for their institution.

```
        - For initial submissions, do not include any information that
        would break anonymity (if applicable), such as the institution
        conducting the review.

16. Declaration of LLM usage
    Question: Does the paper describe the usage of LLMs if it is an
    important, original, or non-standard component of the core methods
    in this research? Note that if the LLM is used only for writing,
    editing, or formatting purposes and does not impact the core
    methodology, scientific rigorousness, or originality of the
    research, declaration is not required.
    Guidelines:
        - The answer NA means that the core method development in this
        research does not involve LLMs as any important, original, or
        non-standard components.
        - Please refer to our MainTrackHandbook for what should or
        should not be described.

Return the output in the following format:
    {
        "paper_title": "",
        "claims": "",
        "claims_justification": "",
        "limitations": "",
        "limitation_justification": "",
        "theory_assumptions_and_proofs": "",
        "theory_justification": "",
        "experimental_result_reproducibility": "",
        "reproducibility_justification": "",
        "open_access_to_data_and_code": "",
        "access_justification": "",
        "experimental_setting_details": "",
        "settings_justification": "",
        "experiment_statistical_significance": "",
        "significance_justification": "",
        "experiments_compute_resources": "",
        "resources_justification": "",
        "code_of_ethics": "",
        "ethics_justification": "",
        "broader_impacts": "",
        "impact_justification": "",
        "safeguards": "",
        "safeguards_justification": "",
        "licenses_for_existing_assets": "",
        "licenses_justification": "",
        "new_assets": "",
        "assets_justification": "",
        "crowdsourcing_and_human_research": "",
        "crowdsourcing_justification": "",
```

```
        "IRB_approval": "",
        "IRB_justification": "",
        "declaration_of_LLM_usage": "",
        "LLM_justification": ""
    }
Return only the JSON object with no additional text, explanations, or
formatting.
```

References

[1] Artificial Analysis. Gemini 3 flash preview (reasoning) vs. llama 3.3 instruct 70b. https://artificialanalysis.ai/models/comparisons/gemini-3-flash-reasoning-vs-llama-3-3-instruct-70b, 2026.

[2] Maitreya Prafulla Chitale, Ketaki Mangesh Shetye, Harshit Gupta, Manav Chaudhary, Manish Shrivastava, and Vasudeva Varma. Autorev: Automatic peer review system for academic research papers, 2026.

[3] Cohen. A coefficient of agreement for nominal scales. *Educational and Psychological Measurement*, 20, 1960.

[4] Conference on Neural Information Processing Systems. About neurips conference. https://neurips.cc/About, 2026. Official NeurIPS conference website.

[5] Conference on Neural Information Processing Systems. Neurips paper checklist. https://neurips.cc/public/guides/PaperChecklist, 2026. Official NeurIPS author guidelines page.

[6] Bastiaan M. Derksen, Wendy Bruinsma, Johan Carel Goslings, and Niels W.L. Schep. The kappa paradox explained. *The Journal of Hand Surgery*, 49(5):482–485, 2024.

[7] Alexander Goldberg, Ihsan Ullah, Thanh Gia Hieu Khuong, Benedictus Kent Rachmat, Zhen Xu, Isabelle Guyon, and Nihar B. Shah. Usefulness of llms as an author checklist assistant for scientific papers: Neurips'24 experiment, 2024.

[8] Google Cloud. Gemini 3 flash. https://docs.cloud.google.com/gemini-enterprise-agent-platform/models/gemini/3-flash, 2026. Model documentation, accessed 15 June 2026.

[9] Jiawei Gu, Xuhui Jiang, Zhichao Shi, Hexiang Tan, Xuehao Zhai, Chengjin Xu, Wei Li, Yinghan Shen, Shengjie Ma, Honghao Liu, Saizhuo Wang, Kun Zhang, Yuanzhuo Wang, Wen Gao, Lionel Ni, and Jian Guo. A survey on llm-as-a-judge. *The Innovation*, 7(6), 2026.

[10] Shashidhar Reddy Javaji, Yupeng Cao, Haohang Li, Yangyang Yu, Nikhil Muralidhar, and Zining Zhu. Can ai validate science? benchmarking llms for accurate scientific claim → evidence reasoning, 2025.

[11] Aobo Kong, Shiwan Zhao, Hao Chen, Qicheng Li, Yong Qin, Ruiqi Sun, Xin Zhou, Enzhi Wang, and Xiaohang Dong. Better zero-shot reasoning with role-play prompting, 2024.

[12] Giuseppe Russo Latona, Manoel Horta Ribeiro, Tim R. Davidson, Veniamin Veselovsky, and Robert West. The ai review lottery: Widespread ai-assisted peer reviews boost paper scores and acceptance rates, 2024.

[13] Rui Li, Jia-Chen Gu, Po-Nien Kung, Heming Xia, Junfeng liu, Xiangwen Kong, Zhifang Sui, and Nanyun Peng. Llm-reval: Can we trust llm reviewers yet?, 2025.

[14] Weixin Liang, Zachary Izzo, Yaohui Zhang, Haley Lepp, Hancheng Cao, Xuandong Zhao, Lingjiao Chen, Haotian Ye, Sheng Liu, Zhi Huang, Daniel A. McFarland, and James Y. Zou. Monitoring ai-modified content at scale: A case study on the impact of chatgpt on ai conference peer reviews, 2026.

[15] Pengfei Liu, Weizhe Yuan, Jinlan Fu, Zhengbao Jiang, Hiroaki Hayashi, and Graham Neubig. Pre-train, prompt, and predict: A systematic survey of prompting methods in natural language processing, 2021.

[16] Ryan Liu and Nihar B. Shah. Reviewergpt? an exploratory study on using large language models for paper reviewing, 2023.

[17] Zhengda Mo, Zhiyu Quan, Eli O'Donohue, and Kaiwen Zhong. Claim automation using large language model, 2026.

[18] Jiefu Ou, William Gantt Walden, Kate Sanders, Zhengping Jiang, Kaiser Sun, Jeffrey Cheng, William Jurayj, Miriam Wanner, Shaobo Liang, Candice Morgan, Seunghoon Han, Weiqi Wang, Chandler May, Hannah Recknor, Daniel Khashabi, and Benjamin Van Durme. Claimcheck: How grounded are llm critiques of scientific papers?, 2025.

[19] Dina Pisarevskaya and Arkaitz Zubiaga. Agent-based automated claim matching with instruction-following llms, 2025.

[20] Shaghayegh Sadeghi, Khashayar Khajavi, Rise Adhikari, and Alexander Tessier. Deepsciverify: Verifying scientific claim–citation alignment via llm-driven evidence escalation, 2026.

[21] Pawin Taechoyotin and Daniel Acuna. Remor: Automated peer review generation with llm reasoning and multi-objective reinforcement learning, 2025.