Large Language Models for Research Review

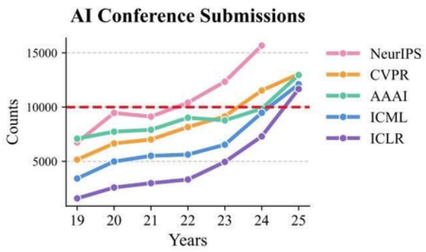
Evaluating Claim-Level Completeness in Machine Learning Research

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?

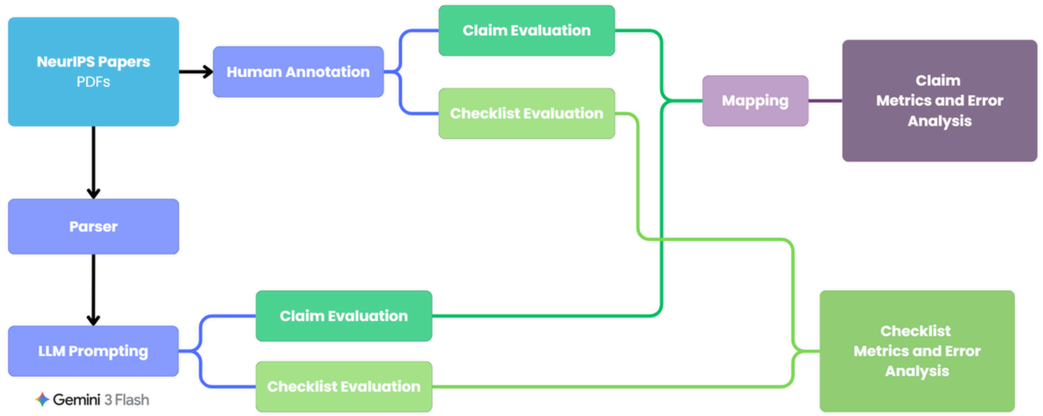
1. Can the LLM accurately extract the primary claims as stated in the abstract and introduction?
2. How well does the LLM's summary of a claim maintain the original semantic meaning without introducing "hallucinations"?
3. Does the LLM's binary "Yes/No" response to the NeurIPS Checklist item align with human judgment?

1 Introduction

- Rapid growth of ML publications → reviewer fatigue
- Consistent peer review is increasingly difficult
- Checklist to assess research quality and transparency
- Claims are the core scientific contributions of a paper
- Claim extraction is a prerequisite for automated review

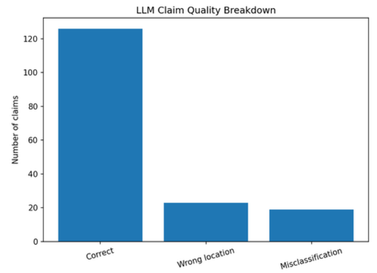


2 Methodology

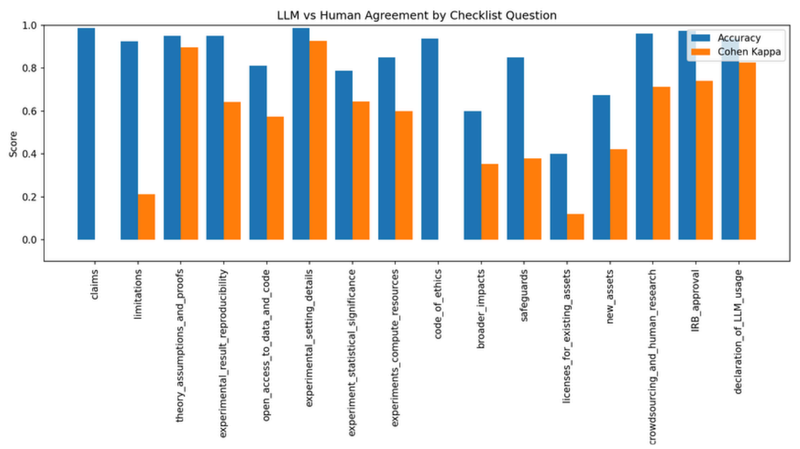


3 Results

Total number of claims by human:	98
Total number of claims by LLM:	168
Covered human claims:	97
Human claim coverage (Recall):	0.99
Correct LLM claims (Precision):	0.75
Redundancy:	1.71
Average number of LLM claims per human claim:	1.31



The LLM is highly effective at identifying relevant scientific claims, achieving almost complete coverage of human-identified claims.
The LLM-extracted claim sets were more granular than the human-annotated ones.



The LLM achieves moderate agreement with human judgment but has a high accuracy compared to the established ground truth.
Disagreements mainly occur in context-dependent items.

4 Discussion

Limitations:

- Small dataset
- No cross-validation for human ground truth
- No access to supplementary material
- API variability limits reproducibility

Societal Impacts:

- Reduced reviewer fatigue and faster scientific evaluation
- More consistent evaluation
- Over-reliance risk leading to incorrect evaluations
- Risk of reinforcing existing inequalities in academics

5 Conclusion & Future Work

Conclusion:

- **RQ1:** The model can reliably identify the main scientific contributions of research papers.
- **RQ2:** Most extracted claims were correct and the majority of errors resulted from misclassification or prompt misinterpretation.
- **RQ3:** The LLM's checklist responses align with human ones moderately well. Agreement is weaker for subjective or context-dependent criteria.

Future Work:

- Scale the evaluation and introduce expert evaluators
- Create more atomic ground-truth claims
- Conduct a structured study of prompt design

Simona Simeonova[1]
S.I.Simeonova@student.tudelft.nl

Supervisors: David Tax,[1] Chenxu Hao[1]
[1]EEMCS, TU Delft, Netherlands

References:
[1] Jaeho Kim, Yunseok Lee, and Seulki Lee. Position: The AI Conference Peer Review Crisis Demands Author Feedback and Reviewer Rewards. In Proceedings of the Forty-second International Conference on Machine Learning, Position Paper Track, 2025.