AUTHORS

Adam Wielinga
Student Researcher
CSE3000

Evaluating the Ability of Large Language Models to Classify Scientific Papers as Empirical or Theoretical using the NeurIPS Checklist

TU Delft

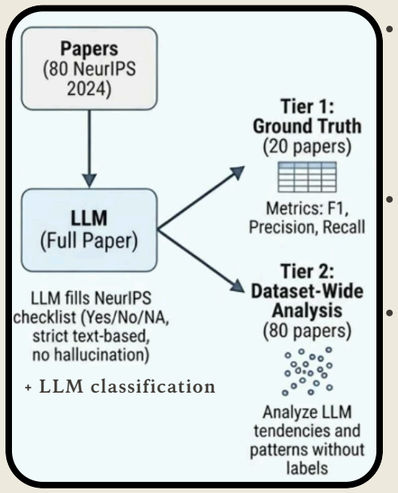
01. Introduction

- Manual evaluation is difficult due to rapid research growth
- LLM use in peer revieuw is an active area
- Differentiating empirical vs theoretical papers is critical in proper evauluation

02. Research questions

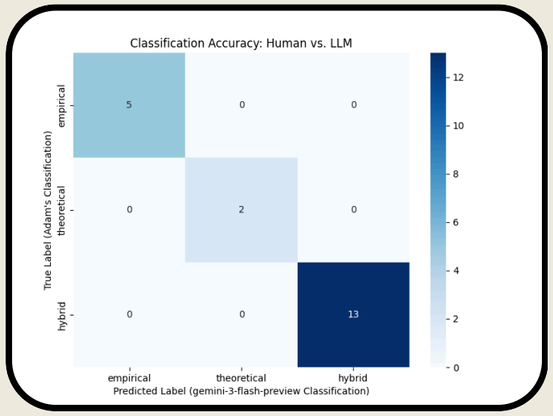
- How accurately does an LLM classify papers compared to human annotation
- How well does the LLM's generated reasoning and granular checklist extraction align with human-annotated rationale?
- How consistent are human annotations when annotating papers

03. Methodology

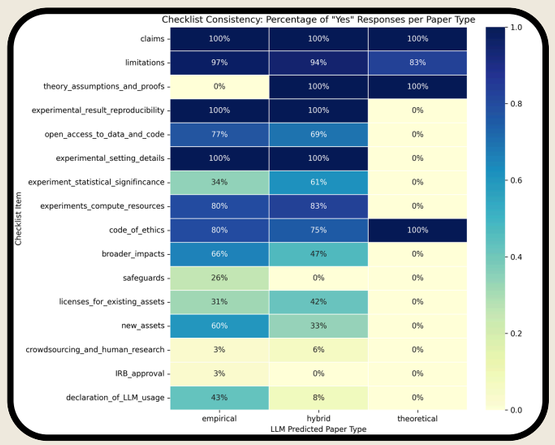


- **Labeling papers:**
  - Checklist
  - Labeling
    - empirical
    - theoretical
    - hybrid
- **LLM Usage**
  - Prompt
  - Checklist
  - Classification
- **Evaluation**
  - **Two-Tier**
    - Ground truth papers
    - Dataset-Wide Analysis
  - Inter-annotator agreement
    - Cohen's Kappa

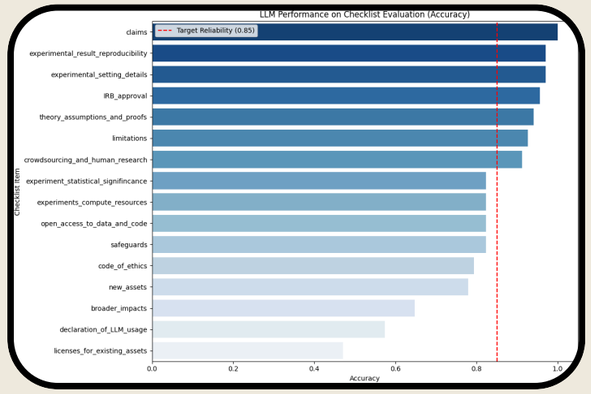
04. Results



- The LLM achieved 100% classification accuracy for the ground truth subset



- The distribution chart shows the classifications are determined by logical feature groupings.
- The LLM maps its final methodological judgment directly to the presence or absence of these key technical indicators



- The LLM achieved high accuracy (92-100%) for questions with clear structural indicators
- However on questions that did not contain clear dedicated sections in the paper, it scored a lot lower, indicating limited contextual reasoning

- Overal human consensus remained strong.
- Slight disagreement was present in some checklist items which helps to give context to the inherent ambiguity of some checklist items

05. Discussion

- **Cohens Kappa:**
  - The ambiguity inherent to some of the checklist items only justifies part of the LLM's lower accuracy.
- **Limitations**
  - Limited to 20 ground truth labels
  - Limited to using free tier LLM api which causes lesser comprehensive ability
  - Certain CheckList items in supplemental material, unvisible to AI

06. Conclusion

- We establish that LLMs provide a highly consistent baseline for classifying paper typologies and extracting explicit methodological data,
- Yet, their inherent dependence on structural indicators and their limited contextual reasoning restrict their ability to be used independent.