## AUTHOR

Alexia-Elena Predoi


# ANALYSIS OF RESULTS IN THE ML RESEARCH FIELD

Investigating the Efficacy of LLMs in Extracting Research Limitations


## SUPERVISORS

David M.J. Tax
Chenxu Hao
Hayley Hung
Nergis Tömen


## INTRODUCTION

- **Core Need**: Declaring limitations is vital for scientific integrity
- **The Problem**: Reviewers must manually search dense, complex papers to isolate self-acknowledged system weaknesses
- **The Gap**: It is unknown if LLMs can deeply comprehend context or if they rely on shallow keyword matching

## OBJECTIVE

- **Goal**: Evaluate LLMs on targeted limitation extraction
- **Impact**: Quickly surfaces a paper's true weaknesses to save reviewers time

## RELATED WORK

- **Reasoning Gaps**: CLAIM-BENCH reveals LLMs often fail to link claims in the introduction to evidence in the results [1]
- **Checklist Utility**: CMU/NeurIPS pilots show that AI assistants are prone to "superficial agreement" and can be "gamed" by authors [2]
- **Pipeline Necessity**: Research confirms single-pass prompting is insufficient; structured pipelines (like the one proposed) are required for scientific accuracy [1], [3]

## RESEARCH QUESTIONS

1. **Explicit Detection**: Can the LLM detect a dedicated limitations section and precisely extract the core limitation phrases?
2. **Implicit Detection**: In the absence of a dedicated section, can the LLM locate limitation phrases throughout the full text?
3. **Contextualization**: Does the LLM accurately explain why an identified factor constitutes a structural limitation?

## METHODOLOGY

- **Data**: 78 NeurIPS papers
- **Ground Truth**: Manual limitations extraction for all 78 papers
- **Pipeline**: Python-based LLM processing of 78 papers
- **Prompt design strategy**:
  - *What Works*: Persona alignment ("expert peer-reviewer"), enforcing verbatim string extractions, and explicit lexical triggers (e.g. "Our method is limited by...")
  - *What Fails*: Allowing the model to "infer" or guess a weakness, and broad context windows that leak future work

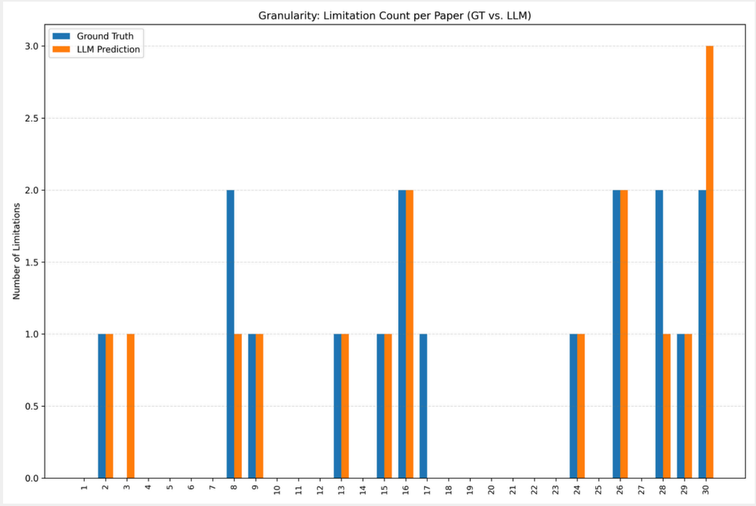
## RESULTS

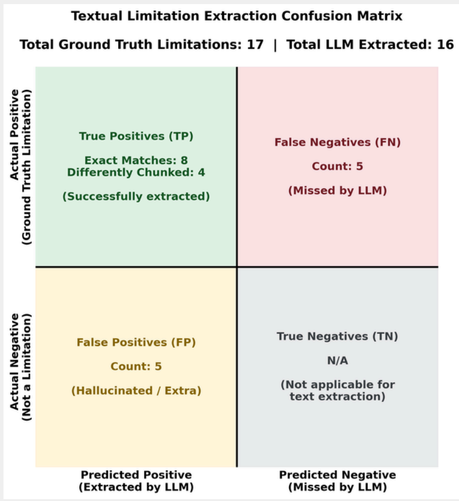
**RQ1: Explicit Auditing (Dedicated Section)**

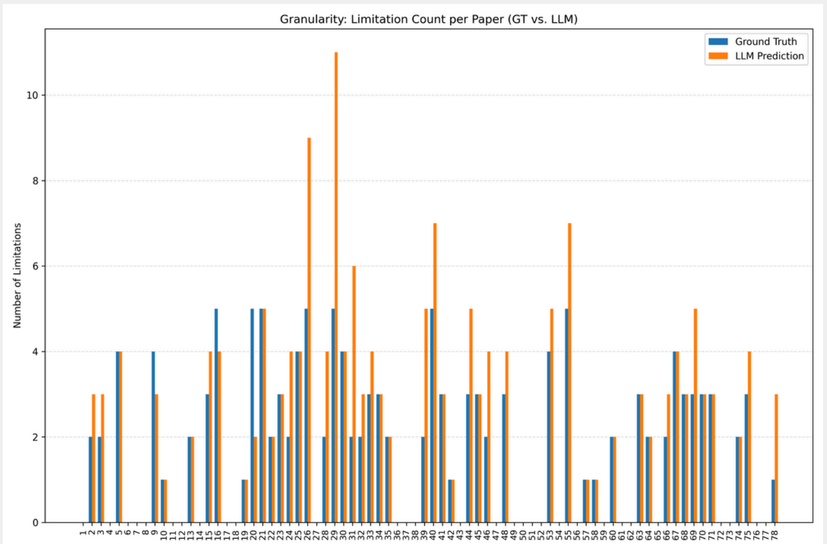
- **Section Detection**: 100% Accuracy
- **Text Extraction**: 0.89 Recall and 0.68 Precision
- **The Catch**: Highly effective at catching limitations, but prone to over-extracting (confuses future work with actual limitations)
- **Chunking Issue**: The model frequently splits single paragraphs into multiple bullets

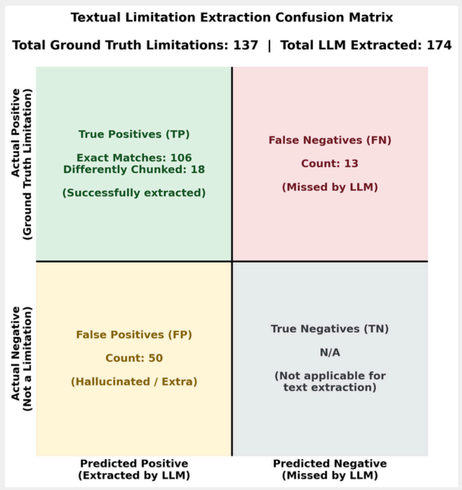
**RQ2: Implicit Auditing (No Dedicated Section)**

- **Text Extraction**: Recall dropped to 0.71 and Precision to 0.69
- **Omission problem**: The model struggled with critical omissions when limitations were buried deep within unstructured text profiles (e.g. Appendices)




## CONCLUSION

- **Takeaway**: LLMs act as great assistants for human reviewers by navigating layouts and surfacing explicit text
- **The "Chunking Problem"**: Strict word-for-word scoring fails when the LLM splits or merges text blocks
- **Future Work**: Shift to semantic metrics (partial credit), implement Chain-of-Thought reasoning to cut false positives, and scale beyond NeurIPS

## DISCUSSION

- ***Model Rate-Limiting***: Use of Gemini 3.0 Flash (Free Tier) results in periodic "Server Busy" errors and rate limits
- ***Sample Size***: Current analysis is restricted to 78 papers
- ***Subjective Ground Truth***: The annotators lacked highly specialized domain expertise
- ***Prompt Sensitivity***: Model outputs can be highly sensitive to specific wording

## REFERENCES


[1] S. R. Javaji et al., "Can AI validate science? Benchmarking LLMs on claim $\rightarrow$ evidence reasoning in AI papers," in Proc. IJCNLP-AACL, 2025, pp. 2355–2379.
[2] A. Goldberg et al., "Usefulness of LLMs as an author checklist assistant for scientific papers," CMU CSD PhD Blog, 2025.C. Oesterheld, "LLMs as author checklist assistants: NeurIPS'24 experiment," CMU CSD Blog, 2025.
[3] H. Ye et al., "How should we responsibly adopt LLMs in peer review?" in Proc. ACL, 2026.