ACM Conference on Reproducibility and Replicability (ACM REP '25)

LLMs in Citation Intent Classification: Progress, Precision, and Reproducibility Challenges

Alex Fogelson, Ana Trišović, Neil Thompson
MIT FutureTech

1. Introduction

Understanding the intent behind scientific citations is critical for advancing scholarly search, literature summarization, and the construction of knowledge graphs. Citations convey far more than acknowledgment—they indicate methodological similarities, extensions of theoretical work, critiques and criticisms, and much more. Accurately classifying these intentions enables richer machine understanding of scientific discourse and has become a key goal in the field of scientometrics and natural language processing [2].

We investigate the use of LLMs for multi-class citation intent classification amongst scientists referencing AI foundation models, focusing on the categories *context*, *uses*, and *extends*. While both fine-tuned models (e.g., SciBERT) and prompt-based LLMs show promise, recent studies highlight ongoing challenges [4, 3].

Our experiments compare open-weight models (e.g., LLaMA 3.1) with proprietary systems (e.g., GPT-4.1-mini), revealing substantial disagreement across classification outputs—even with prompt tuning and threshold calibration. Citation intent classification remains difficult, particularly in ambiguous cases.

Notably, the top-performing models are proprietary. Despite access to a large number of GPUs, open models fall short in accuracy, as well as efficiency. This dependency raises concerns about transparency, replicability, and reproducibility, as closed models often lack versioning and auditability—undermining scientific reliability.

Our contributions are as follows:

- We explore citation intent classification using multiple LLMs, both open and proprietary.
- Proprietary models outperform open models, but at the cost of transparency and replicability.
- Results reveal significant disagreement across model predictions, raising methodological concerns.

Model	F1 Score
GPT-4.1-mini	0.67
LLaMA 3.1	0.50
LLaMA 3.1 + GPT-4o	0.52

Table 1: Performance of citation classification methods

2. Methods and Results

Data. We construct our dataset by extracting citation contexts from the Semantic Scholar citation graph and the S2ORC corpus. For each cited instance of a foundation model, we collect a three-sentence window surrounding the citation. Using a 3-label classification schema—*context*, *uses*, and *extends*—we annotate and evaluate approximately 300 examples.

Approaches. We implement and compare multiple classification pipelines: (1) an open-source model (LLaMA 3.1) [1], (2) a hybrid system combining LLaMA with GPT-4o-mini [5] for reclassification, and (3) a fully proprietary method using GPT-4.1-mini. With each model set, we iterated through combinations of few-shot prompting, question-sets with thresholding, neural network classifiers on LLM outputs indicators, and combinations of these techniques. For each model, we report results from the optimal scaffolding per model.

Disagreement. Despite using state-of-the-art models, we observe limited consistency across outputs. Pairwise inter-model F1 scores average around 0.49–0.50, indicating that citation intent classification remains unstable and sensitive to model choice and scaffolding.

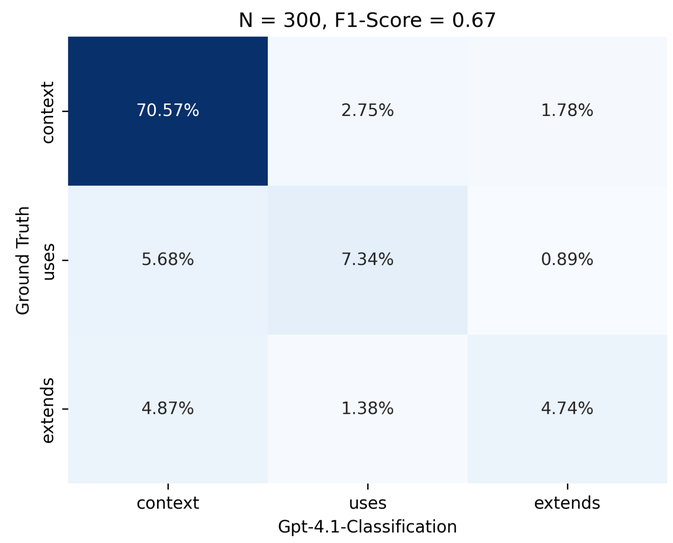

Figure 1: Confusion matrix of best-performing model using zero-shot prompting - GPT-4.1-mini vs. Ground Truth.

3. Discussion

Even the best-performing models show consistent, model-specific errors. Smaller models frequently made misattribution errors (classifying citation as model usage when referencing other authors deployment), while all systems exhibited classifier sycophancy (a tendency to generously interpret sentences toward positive classifications). None of our evaluated methods achieved satisfactory performance without incorporating proprietary models, posing threats to reproducibility due to silent updates, lack of versioning, and opaque training pipelines.

4. Conclusions

Citation intent classification remains a non-trivial challenge for LLMs, with no clear best model or approach. The reliance on proprietary systems hinders methodological transparency and reproducibility. Open-source alternatives are still lagging in performance, especially under resource constraints, but offer clearer paths toward scientific openness.

5. Research Directions

- Improve open models' classification accuracy via training on expanded citation datasets
- Investigate cross-discipline citation behaviors to refine label taxonomy
- Explore explainable AI techniques to interpret citation intent predictions

References

[1] Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Amy Yang, Angela Fan, et al. The Llama 3 Herd of Models. *arXiv preprint arXiv:2407.21783*, 2024.

[2] David Jurgens, Srijan Kumar, Raine Hoover, Dan McFarland, and Dan Jurafsky. Measuring the evolution of a scientific field through citation frames. *Transactions of the Association for Computational Linguistics*, 6:391–406, 2018.

[3] Arina Kostina, Marios D Dikaiakos, Dimosthenis Stefanidis, and George Pallis. Large language models for text classification: Case study and comprehensive review. *arXiv preprint arXiv:2501.08457*, 2025.

[4] Anne Lauscher, Brandon Ko, Bailey Kuehl, Sophie Johnson, David Jurgens, Arman Cohan, and Kyle Lo. Multicite: Modeling realistic citations requires moving beyond the single-sentence single-label setting. *arXiv preprint arXiv:2107.00414*, 2021.

[5] OpenAI. Introducing GPT-4.1 in the api, 2024. Accessed: 2025-05-27.