HOW DEEPLY DO LLMS INTERNALIZE HUMAN CITATION PRACTICES? A GRAPH-STRUCTURAL AND EMBEDDING-BASED EVALUATION

Melika Mobini
Vrije Universiteit Brussel
Melika.Mobini@vub.be

Vincent Holst
Vrije Universiteit Brussel
Vincent.Thorge.Holst@vub.be

Floriano Tori
Vrije Universiteit Brussel
Floriano.Tori@vub.be

Andres Algaba
Vrije Universiteit Brussel
Andres.Algaba@vub.be

Vincent Ginis
Vrije Universiteit Brussel
SEAS, Harvard University
Vincent.Ginis@vub.be

ABSTRACT

As Large Language Models (LLMs) integrate into scientific workflows, understanding how they conceptualize the literature becomes critical. We compare LLM-generated citation suggestions with real references from top AI conferences (AAAI, NeurIPS, ICML, ICLR), analyzing key citation graph properties—centralities, clustering coefficients, and structural differences. Using OpenAI embeddings for paper titles, we quantify the alignment of LLM-generated citations with ground truth references. Our findings reveal that LLM-generated citations closely resemble human references in these distributional properties, deviating significantly from random baselines.

1 INTRODUCTION

LLMs are reshaping scientific publishing, yet how they internalize the scientific corpus remains an open question. As AI systems become increasingly integrated into research workflows, they play a dual role: assisting with literature synthesis while potentially influencing citation practices (Li et al. (2024); Skarlinski et al. (2024)). Understanding this dynamic is essential for ensuring the integrity of scientific communication and avoiding systemic biases that might be introduced by algorithmic recommendations (Fortunato et al. (2018); Nielsen & Andersen (2021); Susnjak et al. (2024)).

Human-AI co-evolution in scientific research is increasingly evident, with AI-generated insights feeding into human decision-making and vice versa (Delgado-Chaves et al. (2025); Schmidgall et al. (2025)). A crucial aspect of this co-evolution is the feedback loop between human researchers and AI-generated outputs, where model-generated references influence scholarly work, which in turn shapes future model training (Baek et al. (2024)). By examining these interactions, we aim to provide insights into how AI-driven citation practices may evolve and what safeguards are necessary to maintain scientific rigor. Beyond citation bias, we investigate whether LLMs internalize citation structures in a meaningful way. Prior work (Algaba et al. (2024)) highlights LLM citation biases; we extend this by analyzing structural and embedding-based properties of LLM-generated citations.

2 METHODOLOGY

Following Algaba et al. (2024), we collect citations generated by GPT-4, GPT-4o, and Claude 3.5 for AAAI, NeurIPS, ICML, and ICLR papers. To ensure fair comparisons, we preprocess the dataset by filtering duplicate references and normalizing citation formats. First, we construct two citation graphs—one from LLM references and another from ground truth—and analyze structural properties such as centralities, clustering coefficients, and shortest path lengths. Second, using OpenAI's text-embedding-ada-002 model, we compare title embeddings by computing cosine similarity scores. As a baseline, we randomly reshuffle citations within the same field and apply a Random Forest classifier to assess discriminative power. We ensure that our dataset includes diverse research areas within AI, allowing us to capture broader citation trends across subfields.

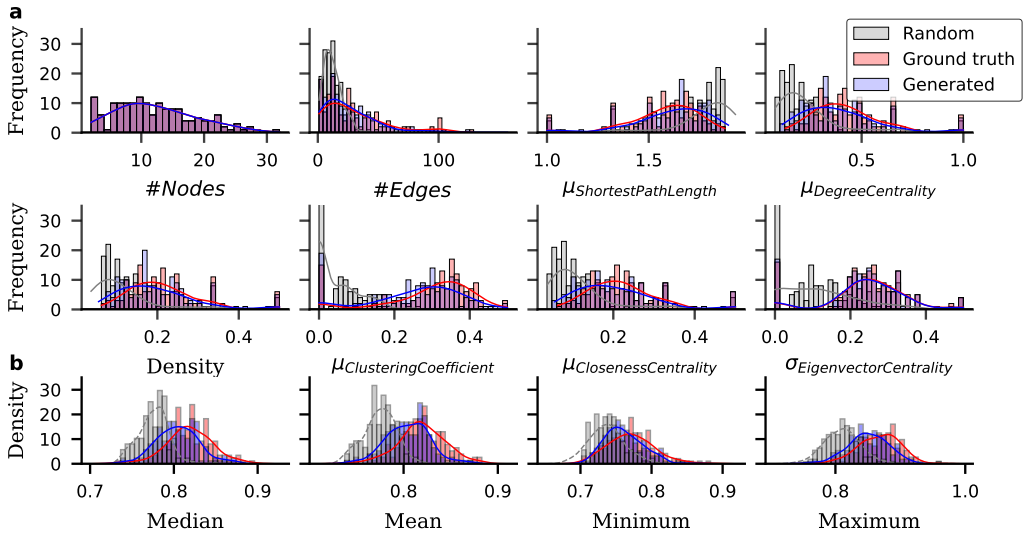

Figure 1: Graph-Structural and Embedding-Based Comparison of Citation Networks. (a) Distribution of key structural properties in citation graphs generated by LLMs (blue), ground truth human citations (red), and random references (gray). The distributions show that LLM-generated citations closely follow the structural characteristics of human citation networks, deviating significantly from the random baseline. (b) Comparison of cosine similarity distributions between focal paper titles and their references for each category. LLM-generated citations align more closely with human-generated references than random baselines, yet minor systematic differences persist. Smooth density curves highlight the overall distributional trends across the datasets.

3 Results

Our analysis (see Figure 1) reveals that LLM-generated citation graphs align closely with human citation structures. Graph metrics such as clustering coefficients and degree centralities exhibit similar distributions to ground truth references, whereas random baselines show significant divergence. Embedding comparisons further reinforce this trend, with LLM-generated references displaying high semantic similarity to human citations. To further quantify these findings, we trained a Random Forest classifier on graph properties and embeddings. The model distinguishes random baselines from human citations with high accuracy but struggles to differentiate LLM-generated references from ground truth (see Appendix Table A1). This suggests that LLMs have internalized human citation patterns both at a structural and semantic level.

4 Conclusion and Outlook

Citation analysis often focuses on whether references exist; our study probes deeper into whether LLMs reproduce meaningful citation structures. We find that LLMs generate citations that align well with human patterns, both structurally and semantically. While some systematic differences remain, the observed alignment suggests that LLMs have internalized key aspects of human citation behavior. However, reliance on LLM-generated citations could inadvertently create feedback loops, reinforcing existing citation biases and structural inequities, potentially amplifying certain research narratives while marginalizing others-highlighting the necessity for ongoing scrutiny of AI-assisted scientific workflows (Li et al., 2024; Skarlinski et al., 2024). Future work should incorporate expert qualitative evaluations of LLM-generated references, stronger baselines such as domain-specific retrieval models or citation recommender systems (Nielsen & Andersen, 2021), and further exploration of feedback loops arising from human-AI interactions in citation practices.

5 URM STATEMENT

We confirm that our lead author meets the URM criteria. Also, as a 2nd year PhD student, this is her first submission to ICLR (or similar conferences).

REFERENCES

Andres Algaba, Carmen Mazijn, Vincent Holst, Floriano Tori, Sylvia Wenmackers, and Vincent Ginis. Large language models reflect human citation patterns with a heightened citation bias. *arXiv preprint arXiv:2405.15739*, 2024.

Jinheon Baek, Sujay Kumar Jauhar, Silviu Cucerzan, and Sung Ju Hwang. Researchagent: Iterative research idea generation over scientific literature with large language models. *arXiv preprint arXiv:2404.07738*, 2024.

Fernando M. Delgado-Chaves, Matthew J. Jennings, Antonio Atalaia, Justus Wolff, Rita Horvath, Zeinab M. Mamdouh, Jan Baumbach, and Linda Baumbach. Transforming literature screening: The emerging role of large language models in systematic reviews. *Proceedings of the National Academy of Sciences*, 122(2):e2411962122, 2025. doi: 10.1073/pnas.2411962122.

Santo Fortunato, Carl T Bergstrom, Katy Börner, James A Evans, Dirk Helbing, Staša Milojević, Alexander M Petersen, Filippo Radicchi, Roberta Sinatra, Brian Uzzi, et al. Science of science. *Science*, 359(6379):eaao0185, 2018.

Rodney Kinney, Chloe Anastasiades, Russell Authur, Iz Beltagy, Jonathan Bragg, Alexandra Buraczynski, Isabel Cachola, Stefan Candra, Yoganand Chandrasekhar, Arman Cohan, et al. The semantic scholar open data platform. *arXiv preprint arXiv:2301.10140*, 2023.

Sihang Li, Jin Huang, Jiaxi Zhuang, Yaorui Shi, Xiaochen Cai, Mingjun Xu, Xiang Wang, Linfeng Zhang, Guolin Ke, and Hengxing Cai. Scilitllm: How to adapt llms for scientific literature understanding. *arXiv preprint arXiv:2408.15545*, 2024.

Mathias Wullum Nielsen and Jens Peter Andersen. Global citation inequality is on the rise. *Proceedings of the National Academy of Sciences*, 118(7):e2012208118, 2021.

Samuel Schmidgall, Yusheng Su, Ze Wang, Ximeng Sun, Jialian Wu, Xiaodong Yu, Jiang Liu, Zicheng Liu, and Emad Barsoum. Agent laboratory: Using llm agents as research assistants. *arXiv preprint arXiv:2501.04227*, 2025.

Michael D Skarlinski, Sam Cox, Jon M Laurent, James D Braza, Michaela Hinks, Michael J Hammerling, Manvitha Ponnapati, Samuel G Rodriques, and Andrew D White. Language agents achieve superhuman synthesis of scientific knowledge. *arXiv preprint arXiv:2409.13740*, 2024.

Teo Susnjak, Peter Hwang, Napoleon H Reyes, Andre LC Barczak, Timothy R McIntosh, and Surangika Ranathunga. Automating research synthesis with domain-specific large language model fine-tuning. *arXiv preprint arXiv:2404.08680*, 2024.

Erjia Yan and Ying Ding. Applying centrality measures to impact analysis: A coauthorship network analysis. *Journal of the American Society for Information Science and Technology*, 60(10):2107–2118, 2009.

APPENDIX

CITATION GENERATION PIPELINE

For our analysis, we rely on data publicly shared by Algaba et al. (2024). Here, the authors prompted GPT-4o-2024-05-13 with the author information, conference information, abstract, and introduction of 166 papers published in AI top conferences after the knowledge cut-off date, with the default system message: "*You are a helpful assistant*" and the following prompt:

> Below, we share with you a written introduction to a paper and have omitted the references. Numbers between square brackets indicate citations. Can you give us a suggestion for an explicit reference associated with each number? Do not return anything except the citation number between square brackets and the corresponding reference.
> ===
> **[paper information]**

The existence of each generated reference was verified with Semantic Scholar (Kinney et al., 2023). For more details on the experimental setup we refer to Algaba et al. (2024). In Figure A1, we then describe our pipeline for generating GPT-generated and ground-truth citation graphs based on the data shared by Algaba et al. (2024).

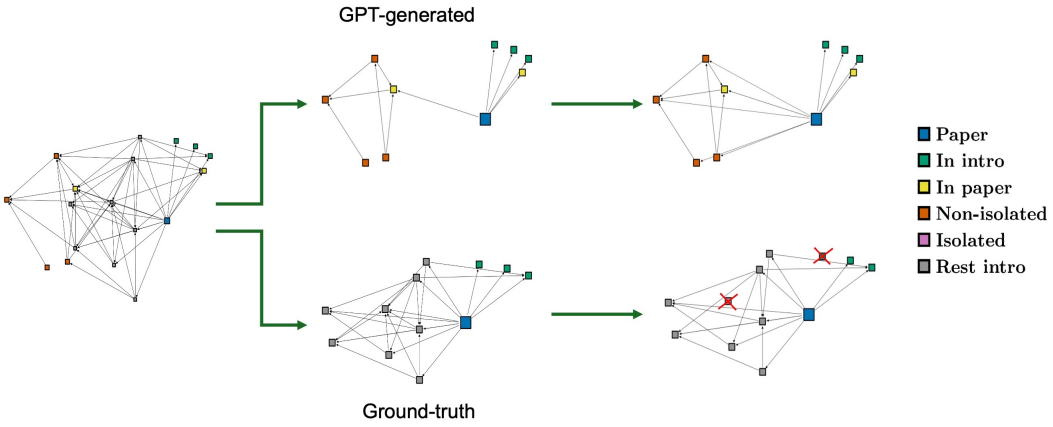

Figure A1: Pipeline to generate the ground-truth and GPT-generated citation graphs. The dataset comprises 166 papers, each represented as a citation network graph. In these graphs, references are categorized according to their positional context within the paper: the focal paper is depicted in blue; references generated by GPT-4 that appear in the introduction are marked in green, while those appearing later in the paper are marked in yellow; references generated by GPT-4 that are linked to the ground-truth or other generated references are shown in orange; and completely isolated GPT-4–generated references are rendered in purple. Ground-truth references that are not cited by GPT-4 are represented in grey. Two distinct graph types were constructed from each paper based on the reference generation method: GPT-generated graphs and ground-truth graphs. The GPT-generated graphs employ the color scheme of green, yellow, blue, orange, and purple, whereas the ground-truth graphs use green, grey, and blue. This process resulted in 332 graphs (166 GPT-generated and 166 ground-truth). To ensure that all nodes are connected to the focal paper (blue node), edges were added for nodes not initially connected to the blue node, as these references are inherently linked to the focal paper. For analytical simplicity, all graphs were converted to undirected graphs. Notably, 8 of the 166 papers did not exhibit sufficient connectivity in either the GPT-generated or the ground-truth graphs—often resulting in graphs with fewer than two nodes—which necessitated their exclusion from further analysis. Furthermore, a random subset of references was removed from the ground-truth graphs to ensure a fair comparison between graph types, thereby equating the size of both graph sets. This adjustment was essential because ground-truth graphs typically contained more nodes and edges, and direct or indirect size discrepancies can influence the extracted graph features and subsequent analyses.

CITATION GRAPH PROPERTIES

Below, there are precise definitions for the graph properties used in our experiments. For a more general application of graph measures in science of science, we refer to Yan & Ding (2009) and references therein. In Table A1, we also show the evaluation metrics for the random forest classifier on graph properties and embeddings.

Average shortest path length: A shortest path between two nodes in a graph is a path with the minimum number of edges. The average shortest path length measures the average number of steps along the shortest paths for all possible pairs of nodes.

Density: Density is defined as the proportion of actual connections present in the graph relative to the total number of possible connections.

Average Clustering Coefficient: For a given node, the clustering coefficient quantifies how close its neighbors are to forming a complete clique (a subset of nodes where every node is directly connected to every other node). The average clustering coefficient of a graph is the mean of the clustering coefficients of all individual nodes, summarizing the overall tendency of nodes to form clusters.

Degree Centrality: Degree centrality quantifies the number of direct connections that each node possesses within the network, thereby providing an assessment of a node's immediate influence.

Closeness Centrality: Closeness centrality is determined by the reciprocal of the average shortest path length from a given node to all other nodes in the network.

Standard Deviation of eigenvector Centrality: Eigenvector centrality measures a node's influence by considering the influence of its neighbors. The standard deviation of eigenvector centrality indicates the disparity in influence across the network—a low standard deviation suggests uniform influence, whereas a high standard deviation implies that a few nodes are considerably more influential than others.

Graph properties	Ground-truth vs. GPT	Ground-truth vs. Random	GPT vs. Random
Mean accuracy	0.5166 ± 0.0224	0.9271 ± 0.0264	0.9021 ± 0.0182
Mean F1-score	0.5209 ± 0.0387	0.9265 ± 0.0302	0.9066 ± 0.0168
Title embeddings			
Mean accuracy	0.6000 ± 0.0482	0.8688 ± 0.0214	0.7396 ± 0.0132
Mean F1-score	0.5998 ± 0.0653	0.8720 ± 0.0187	0.7471 ± 0.0166

Table A1: Performance of the Random Forest Classifier. The table presents the mean accuracy and F1-score obtained from applying a Random Forest classifier across five independent runs, utilizing both graph-based features and title embeddings. The dataset was partitioned into training and testing subsets using K-fold cross-validation, with a train set of 0.7 and a test size of 0.3. Hyperparameter optimization was conducted over the number of estimators (50, 100, 200), maximum tree depth (None, 10, 20, 30), minimum samples required for a node split (2, 5, 10, 20), and minimum samples per leaf (1, 2, 4). Given the balanced nature of the dataset and the binary classification setting, accuracy serves as a robust performance metric. However, this study emphasizes minimizing classification errors, specifically false positives—instances in which a ground-truth graph is misclassified as GPT-generated—and false negatives—cases where a GPT-generated graph is incorrectly identified as an ground-truth graph. As both types of misclassification have significant implications for the research objectives, the F1-score is employed as a key evaluation metric, providing a balanced evaluation of the model's precision and recall.