

Cite2Root: An Automated Citation Verification Platform

Dong Chen
University of British Columbia

Cite2Root is an AI-powered platform addressing the issue of mis-citation in academia and other media domains(future), a critical problem undermining the integrity of scholarly and public information. By automating the verification of claims and tracing their sources, Cite2Root enhances trust in citations and mitigates the risk of information contamination. Leveraging advanced document parsing, metadata retrieval, and language models, the platform provides a scalable solution to maintain robust connections to the source of truth, crucial in an era where generative AI amplifies disinformation. Cite2Root's unique multi-layer tracing approach also ensures that citation networks remain trustworthy over time, with applications extending beyond academia to journalism and media.

1 Problem Overview

In the ever-expanding digital information sphere, the rise of generative AI and disinformation has blurred the line between truth and fiction, creating unprecedented challenges in establishing trust. This evolving landscape demands that we anchor ourselves firmly to the source of truth, which is becoming increasingly critical as the gap between reality and virtual representations narrows. Without reliable anchors, the ability to distinguish between credible and manipulated information deteriorates, leaving individuals vulnerable to misinformation and echo chambers.

One of the most vital grounds lies in academia, where research relies on citations as anchors to foundational truths, providing a pathway to verify and trace knowledge to its origin. However, even in this rigorous environment, mis-citations occur. These errors mislead individual studies, accumulate over time, and risk contaminating entire information pathways. Mis-citation undermines the integrity of academic inquiry, eroding the trust placed in scholarship. To address this issue, we must build systems capable of error correction to restore and maintain robust connections to the source of truth, reinforcing the credibility of scholarly and public information landscapes.

2 Our Solution

Cite2Root is an AI-powered citation verification platform designed to ensure the integrity of information by automating claim-source alignment and tracing the origins of cited statements.

- a. **Technical Foundation:** Cite2Root combines document parsing, metadata retrieval, and language model-based evaluation to assess citation credibility. The method is validated in Takehiko et al. (2024).
- b. **Addressing AI Safety Challenges:** By mitigating the propagation of mis-citations, the platform directly enhances trust in AI-driven knowledge systems, combating the cascading risks of misinformation.

- c. **Competitive Advantages:** Cite2Root's multi-layer tracing feature uniquely identifies and corrects contamination points in citation chains. Its modular design ensures scalability across domains.
- d. **Diagrams and Proof-of-Concept Results:** Preliminary results from pilot testing demonstrate the platform's ability to identify unsupported claims in academic papers, with screenshots available in the appendix.
- e. **Innovation and Scalability:** The system's architecture is adaptable for diverse fields, from academia to journalism, ensuring long-term viability.

3 Pilot Experiment or Demo

- 1. **Goal:** Build a minimum viable product (MVP) capable of parsing academic PDFs, retrieving references, and assessing the credibility of every citation in the article.
- 2. **Methods:** Integrated pdf plumber for text extraction, CrossRef/JINA API for metadata, and frontier models like GPT-4o, and Claude for claim verification.
- 3. **Results:** Flagged unsupported claims in a dataset of ten academic papers, demonstrating proof-of-concept viability.
- 4. **Technical Challenges:** Addressed inconsistencies in citation formats and optimized LLM prompts for accuracy.
- 5. **Resources:** See GitHub repository: [GitHub: Cite2Root](#).

4 Potential Pitfalls and Mitigation Strategies

Cite2Root has immense potential to transform how citations are verified, but its success depends on overcoming several key challenges:

- 1. **Overreliance on AI Outputs:** While GPT-4 and similar models are powerful tools, they can occasionally misinterpret nuanced claims or hallucinate unsupported conclusions.
 - *Mitigation:* Create a fine-tuned model targeting to increase the accuracy of verification; also, implement a human-in-the-loop review system for flagged results, particularly in high-stakes scenarios such as medical or legal research. By combining automated and manual processes, we aim to ensure greater reliability and transparency.
- 2. **Scalability:** Processing large-scale citation networks and multi-layer tracing can strain computational resources, especially with an expanding user base.
 - *Mitigation:* Optimize computational pipelines by prioritizing high-impact citations and caching frequently accessed data. Implement cloud-based solutions to handle peak loads and ensure smooth scaling.

Addressing these pitfalls ensures Cite2Root's robustness, accuracy, and scalability, making it a trusted tool for verifying citations across domains.

5 Process

If given unlimited resources and one year, the roadmap for Cite2Root would include:

Timeframe	What will you do?
Next 3 months	A commercially usable gadget(likely via a website), use existing models for citation screening, with only one layer tracing
2025	Develop multi-layer citation tracing and enhance LLM performance with a fine-tuned screening model. With customers from all fields of academia.
2026	Scale platform to journalism and public media use cases, addressing diverse citation formats and contexts.
2027	Deploy a fully commercialized version with industry partnerships, ensuring global adoption and impact.

6 Impact on AI Safety & Key Risks

1. **Safety Benefits:** By preventing the propagation of misinformation, Cite2Root strengthens trust in the complex information landscape and ensures the information receiver is close to the source of the truth.
2. **Risks and Mitigation:**
 - (a) Risks: Mis/disinformation and loss of information autonomy due to lack of navigation tool.
 - (b) Mitigation: A verification system that traces the evolution of information that puts the receivers closer to the source of the truth.
3. **Alignment with Standards:** We aim to become the industrial standard for the verification pipeline.
4. **Monitoring and Evaluation:** Continuous performance tracking via accuracy metrics and user feedback.
5. **Success Metrics:**
 - 1. Adoption rate on both narrow domains and general publics
 - 2. Accuracy in flagging mis-citations/citation contamination
 - 3. Total number of mis-citations captured
 - 3. Total number of citation contamination prevented

7 Appendix

1. **Technical Documentation:** See detailed implementation notes in GitHub: Cite2Root
2. **Method Validation:** Takehiko et al. (2024).