Academic RefChecker: an AI-powered academic reference-validation tool

31 July 2025

Summary

Academic RefChecker is an open-source command-line utility that audits the references of academic manuscripts.
Given a PDF, LaTeX source, plain-text bibliography, or arXiv ID, the program:

1. **Extracts** every citation with a large-language-model (LLM) parser
2. **Queries** authoritative bibliographic APIs—Semantic Scholar, OpenAlex, and Crossref
3. **Reports** discrepancies such as misspelled authors, wrong publication years, malformed DOIs, incorrect arXiv IDs, and broken URLs

A single run produces a color-coded report together with drop-in corrected BibTeX entries, enabling authors, reviewers, and editors to safeguard citation integrity in seconds. Academic RefChecker is released under the MIT licence and currently comprises approximately 11k lines of tested Python with continuous integration on GitHub.

Statement of need

Reference lists are surprisingly error-prone—studies report mismatch rates of 10 – 25 % in several disciplines. Existing tools concentrate on style rather than factual accuracy, and rule-based parsers break on the myriad formats encountered "in the wild". LLMs, however, excel at normalising heterogeneous strings and rescuing information from layout-mangled PDFs. Academic RefChecker operationalises this capability while adding deterministic, multi-source verification—something no existing open-source package provides. The result is a lightweight checker that drops into author workflows (pre-submission) or editorial pipelines (peer review), finally closing a long-standing quality gap in scholarly publishing.

Academic RefChecker

Architecture

The pipeline has four stages:

1. **Ingestion** — accepts local files, URLs or academic identifiers.

2. **Bibliography localisation** — heuristics detect section boundaries (“References”, “Bibliography”, etc.).

3. **LLM-based parsing** — converts free-text citations into structured form for validation.

4. **Cross-validation** — resolves each field against Semantic Scholar (Allen Institute for AI 2024), OpenAlex (Priem, Piwowar, and Orr 2022), and Crossref (Crossref 2025), with retry logic that tolerates rate limits and partial matches using similarity heuristics.

Features

- Multi-format input: PDF, LaTeX, arXiv, plaintext.

- Pluggable LLM back-ends: OpenAI o3, Claude Sonnet 4, Gemini 2.5, local vLLM.

- Error taxonomy: author, title, venue, year, DOI, arXiv ID, URL; each assigned a severity level.

- Offline acceleration: optional SQLite mirror of Semantic Scholar (Allen Institute for AI 2024) records for sub-second look-ups.

- Extensibility: modular back-ends make it easy to add PubMed or DOAJ look-ups.

Usage examples

```
# Install from PyPI
pip install academic-refchecker

# Check a canonical paper by arXiv ID
academic-refchecker --llm-provider openai --llm-model gpt-4.1 --paper 1706.03762

# Audit a local manuscript and save a report
academic-refchecker my_draft.pdf --output results.txt
```

On the *Attention Is All You Need* bibliography (40 items), Academic RefChecker completed in approximately 2 minutes and surfaced multiple errors and warnings.

Comparison with related work

Crossref's Simple Text Query (Crossref 2025) provides free DOI lookup by pasting references but requires manual input and only validates DOIs, not authors, venues, or other metadata fields. Recite (Recite 2025) offers commercial reference validation focused on citation-reference consistency checking but lacks multi-source verification and comprehensive metadata validation. Scite.ai (Scite 2025) provides "Smart Citations" and reference validation features, identifying retracted papers and citation contexts, but operates as a proprietary research platform rather than a standalone validation tool. Amazon Science's (Hu et al. 2024) RefChecker focuses on hallucination detection in LLM outputs using knowledge triplets, not bibliographic metadata accuracy.

Academic RefChecker is, to our knowledge, the first open-source package that combines LLM-powered reference extraction from multiple input formats (PDF, LaTeX, arXiv) with comprehensive, multi-source factual validation across complete reference metadata—offering holistic, reference-by-reference verification against authoritative scholarly indices in a single automated workflow.

Acknowledgements

Development was accelerated with GitHub Copilot Agent, Cursor, and Claude Code. We thank the maintainers of Crossref, Semantic Scholar, and OpenAlex for their open APIs, and early testers for bug reports and feature suggestions.

References

Allen Institute for AI. 2024. "Semantic Scholar Academic Graph Api." https://www.semanticscholar.org/product/api.

Crossref. 2025. "Crossref Rest Api." https://www.crossref.org/documentation/retrieve-metadata/rest-api/.

Hu, Xiangkun, Dongyu Ru, Lin Qiu, Qipeng Guo, Yun Luo, Pengfei Li, Yue Zhang, and Zheng Zhang. 2024. "RefChecker: Reference-Based Fine-Grained Hallucination Checker and Benchmark for Large Language Models." *arXiv Preprint arXiv:2405.14486*. https://arxiv.org/abs/2405.14486.

Priem, Jason, Heather Piwowar, and Richard Orr. 2022. "OpenAlex: A Fully-Open Index of Scholarly Works, Authors, Venues, Institutions, and Concepts." *arXiv Preprint arXiv:2205.01833*. https://arxiv.org/abs/2205.01833.

Recite. 2025. "Recite Reference Checker." https://reciteworks.com/.

Scite. 2025. “Scite Reference Check.” https://scite.ai/.