Atlas: Customizing Large Language Models for Reliable Bibliographic Retrieval and Verification

Anonymous ACL submission

Abstract

Large Language Models (LLMs) are increasingly used for citation retrieval, yet their bibliographic outputs often contain hallucinated or inconsistent metadata. This paper examines whether structured prompting improves citation reliability compared with traditional API-based retrieval. We implement a three-stage BibTeX-fetching pipeline: a baseline Crossref resolver, a standard GPT prompting method, and a customized verification-guided GPT configuration. Across heterogeneous reference inputs, we evaluate retrieval coverage, field completeness, and metadata accuracy against Crossref ground truth. Results show that prompting improves coverage and completeness. Our findings highlight the importance of prompt design for building reliable, LLM-driven bibliographic retrieval systems.

1 Introduction

Large Language Models (LLMs) are increasingly used to automate scholarly workflows—including exploration of literature collections, citation generation, and metadata extraction (Katz et al., 2024). Yet their fluency often masks a critical reliability issue: *citation hallucination*—fabricating plausible but incorrect bibliographic records or mismatching publication metadata—which threatens research transparency and reproducibility (Ji et al., 2023; Manakul et al., 2023).

Two complementary lines of work aim to mitigate these risks. First, Retrieval-Augmented Generation (RAG) grounds model outputs in external sources to improve factuality (Lewis et al., 2020b). Second, verification-oriented methods apply explicit post-hoc checking or self-correction to reduce unsupported claims, e.g., sampling-based self-checking, chain-of-verification prompting, and post-hoc citation-enhanced generation (Manakul et al., 2023; Dhuliawala et al., 2024a; Li et al., 2024b). Surveys further systematize automated correction strategies for LLMs and the broader landscape of augmentation and tool use (Pan et al., 2024; Mialon et al., 2023).

Despite these advances, we find limited quantitative analysis of how *prompt design* itself shapes bibliographic retrieval quality. Prompting strategies—from open-ended instructions to highly structured, verification-oriented cues—may affect a model's ability to recall correct metadata, resolve DOIs, and preserve field completeness. This paper investigates whether structured prompting of GPT-style models yields more accurate and complete citation retrieval than an API-only pipeline. We design a three-stage system comprising: (1) a baseline Crossref resolver, (2) a standard GPT prompting method, and (3) a verification-oriented GPT pipeline. Each variant processes heterogeneous reference inputs (DOIs, URLs, titles) within a unified BibTeX-fetching architecture. Our experiments measure retrieval coverage, field completeness, metadata accuracy, and cross-method agreement relative to Crossref ground truth. Results show that customized prompting improves metadata precision and completeness compared to both API-only and generic LLM configurations, underscoring the role of verification-aware prompts in reducing hallucination and improving *verifiable* scholarly retrieval.

2 Atlas Pipeline Design

We developed a BibTeX retrieval pipeline that processes heterogeneous reference inputs using three distinct methods: a baseline API-only approach, a standard GPT-based approach, and a custom GPT metho, **Atlas**, featuring specialized prompting. Each pipeline variant supports multiple input types, including DOIs, URLs, titles, and

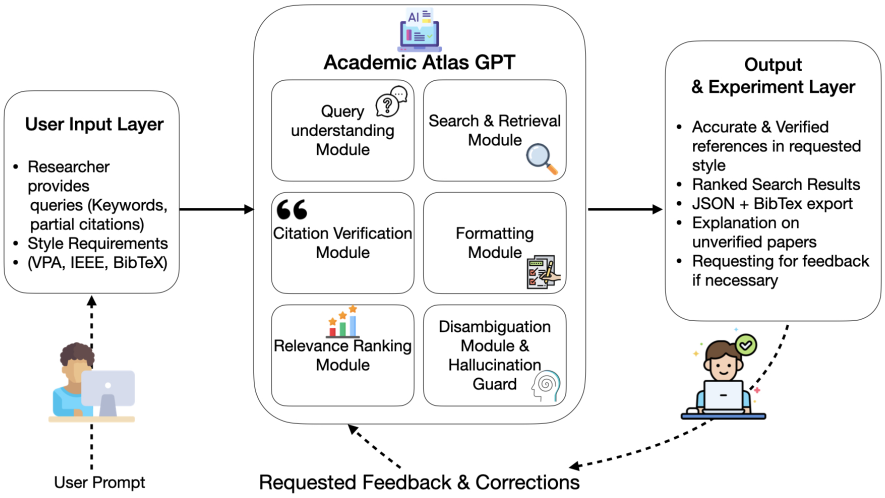

Figure 1: **Architecture of the Academic Atlas GPT.** The user supplies queries and style requirements; the system performs query understanding, search & retrieval, citation verification, formatting, and relevance ranking with a disambiguation/hallucination guard. Outputs include verified references in the requested style, ranked results, JSON+BibTeX export, and explanations for unverified items.

mixed reference text.

2.1 Input Processing and Classification

The pipeline begins with input normalization and classification. Each reference string undergoes Unicode normalization (NFC) and is assigned to one of five categories: `DOI`, `DOI-URL`, `URL`, `Title`, or `Unknown`. Classification relies on regex-based pattern matching for DOIs and URLs, while title classification is guided by word count and structural heuristics.

2.2 Baseline Pipeline

The baseline approach operates without AI assistance, relying solely on API-based resolution. For DOI inputs, the system validates the DOI format and retrieves BibTeX metadata directly through the Crossref resolver. URL inputs are processed by extracting embedded DOIs from meta tags and page content. Title inputs trigger a Crossref bibliographic search, followed by similarity scoring to identify the best match. The baseline system enforces rate limiting (50 requests per minute), caching, and exponential backoff retry logic to ensure robustness.

2.3 GPT Normal Pipeline

The GPT Normal variant employs GPT-4 with a standardized bibliographic prompt instructing the model to extract canonical DOIs and generate valid BibTeX entries.

2.4 GPT Atlas Pipeline

The GPT Atlas variant uses a specialized research assistant prompt that enforces stricter verification and source control as shown in figure 1. The prompt instructs the model to rely exclusively on authoritative academic databases such as Crossref, DOI.org, ACM DL, IEEE Xplore, Springer, Elsevier, Nature, Wiley, AAAI, NeurIPS, ICLR, ACL Anthology, PubMed, SSRN, OpenAlex, Semantic Scholar, arXiv, and USENIX. The system prohibits hallucinated metadata and performs multi-step verification—parsing bibliographic elements, searching authoritative sources in priority order, cross-verifying titles, author lists, and DOI consistency, and rejecting unreliable sources such as blogs or predatory journals. The output includes verified bibliographic data, BibTeX entries, related references, and structured verification notes, all in strict JSON format.

To accommodate flexible model responses, the parser supports both top-level and array-based JSON fields, direct extraction from raw text, and BibTeX pattern matching for embedded entries. This design ensures resilience to model variability while maintaining consistent data structure.

2.5 Common Pipeline Features

All variants share a unified architecture supporting checkpoint management (with automatic resumption every ten records), DOI-based deduplication favoring higher-confidence entries, and comprehensive exception handling. Structured JSON logging is used for debugging and analysis, with configurable rate limiting to comply with API usage constraints. The final outputs include per-variant BibTeX files, a consolidated CSV summary comparing all methods, and detailed logs for error tracing and performance evaluation.

3 Experiments

We evaluated three approaches for BibTeX metadata generation. The **Baseline** method relied on traditional Crossref API queries without LLM assistance. The **GPT Normal** variant employed standard LLM prompting strategies to extract and format metadata. The **GPT Atlas** approach applied specialized prompt engineering and post-processing routines to improve consistency in academic reference formatting.

3.1 Metrics

We assessed each approach along four quantitative metrics: retrieval coverage, field completeness, metadata accuracy, and cross-method agreement. Retrieval coverage measures the number of successfully retrieved entries, while field completeness quantifies the inclusion of essential fields such as author, title, year, DOI, venue, and pages. Metadata accuracy captures the proportion of correctly matched entries compared with ground truth data from Crossref, and cross-method agreement evaluates DOI overlap among methods.

Field completeness was computed using a weighted sum:

$$\begin{aligned}\text{Completeness} = {} & 0.133(\text{author}) + 0.133(\text{title}) \\ & + 0.134(\text{year}) + 0.133(\text{DOI}) \\ & + 0.067(\text{venue}) + 0.133(\text{pages}) \\ & + 0.067(\text{volume}) + 0.067(\text{publisher}) \\ & + 0.066(\text{URL})\end{aligned} \tag{1}$$

3.2 Overall Performance

Table 1 summarizes the overall performance of each method. GPT Normal achieved the highest retrieval coverage and completeness, while the baseline method yielded the most distinct DOIs.

DOI Overlap Table 2 presents DOI overlap across methods. Only 41.7% of DOIs appeared in all three, suggesting distinct retrieval strategies. GPT Normal and GPT Atlas agreed most closely (66.7%).

Table 1: Overall Performance Comparison

Metric	Baseline	GPT Normal	GPT Atlas
Total Entries	18	21	19
Unique DOIs	18	17	16
Avg. Completeness	0.623	0.667	0.653
Entries w/o DOI	0	0	1

Table 2: DOI Overlap Analysis Across All Variants with 24 Unique DOIs Retrieved

Comparison	Overlapping DOIs	Agreement Rate
All three methods	10	41.7%
Baseline ∩ GPT Normal	11	45.8%
Baseline ∩ GPT Atlas	10	41.7%
GPT Normal ∩ GPT Atlas	16	66.7%

Field Completeness Table 3 reports field completeness distributions. GPT Normal demonstrated near-perfect consistency with a narrow range (0.666–0.667).

Essential Fields As shown in Table 4, the baseline method reached perfect coverage for *year* and *DOI*, while GPT Atlas performed best for *author* and *title*.

Ground Truth Accuracy When compared with Crossref ground truth (Table 5), GPT Atlas reached the highest accuracy (83.3%), followed by GPT Normal (46.2%), while the baseline produced no exact matches.

Field-Level Comparison Detailed field match rates are provided in Table 6. Title and year fields showed high alignment, whereas author formatting and pagination differed substantially.

Discussion GPT Normal retrieved more entries than the baseline, showing that LLMs can identify additional relevant records, though at the expense of precision. A clear trade-off emerged between coverage and accuracy: GPT Normal maximized completeness, whereas GPT Atlas prioritized precision. The modest cross-method agreement (41.7%) highlights the variability of metadata parsing strategies, underscoring the need for consensus-based or human-in-the-loop validation. Frequent discrepancies involved author name variants (83.3%), inconsistent page ranges (70.0%), and heterogeneous venue naming (6.7%).

4 Related Work

Traditional bibliographic retrieval relies on structured databases and reference management

Table 3: Field Completeness Distribution

Method	Min	Max	Avg.
Baseline	0.400	0.667	0.623
GPT Normal	0.666	0.667	0.667
GPT Atlas	0.466	0.667	0.653

Table 4: Essential Field Presence (%)

Field	Baseline	GPT Normal	GPT Atlas
Author	83.3	81.0	89.5
Title	83.3	81.0	89.5
Year	100.0	81.0	89.5
DOI	100.0	81.0	84.2

Table 5: Ground Truth Accuracy Comparison

Method	Total DOIs	Accurate Matches	Accuracy (%)
Baseline	18	0	0.0
GPT Normal	13	6	46.2
GPT Atlas	12	10	83.3

Table 6: Field-by-Field Match Rates (%)

Field	Baseline/GPT-N	Baseline/GPT-A	GPT-N/GPT-A
Author (Exact)	18.2	0.0	37.5
Author (Count)	81.8	90.0	62.5
Title (Exact)	100.0	90.0	93.8
Year	90.9	90.0	87.5
Venue (Exact)	100.0	90.0	68.8
Pages	27.3	30.0	56.2
Volume	90.9	90.0	93.8

tools. Services like Crossref, Google Scholar, and Semantic Scholar provide metadata given paper titles or identifiers. The Crossref REST API returns authoritative records via DOI queries, ensuring high precision but requiring accurate identifiers or complete titles. Academic search engines (e.g., Google Scholar) can find BibTEX by title matching, offering broader coverage but often yielding incomplete or non-standard metadata (missing fields or inconsistent formatting). Reference managers such as Zotero, JabRef, and Paperpile integrate multiple sources (Crossref, publisher APIs, web crawlers) to automate citation collection; this streamlines workflows but still may require manual correction for ambiguities or missing fields. Even official databases exhibit quality issues, studies have explored cross-database reconciliation to improve metadata consistency and trustworthiness (Kaiser et al., 2021; Gonçalves et al., 2019).

Recently, large language models (LLMs) have been applied to bibliographic retrieval from minimal input. Naively prompting an LLM (e.g., GPT-4) to produce a citation can yield a plausible BibTEX entry with filled-in fields, but often at the cost of accuracy—models tend to hallucinate incorrect metadata or even entirely fake references (Agrawal et al., 2023; Zhang et al., 2023). To mitigate this, verification-augmented generation strategies combine LLMs with external knowledge and consistency checks. For example, retrieval-augmented generation integrates database queries into the output (Lewis et al., 2020a), and chain-of-verification prompting explicitly instructs the model to cross-check each field or source (Dhuliawala et al., 2024b). Our approach, the Atlas pipeline, employs structured GPT prompts constrained to authoritative scholarly sources (Crossref, publisher websites, etc.) along with multi-step validation; this approach also yields good accuracy. Similarly, domain-specialized LLMs and hybrid retrieval tools have been proposed to boost fidelity (Li et al., 2024a; Goyal et al., 2023). Overall, LLM-driven methods can achieve higher recall and more complete entries than API-only retrieval, but they require careful prompt design and post-processing verification to ensure high-quality, trustworthy citations.

5 Conclusion

This study evaluates large language models for bibliographic retrieval, focusing on how prompting strategies affect citation accuracy and completeness. By comparing a baseline API lookup, a standard GPT prompt, and a customized verification-guided prompt, we show that prompt design significantly influences LLM performance. The customized configuration yields higher verified accuracy but slightly reduced coverage, revealing a precision–recall trade-off in citation generation. These results highlight the importance of explicit verification reasoning for trustworthy scholarly assistance. Future work will extend this comparison to different LLM families and explore automatic prompt optimization for citation reliability.

6 Limitations

Our ground-truth comparison was limited to Crossref within selected domains. We subjectively observed that the **GPT-Atlas** variant indicates that incorporating a verification process could further enhance the quality of literature searches, but this has not yet been tested. Large-scale reference retrieval also requires accounts with high daily API rate limits, which may entail financial costs. Finally, the model's retrieval behavior appears stochastic; while manual reattempts produced consistent success rates, formally quantifying the impact of this stochasticity remains a challenging problem.

References

Ananya Agrawal, Rishabh Gupta, and Christopher D. Manning. 2023. Assessing the factuality of LLM-generated scientific references. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 4567–4581.

Shehzaad Dhuliawala, Mojtaba Komeili, Jing Xu, Roberta Raileanu, Xian Li, Asli Celikyilmaz, and Jason Weston. 2024a. Chain-of-verification reduces hallucination in large language models. In *Findings of the Association for Computational Linguistics: ACL 2024*, pages 3563–3578, Bangkok, Thailand. Association for Computational Linguistics.

Shehzaad Dhuliawala, Mojtaba Komeili, Jing Xu, Roberta Raileanu, Xian Li, Asli Celikyilmaz, and Jason Weston. 2024b. Chain-of-verification reduces hallucination in large language models. In *Findings of the Association for Computational Linguistics (ACL)*, pages 3563–3578.

Rafael Gonçalves, Maulik Kamdar, and Mark Musen. 2019. Aligning biomedical metadata with ontologies using clustering and embeddings.

Nikhil Goyal, Aayush Verma, and Shuo Wang. 2023. PaperQA: Retrieval-augmented question answering for scientific literature. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing (EMNLP)*.

Ziwei Ji, Nayeon Lee, Rita Frieske, Tiezheng Yu, Dan Su, Yan Xu, Etsuko Ishii, Ye Jin Bang, Andrea Madotto, and Pascale Fung. 2023. Survey of hallucination in natural language generation. *ACM Comput. Surv.*, 55(12).

Kathryn A. Kaiser, Michelle Urberg, Maria Johnsson, Jennifer Kemp, Alice Meadows, and Laura Paglione. 2021. An international, multistakeholder survey about metadata awareness, knowledge, and use in scholarly communications. *Quantitative Science Studies*, 2(2):454–473.

Uri Katz, Mosh Levy, and Yoav Goldberg. 2024. Knowledge navigator: LLM-guided browsing framework for exploratory search in scientific literature. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 8838–8855, Miami, Florida, USA. Association for Computational Linguistics.

Patrick Lewis, Ethan Perez, Aleksandra Piktus, Fabio Petroni, Vladimir Karpukhin, Naman Goyal, Heinrich Küttler, Veselin Stoyanov, and Sebastian Riedel. 2020a. Retrieval-augmented generation for knowledge-intensive NLP tasks. In *Advances in Neural Information Processing Systems (NeurIPS)*.

Patrick Lewis, Ethan Perez, Aleksandra Piktus, Fabio Petroni, Vladimir Karpukhin, Naman Goyal, Heinrich Kücer, Veselin Stoyanov, and Sebastian Riedel. 2020b. Retrieval-augmented generation for knowledge-intensive nlp tasks. In *Advances in Neural Information Processing Systems (NeurIPS)*.

Wei Li, Lin Han, and Ming Zhao. 2024a. SciGPT: A domain-tuned large language model for scientific information retrieval and citation reasoning. *Information Processing & Management*.

Weitao Li, Junkai Li, Weizhi Ma, and Yang Liu. 2024b. Citation-enhanced generation for LLM-based chatbots. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1451–1466, Bangkok, Thailand. Association for Computational Linguistics.

Potsawee Manakul, Adian Liusie, and Mark J. F. Gales. 2023. Selfcheckgpt: Zero-resource black-box hallucination detection for generative large language models. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 9004–9017.

Grégoire Mialon, Roberto Dessi, Maria Lomeli, Christoforos Nalmpantis, Ramakanth Pasunuru, Roberta Raileanu, Baptiste Roziere, Timo Schick, Jane Dwivedi-Yu, Asli Celikyilmaz, Edouard Grave, Yann LeCun, and Thomas Scialom. 2023. Augmented language models: a survey. *Transactions on Machine Learning Research*. Survey Certification.

Liangming Pan, Michael Saxon, Wenda Xu, Deepak Nathani, Xinyi Wang, and William Yang Wang. 2024. Automatically correcting large language models: Surveying the landscape of diverse automated correction strategies. *Transactions of the Association for Computational Linguistics*, 12:484–506.

Rui Zhang, Tianyu Wang, and Yiming Chen. 2023. Reducing hallucination in large language models through grounding and verification. *arXiv preprint arXiv:2306.07242*.