

A Self-Healing Architecture for Mitigating Bibliographic Hallucinations in LLM-Generated Academic Texts

Davide Tosi

`davide.tosi@uninsubria.it`

University of Insubria

Edoardo Crescenzo Mauriello

University of Insubria

Andrea Utzeri

University of Insubria

Research Article

Keywords: Academic Writing, Citation Verification, Hallucinations, Large Language Models, LLM, Self-Healing Architecture, Generative AI

Posted Date: August 26th, 2026

DOI: <https://doi.org/10.21203/rs.3.rs-10539555/v1>

License:   This work is licensed under a Creative Commons Attribution 4.0 International License.

[Read Full License](#)

Additional Declarations: No competing interests reported.

A Self-Healing Architecture for Mitigating Bibliographic Hallucinations in LLM-Generated Academic Texts

Crescenzo Edoardo Mauriello^{1†}, Andrea Utzeri^{1†}, Davide Tosi^{2*†}

¹Department of Theoretical and Applied Sciences, University of Insubria, Via Ottorino Rossi, 9, Varese, 21100, Italy.

²Department of Human Sciences and Innovation for the Territory, University of Insubria, Via M. E. Bossi, 10, Como, 22100, Italy.

*Corresponding author(s). E-mail(s): davide.tosi@uninsubria.it;
Contributing authors: cemauriello@uninsubria.it;
autzeri@studenti.uninsubria.it;

[†]These authors contributed equally to this work.

Abstract

Large Language Models (LLMs) have demonstrated remarkable capabilities in academic writing assistance, yet they frequently generate bibliographic hallucinations (i.e., plausible but non-existent citations) that undermine scholarly integrity. Hallucination rates in LLM-generated references can exceed 25%, with some models producing over 90% fabricated citations, posing risks to academic credibility, as evidenced by legal sanctions against professionals who submitted AI-generated documents containing false references. This paper presents a novel self-healing architecture that detects and mitigates bibliographic hallucinations in LLM-generated academic texts. The system implements a multi-layered verification framework, combining citation validation, cross-referencing with authoritative databases, and automated correction, to autonomously identify inconsistencies in bibliographic data (author names, titles, venues, dates) and either correct them using verified sources or flag them for human review. Experimental evaluation shows that the self-healing system significantly reduces hallucination rates while maintaining fluency and coherence. Our baseline evaluation with gpt-4o-mini showed that approximately 40% of generated DOIs were invalid. We tested the approach on more than 1,000 citations across 4 LLM engines, achieving an average identification accuracy of 0.96 and an F1 score of

0.93. Furthermore, optimizing the review (refinement) phase using a sentence-based approach yields significant computational efficiency, saving nearly 67% of tokens and latency compared to full-text refinement. Self-healing LLM thus provides a scalable, transparent, and robust solution for enhancing the trustworthiness of machine-generated scholarly content, contributing to the development of more reliable AI-assisted academic writing tools.

Keywords: Academic Writing, Citation Verification, Hallucinations, Large Language Models, LLM, Self-Healing Architecture, Generative AI

1 Introduction

The adoption of Large Language Models (LLMs) for academic and professional text generation has accelerated rapidly. These systems are now routinely employed to synthesize structured documents, scientific summaries, and literature reviews. However, as their utility expands, so does the critical challenge of ensuring factual integrity. A pervasive issue in this domain is the generation of bibliographic hallucinations: citations that appear structurally sound but reference non-existent DOIs, misattribute authorship, or fabricate publication metadata entirely. Although it is not easy to quantify the phenomenon of fabricated attributions, experimental studies such as the ones by Linardon et al. [1] and Liu et al. [2] still demonstrate how LLMs lack trustworthiness when it comes to creating accurate references.

Unlike Retrieval-Augmented Generative (RAG) systems that pull directly from a database, standard LLMs generate references probabilistically based on linguistic patterns rather than verified records, just as they generate normal text. Consequently, these models lack an intrinsic mechanism to cross-reference their outputs against authoritative indices such as CrossRef[3] or OpenAlex[4]. Even in top-tier architectures, these error rates remain alarmingly high; Linardon et al. (2025) [1] report that nearly 20% of citations in LLM-generated reviews are entirely fabricated, and Walters and Wilder [5] independently found that over 55% of GPT-3.5-generated citations were fabricated (about 20% for GPT-4), corroborating industry-reported estimates such as the AllAboutAI AI Hallucination Report 2026 [6], which puts the average hallucination rate for scientific queries at 16.9%. Taken together, these peer-reviewed and industry figures point to a consistent credibility trap where coherent prose masks factual decay.

To demarcate the scope of this problem, Table 1 presents a taxonomy of bibliographic errors, ranging from subtle metadata¹ inaccuracies (e.g., incorrect publication year) to extrinsic hallucinations where entire papers are fabricated.

Few existing frameworks offer a complete, autonomous pipeline for the post-generation repair of bibliographic data. Current solutions often require human-in-the-loop validation or are limited to flagging errors without correcting them.

¹Metadata is additional information that describes data and can enhance the organization, retrieval, and utilization of data.

Existing approaches to factual grounding in LLM-generated text generally fall into three categories:

- **Factual consistency checking:** external validators assess factual claims in generated text.
- **Retrieval-Augmented Generation (RAG):** integrates external retrieval into the generation process but does not specifically address structured bibliographic elements.
- **Self-verification and reflection:** methods primarily designed for improving reasoning accuracy rather than citation grounding.

Despite these advancements, no existing work explicitly targets the complete pipeline of detection, validation, and correction of bibliographic references in LLM-generated texts. This identifies a clear research gap: the need for post-generation verification mechanisms capable of ensuring the factual integrity of references produced by language models.

To address this problem, the present work proposes **Self-healing LLM**, a modular post-processing architecture that analyzes and corrects the references generated by an LLM. The system operates independently from the internal mechanisms of the model and is structured into four sequential modules:

1. **MONITOR:** identifies all bibliographic elements in the text, including DOIs, author-year references, and citation-like fragments.
2. **PLAN:** verifies each reference by querying trusted external databases such as CrossRef and OpenAlex.
3. **ACT:** removes invalid references and enriches valid ones with additional metadata when available.
4. **REVIEW:** refines the textual output using the LLM itself to repair grammatical and stylistic inconsistencies introduced by reference removal or rewriting.

The whole proposed self-healing pipeline has been evaluated on a ground-truth dataset of 1,000 citations and by testing 4 different LLM backends. Specifically, we studied our self-healing pipeline against **claude-haiku-4-5** (Anthropic), **deepseek-v4-flash** (DeepSeek), **gemini-2.5-flash** (Google), and **gpt-4o-mini** (OpenAI).

Self-healing LLM enhances both factual reliability and linguistic coherence while remaining scalable, transparent, and easily extensible for future developments.

2 Related work

The problem of ensuring the reliability of bibliographic references in automatically generated academic texts lies at the intersection of fact-checking, metadata validation, and LLM self-correction. Although no prior study directly addresses DOI verification within LLM-generated texts, several adjacent research areas provide relevant insights.

Table 1 Taxonomy of bibliographic hallucinations in LLM-generated academic text.

Category	Description
Non-existent DOI	The model produces a DOI that is syntactically valid but not registered in authoritative databases.
Incorrect DOI Metadata	A real DOI is cited but the associated metadata (authors, year, title, venue) is incorrect or partially fabricated. The citation appears credible, but the details do not match the actual publication.
Fabricated or Underspecified Citation	The model generates a citation that does not correspond to any real publication and lacks sufficient bibliographic structure. This often includes references expressed only as <i>author-year</i> , or citations missing title, venue, or DOI, making credibility and verifiability impossible.
Contextually Irrelevant Reference	The cited work exists but is semantically unrelated to the surrounding content. This typically results from purely probabilistic insertion rather than true grounding in the generated text.

Example cases (Hallucinated or erroneous elements highlighted in red):

1. Non-existent DOI:

Ex A: “Recent advancements in attention mechanisms have significantly reduced inference latency (DOI: 10.1145/1234567.8901234).”

Ex B: “As demonstrated by Chen and Wang (2023) in their recent survey, <https://doi.org/10.1016/j.bbr.2023.11456>.”

2. Incorrect DOI metadata:

Ex A: “The foundational work on this architecture was proposed by Smith and Johnson in *Advanced Methods in Neural Computation* (NeurIPS 2018). DOI: 10.3390/publications13040049.” (Real DOI, fabricated authors/title/venue).

Ex B: “In *The Future of AI* (Nature 2022), the authors argue that scaling laws plateau. DOI: 10.1038/s41586-021-03819-2.” (Real DOI, hallucinated title/year).

3. Fabricated / underspecified:

Ex A: “While early models struggled with long-term dependency (Martinez, 2017), modern transformers address this via sparse attention.” (Missing title, venue, and record entirely).

Ex B: “Several studies on LLM bounds (e.g., *Journal of AI Research*, Vol 42) suggest that factual recall is inherently limited.” (Lacks authors, specific year, and article title).

4. Contextually irrelevant:

Ex A: “The Byzantine architecture of the 5th century heavily utilized domed structures and intricate mosaics (Goodfellow et al., 2014).” (Goodfellow is the foundational GAN paper, entirely unrelated to architecture).

Ex B: “Our pipeline reduces hallucination rates by 40%, aligning with theoretical limits (Einstein, Podolsky, and Rosen, 1935, *Phys. Rev.* 47, 777).” (Real citation, absurd context).

2.1 Fact Checking and Knowledge Verification in LLMs

Large language models (LLMs) such as GPT-4 and Claude 3 have demonstrated impressive text-generation capabilities, but they are prone to factual errors and hallucinations. Research on fact-checking for LLMs has explored methods to improve factual accuracy, such as retrieval augmented generation (RAG), which grounds responses

in external knowledge sources [7], and post-hoc factual verification, where generated content is cross-checked [8, 9] or simply refined through a chain of thought [10].

Recent studies also examined the reliability of AI-generated bibliographic searches and literature reviews, comparing generative approaches with traditional human-led systematic reviews, highlighting common patterns of incomplete or underspecified citations [11, 12]. Empirical evaluations confirm the severity of this problem across disciplines: Walters and Wilder [5] found that over 55% of GPT-3.5-generated citations were fabricated and about 20% for GPT-4, while Chelli et al. [13] reported hallucination rates of up to 91.4% for some models, with accurate DOIs present in fewer than 20% of non-hallucinated references. Mugaanyi et al. [14] further showed that hallucination is not uniform across fields, with DOI fabrication reaching 89.4% in the humanities versus 29.1% in the natural sciences, underscoring the need for domain-agnostic verification pipelines. Similarly, Athaluri et al. [15] identified a high incidence of broken or misattributed DOIs even when the underlying article existed, a pattern directly addressed by the verification pipeline proposed in this work. The scale of the problem is growing: a large-scale audit of 2.5 million biomedical papers published in *The Lancet* [16] found that the rate of fabricated references increased more than 12-fold between 2023 and early 2026, coinciding with the widespread adoption of LLM-assisted writing tools.

2.2 Bibliographic Metadata and DOI Validation

In the domain of scholarly communication, reliable metadata management is crucial. Tools such as CrossRef [3] and OpenAlex [4] offer APIs for querying and verifying bibliographic records, including DOI metadata. Recent surveys [17] provide a taxonomy of hallucination mitigation techniques in LLMs, highlighting the need for integrating real-time bibliographic verification within the text generation pipeline.

2.2.1 Cross-Country DOI Fabrication in LLMs

A recent contribution that directly aligns with the objectives of this work is the cross-country analysis conducted by Kim, Kipchumba and Min[18], which systematically evaluates the accuracy of DOI citations generated by multiple LLMs. Their study demonstrates that DOI fabrication is not only frequent but also exhibits significant geographic and model specific variation, revealing structural weaknesses in how LLMs internalize and reproduce bibliographic metadata. This evidence reinforces the need for automated verification pipelines capable of validating DOIs in real time, as relying solely on model-generated references even for well established scientific publishers remains unreliable.

2.3 Self-Healing and Reflective Architectures in LLMs

Recent research has investigated self-reflective or self-healing mechanisms, where a model iteratively refines its own outputs. The SelfCheckGPT framework [19] introduced a probabilistic method to detect hallucinations through self-consistency checks, while Reflexion [20] proposed a reinforcement learning loop where the model critiques

and improves its prior responses. Other studies [21], [17] discuss architectural patterns to mitigate hallucinations in generated content. The architecture developed in this work adopts a similar principle, but applies it to the domain of bibliographic integrity. Unlike prior systems that rely solely on internal reflection, this approach integrates external factual grounding through CrossRef, enhancing both accuracy and traceability.

2.4 Citation Correction and Attribution in LLM-Generated Text

A parallel line of research has focused on improving the quality of citations produced by LLM-based systems, moving beyond generic hallucination detection toward citation-specific verification and correction mechanisms.

CiteFix [22] addresses citation misattribution in Retrieval-Augmented Generation (RAG) environments. In this setting, the supporting evidence is already available through the retrieval stage, yet the language model may associate a generated claim with an incorrect source. CiteFix introduces a post-processing attribution layer that re-evaluates the alignment between generated statements and retrieved documents using lexical, semantic, and learned similarity signals. Its objective is therefore not to verify the existence of a citation, but to improve source assignment within a predefined retrieval set. Consequently, the approach operates under a closed-world assumption and cannot address cases in which the cited publication itself does not exist.

CiteGuard [23] approaches citation reliability as a citation attribution problem. Given a textual claim and an implicit or missing reference, the system employs an iterative retrieval-and-validation process over scholarly databases to identify publications that best support the statement. Rather than correcting existing references, CiteGuard aims to recover the citation that a human author would plausibly associate with a given excerpt. This paradigm is particularly effective when a valid supporting source is assumed to exist and can be discovered through literature search. However, the framework is not designed to detect or repair fabricated bibliographic records generated directly by a language model.

The present work addresses a complementary problem that is largely orthogonal to both approaches: the autonomous generation of academic text by an LLM in the absence of an external retrieval context. In such scenarios, references may appear syntactically plausible while containing non-existent DOIs, fabricated metadata, invented authorship, or hallucinated publication venues. Rather than performing citation re-ranking or citation discovery, the proposed Self-Healing LLM focuses on post-generation bibliographic verification and repair. The architecture combines automatic citation extraction, validation against authoritative scholarly registries such as CrossRef [3] and OpenAlex [4], metadata consistency checking, and targeted textual refinement within a unified MPAR pipeline. Unlike approaches that assume the availability of retrieved evidence or an existing ground-truth citation, the proposed framework operates directly on raw LLM-generated content and is specifically designed to mitigate bibliographic hallucinations. Table 2.4 summarises the main differences among the three approaches. This is a *qualitative, conceptual* comparison rather than a direct experimental benchmark: CiteFix and CiteGuard target different scenarios

Dimension	CiteFix	CiteGuard	Self-healing LLM
Primary objective	Citation reassignment	Citation attribution	Citation verification & repair
Target scenario	RAG misattribution	Missing citation retrieval	Hallucinated references
Input assumption	Retrieved documents	Claim/excerpt	Raw LLM-generated text
Requires retrieval context	✓	✓	×
Handles fabricated references	×	×	✓
External scholarly source	Retrieved corpus	Semantic Scholar	CrossRef, OpenAlex
Bibliographic metadata validation	×	×	✓
Textual repair after correction	×	×	✓
Model-agnostic design	✓	✓	✓

Table 2 Comparison of citation attribution and correction approaches.

(RAG misattribution and citation retrieval, respectively), assume different inputs, and are evaluated with different metrics, so a head-to-head experimental comparison against the proposed pipeline on a shared dataset would not be meaningful.

3 Materials

This section presents the implementation details of the Self-Healing LLM developed to automatically detect, verify, and correct bibliographic references within AI-generated academic texts. The approach combines traditional programming techniques with AI-based text refinement, integrating several Python modules into a cohesive Self-Healing Layer architecture Monitor-Plan-Act-Review (MPAR).

3.1 Tools and Libraries

The implementation relies primarily on the Python programming language and the following tools and libraries:

- **re (Regular Expressions):** Used for pattern matching and text extraction, particularly for identifying DOI strings in different URL formats.
- **requests:** Employed to perform HTTP requests to the Crossref and OpenAlex API, which serves as the authoritative source for DOI metadata verification.
- **typing:** Provides type annotations to ensure code clarity and maintainability.
- **time:** Used for performance measurements and latency analysis.
- **openai:** The official OpenAI Python SDK, used to interface with GPT-based language models for syntactic and stylistic text refinement.
- **Custom modules (monitoring, plan, act, review, api_openai):** Each encapsulates a specific responsibility within the pipeline, promoting modularity and scalability.

3.2 Methodology

To address the challenge of bibliographic hallucinations, we developed a Self-healing LLM, a post-processing architecture based on the *Monitor-Plan-Act-Review* (MPAR) paradigm. The system is implemented in Python and operates as an autonomous

agentic loop that audits and repairs the output of a LLM without requiring access to the weights of the model.

The Self-Healing LLM consists of four sequential modules, orchestrated by a central integration layer. The workflow is designed to ensure that every bibliographic reference is detected, verified against authoritative metadata, and corrected or removed if found to be fraudulent.

The architecture is structured as a multi-stage pipeline, where each stage is responsible for a specific task within the overall self-healing process. The workflow progresses through interconnected components as shown in Fig. 1.

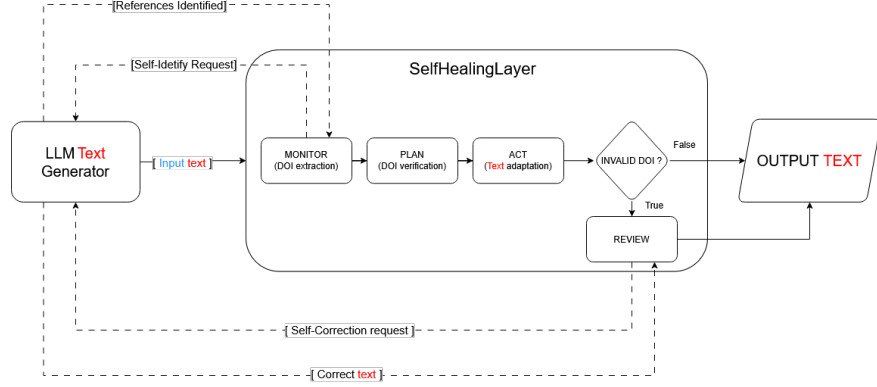


Fig. 1 Overview of the proposed Self-Healing Layer architecture based on the Monitor–Plan–Act–Review (MPAR) paradigm.

3.2.1 Monitor (Extraction)

The Monitor module constitutes the initial stage of the pipeline and initiates the self-healing process by steering the LLM toward a controlled self-auditing behavior. Instead of relying on predefined patterns or handcrafted parsing rules, it leverages the linguistic reasoning capabilities of the LLM to identify all bibliographic elements present in the generated text. To achieve this, the system prompts the model to re-examine its output and explicitly extract every fragment that resembles a citation, including DOIs, author-year references, bracketed numeric citations, and partially formed bibliographic strings. Although LLMs often hallucinate references that appear syntactically plausible but lack factual grounding, they remain notably effective at detecting citation-like structures, even those they have incorrectly produced.

The Monitor module exploits this property by imposing a strictly constrained output format. The LLM must return a single, well-structured JSON array describing each detected reference, including its textual span, surface form, and any explicit metadata (e.g., authors, year, DOI) without inferring or completing missing information. A lightweight extraction layer then isolates and validates the JSON block. The result is a normalized list of all bibliographic candidates, independent of correctness or factual grounding. This structured representation serves as the foundation for the

subsequent modules, which verify, correct, or remove citations through rule-based logic and external metadata services.

3.2.2 Plan (Verification)

The Plan module provides the factual grounding required for corrective actions. Building on the references extracted during monitoring, it evaluates the credibility of each citation by comparing it against authoritative bibliographic registries, primarily the Crossref metadata service. Its purpose is to transform a noisy set of citation candidates into an assessed collection, where each entry is explicitly annotated with a verification status.

This stage is essential because LLMs tend to generate citations in highly heterogeneous forms. Beyond fabricating DOIs or producing plausible but nonexistent entries, models frequently output author-year patterns without accompanying metadata, often to maintain narrative coherence rather than factual accuracy. Although such patterns resemble legitimate scholarly references, they cannot be treated as valid because they omit mandatory elements such as title, venue, or persistent identifier. The Plan module detects these incomplete structures and classifies them as unverifiable, preventing them from being interpreted as credible records.

For citations containing a DOI, the module performs hierarchical validation. It first checks whether the DOI resolves to an existing Crossref record. If it does, the module compares metadata fields (authors, title, publication year) provided by the LLM with canonical metadata obtained from the registry. Title similarity is evaluated through fuzzy matching, while author lists are compared through proportional overlap. A conservative title-similarity threshold of 0.97 is applied, ensuring that only references exhibiting near-perfect alignment with ground-truth metadata are marked as verified.

When mismatches occur, or when a reference lacks sufficient metadata for confirmation, the Plan module annotates it as inconsistent or unverifiable. This verification layer thus complements the Monitor stage: while the LLM identifies citation-like fragments, the Plan module determines whether those fragments correspond to real publications or hallucinated constructs. Its output provides the diagnostic basis for the Act module, which applies the appropriate repair strategy.

3.2.3 Act (Correction)

The Act module operationalizes the verification results by applying deterministic transformations to the text. While the Plan module establishes whether each citation is valid, consistent, or unverifiable, Act enforces the appropriate corrective action, ensuring that the final document contains only reliable and properly formatted references.

When a reference is confirmed as valid, typically through a verified DOI, the module replaces the noisy or incomplete citation with a normalized version constructed from authentic Crossref metadata. The corrected reference is then formatted according to a standard citation style (e.g., APA), ensuring accuracy and uniformity. For references associated with invalid DOIs, Act removes the fabricated entry entirely and inserts a placeholder to signal that the original citation lacked factual support.

A separate class of cases concerns unverifiable references, such as author-year patterns that cannot be linked to any real publication. Although linguistically plausible, these references cannot be trusted as bibliographic records. Instead of removing them, which might alter the meaning or structure of the text, Act appends a clear [UNVERIFIABLE] tag to the original expression. This preserves the continuity of the discourse while transparently indicating the lack of verifiable metadata.

Together, the Plan and Act modules form the corrective engine of the self-healing pipeline. Whereas the LLM contributes to generation and detection, these two evidence-based stages provide verification and repair, ensuring that the final bibliography is consistent, grounded, and free of hallucinated content.

3.2.4 Review (Refinement)

The Review module constitutes the final stage of the pipeline and restores the linguistic coherence of the corrected text. The interventions applied during the Act stage, such as removing invalid citations or inserting placeholders for unverifiable content, may leave behind grammatical artifacts (e.g., "As shown in [[OMIT]], the results..."). The purpose of the Review module is to eliminate these disruptions and regenerate fluent academic prose while preserving the original meaning.

Like the Monitor stage, Review leverages the generative abilities of the LLM, but under far stricter constraints. The model is instructed to rewrite only the sentences affected by earlier corrections, guided by explicit diagnostic information produced during the Plan and Act stages. In this way, the Review module completes the self-correction loop: the LLM revises its own text, but only within boundaries shaped by verified evidence.

To reduce computational cost and avoid unnecessary rewriting of unmodified content, the system adopts a sentence-level refinement strategy. It isolates sentences containing placeholders such as [[OMIT]] or tags marking unverifiable references, deduplicates them, and processes them in a single batch. Each sentence is then rewritten to remove the placeholder and restore grammaticality without altering the surrounding context. This targeted approach maintains the integrity of the corrected document while limiting the involvement of the LLM to the specific segments requiring refinement.

Through this final stage, the pipeline produces a coherent, publication-ready output. The generative capabilities of the LLM are used only where appropriate, while the evidence-based modules ensure that all corrections remain grounded, constrained, and aligned with verified bibliographic information. Together, these components enable a controlled form of self-revision that compensates for the inability of the model to self-validate in isolation.

3.2.5 Benchmark Methodology and Metrics

To assess the verification pipeline, we utilized a curated Ground-Truth Dataset containing verified references (labeled *TRUE REF*) and fabricated/corrupted citations (labeled *FALSE REF*).

The performance was evaluated as a binary classifier using a standard confusion matrix to assess the relationship between the predictions and the ground truth. In

standard notation, (T) rue/(F)alse defines correct identification, meaning that prediction and ground truth are aligned; (P)ositive/(N)egative defines the prediction of the target class, in our case, real or fabricated. This framework yields four distinct outcomes: True Positives and True Negatives, representing accurate classifications, and False Positives and False Negatives, representing classification errors.

Based on these definitions, we computed standard performance metrics such as:

Accuracy to measure the overall proportion of correct predictions; Precision to evaluate how many predicted positives are actually correct; Recall to compute how many real positives the model correctly detects; and the F1-score as the harmonic mean of precision and recall.

4 Experimental Results

To evaluate the *Self-healing LLM* pipeline, we designed a three-phase experimental protocol. First, we established a baseline of hallucination rates in unverified LLM outputs. Second, we assessed the accuracy of our verification pipeline across four state-of-the-art language models using an expanded benchmark dataset. Third, we measured the computational efficiency of our optimized refinement strategy.

4.1 Ground Truth Dataset

To construct the evaluation dataset, we first established a ground-truth corpus of 987 pristine academic citations, programmatically extracted from the bibliographies of highly cited review papers spanning a broad and multidisciplinary range of fields, from discipline-specific venues in engineering, medicine, the humanities, and economics to general-purpose scientific outlets, ensuring that all baseline references possessed complete, verified metadata. Every citation in the corpus carries a resolvable DOI, spans publication years from 1924 to 2026, is drawn from 421 distinct venues, and predominantly consists of journal articles (870) and conference proceedings (109), with a small number of book chapters; no preprints were included. To simulate real-world LLM hallucinations and formatting errors, we applied a controlled perturbation function to this ground truth. This programmatic corruption systematically injected specific error modalities: subtly altering titles or publication years and injecting erroneous authors to simulate Incorrect DOI Metadata; truncating citations to minimal, unverifiable author-year pairs to create Fabricated or Underspecified Citations; and replacing legitimate DOI prefixes with syntactically valid but unregistered strings (e.g., "10.fake/") to generate Non-existent DOIs. A control group of unchanged citations was retained to serve as valid references. The benchmark dataset and pipeline code described in this section are available from the corresponding author upon reasonable request.

It is important to distinguish this synthetically corrupted benchmark from the real-world baseline reported below (Section 4, Baseline Evaluation), which analyses DOIs organically hallucinated by `gpt-4o-mini` during unconstrained generation, without any programmatic perturbation. The two evaluations serve complementary purposes: the real baseline establishes that bibliographic hallucination is a genuine, naturally occurring failure mode at a rate consistent with prior literature, while the synthetically corrupted benchmark provides the labeled, class-balanced ground truth needed

to systematically measure detection and correction performance across error categories and models. The corruption modalities were designed to emulate the categories of Table 1 and the failure patterns documented in prior work on LLM citation fabrication (e.g., truncated author-year pairs mirror the underspecified citations reported by Walters and Wilder [5], and fake DOI prefixes mirror the non-resolving DOIs observed in our own baseline and in Athaluri et al. [15]). Nonetheless, we acknowledge that programmatic corruption cannot fully reproduce the statistical signature of organically generated hallucinations, and we treat the two results as complementary rather than directly comparable; this distinction is revisited in the Limitations discussion.

4.2 Baseline Evaluation: Reliability of Generated DOIs

Using our *Monitor* module, we extracted 651 DOI-like strings from the generated text. Each DOI was verified against the CrossRef and OpenAlex API services [24, 25]. As shown in Table 3, 40% of the generated identifiers were non-existent despite being syntactically valid, confirming that without external verification, nearly half of the citations produced by the model would lead to dead links or “404 Not Found” errors.

Table 3 Baseline validity analysis of DOIs generated by gpt-4o-mini.

Metric	Count	Percentage
Total DOIs Extracted	651	100%
Valid DOIs (Existing)	389	60%
Invalid DOIs (Hallucinations)	262	40%

The first experiment aimed to quantify the prevalence of bibliographic hallucinations in a standard generation setting. We prompted **gpt-4o-mini** to generate 100 academic-style paragraphs on diverse scientific topics, specifically requesting the inclusion of bibliographic references.

4.3 Benchmark Verification Performance

To rigorously evaluate the pipeline across different LLM backends, we compiled a benchmark of 1,000 citations per model, drawn from a dataset of validated and synthetically corrupted bibliographic references spanning four perturbation categories: *valid references* (no perturbation), *Incorrect DOI Metadata*, *Fabricated or Underspecified Citations*, and *Non-existent DOIs*. Samples were randomly drawn with a fixed seed and organised into groups of three citations each, with shuffling applied to ensure a uniform distribution of valid and corrupted references across groups. Each group was embedded into a generated scientific paragraph, whose references were drawn from the dataset described above, so as to evaluate the monitor’s ability to extract and verify citations from naturalistic text rather than from isolated structured input. To stress-test the title-based verification fallback, a controlled proportion of

the benchmark—20% of valid references and 15% of corrupted ones—consists of references lacking a DOI identifier, requiring the pipeline to rely exclusively on title-based fuzzy matching.

The benchmark was evaluated independently against four LLM backends: `claude-haiku-4-5` (Anthropic), `deepseek-v4-flash` (DeepSeek), `gemini-2.5-flash` (Google), and `gpt-4o-mini` (OpenAI). The aggregated performance metrics along with the confusion matrices are reported in Table 4.

To contextualise the four adopted LLM backends, we refer to the Vectara Hallucination Leaderboard [26], which reports baseline hallucination rates. As for the four adopted models, the hallucination rate spans from approximately 6% to 9.8%. It must be noted, however, that those figures are measured on open-ended document summarization tasks, and therefore they capture a general propensity for unfaithful generation rather than errors specific to bibliographic reference handling. The two measures are therefore not directly comparable. Nevertheless, they provide a meaningful contextual baseline. On the one hand, they highlight the need for a self-healing mechanism that automatically detects and corrects hallucinated citations. On the other hand, they characterize the intrinsic reliability of each model in grounded generation settings, thereby providing a reference point against which the effectiveness of the proposed self-healing verification layer can be qualitatively assessed. Within the domain-specific task of citation verification, the pipeline achieves false-positive rates between 6.0% and 6.7% across all four backends (i.e., the rate of corrupted citations incorrectly passed as valid), alongside F1-scores of 0.936–0.948 for three of the four models, and overall accuracies exceeding 96%. These results confirm that the self-healing mechanism provides consistent and quantifiable verification gains, and that this holds regardless of the underlying model provider. Three out of the four LLM providers (Claude Haiku, Gemini Flash, and DeepSeek) demonstrate robust citation verification performance, achieving F1-scores of 0.936–0.948 and accuracies exceeding 96.0%. In contrast, `gpt-4o-mini` exhibits a notable performance gap, with a recall of 0.849 and a correspondingly higher false-negative rate, resulting in an F1-score of 0.889. These results indicate that model-specific differences in information extraction and reasoning capabilities influence the downstream verification pipeline, yet the self-healing architecture succeeds in constraining false-positive propagation to comparable levels across all providers. The pipeline demonstrates strong robustness across most corruption patterns, with false positives concentrated in scenarios where references lack explicit DOI information. In such cases, the verification mechanism relies on title-based matching heuristics, which exhibit inherently lower discriminative power compared to deterministic DOI validation. This trade-off reflects the fundamental challenge of reference verification when authoritative identifiers are absent or invalid, a limitation that generalises beyond this specific implementation to the broader domain of citation analysis.

4.4 Citation Correction and Refinement Efficiency

Following the verification phase, the *Act* module intercepts all references flagged as invalid by the *Plan* and applies targeted corrections directly on the source text using exact character offsets. Each citation is handled according to its verification status:

Table 4 Confusion matrices and performance metrics for all evaluated models.

Model	Confusion Matrix				Precision	Recall	F1	Accuracy
	TP	FN	FP	TN				
claude-haiku-4-5 (Anthropic)	298	14	21	667	0.934	0.955	0.945	0.965
gemini-2.5-flash (Google)	296	16	19	669	0.940	0.949	0.944	0.965
deepseek-v4-flash (DeepSeek)	298	14	19	669	0.940	0.955	0.948	0.967
gpt-4o-mini (OpenAI)	265	47	19	669	0.933	0.849	0.889	0.934
Average					0.937	0.927	0.931	0.958

references confirmed via CrossRef are replaced with a canonical, metadata-complete form reconstructed from the retrieved record; references carrying an invalid DOI are removed and marked for subsequent refinement; references that could not be verified due to insufficient metadata are annotated as [UNVERIFIABLE]; and references for which the extracted metadata contradicts the CrossRef record are flagged as [CONTRADICTION]. Replacements are applied in reverse offset order to preserve positional consistency throughout the document. Crucially, this correction step operates exhaustively, every citation identified as hallucinated or inconsistent during the Plan phase is intercepted and acted upon.

The subsequent *Review* phase repairs the text after citation removal, and is the most computationally expensive component of the pipeline. We compared two implementation strategies across 30 test paragraphs to measure latency and token usage.

1. **Full-text Refinement:** The entire paragraph is re-submitted to the LLM.
2. **Sentence-based Refinement (Optimized):** Only sentences containing annotation placeholders (e.g., [[OMIT_INVALID_DOI]]) are extracted and processed.

The results, averaged in Table 5, show that the optimized strategy reduces the computational burden by approximately two-thirds.

Table 5 Efficiency Comparison (Average per Paragraph).

Metric	Full-text	Sentence-based	Reduction
Total Tokens	2072	675	67.4%
Cost (USD)	\$0.000736	\$0.000229	68.9%
Latency (s)	22.42	7.21	67.8%

By isolating the context to only the sentences affected by citation removal, the optimized strategy achieves a **~68% reduction in both cost and latency**. Crucially, this optimization did not degrade the linguistic quality of the output, as the LLM successfully repaired the local grammatical structure (e.g., removing “as seen in [OMIT]”) without needing to process the surrounding unaffected text.

4.5 Limitations

While the results above demonstrate consistent verification performance across four LLM backends, several limitations should be acknowledged.

First, the verification pipeline depends entirely on the coverage of CrossRef and OpenAlex. Both registries are strongest for journal articles and conference proceedings with registered DOIs; coverage is comparatively weaker for books, non-English scholarship, older publications, and preprints that have not yet been indexed. In domains or venues underserved by these registries, valid citations may be more likely to be misclassified as unverifiable, and the pipeline’s effectiveness cannot be assumed to generalize uniformly across all fields.

Second, as discussed in Section 4, the main benchmark relies on programmatically corrupted citations rather than citations that LLMs actually hallucinated unprompted. Although the corruption modalities were designed to emulate the taxonomy in Table 1 and patterns reported in prior empirical studies, synthetic corruption cannot fully reproduce the statistical signature of naturally occurring hallucinations, so the reported accuracy and F1 figures should be interpreted as an upper-bound estimate of production performance rather than a direct measurement of it. Evaluating the pipeline on real LLM-generated academic text with human-verified citation labels remains an important direction for future work.

Third, the current evaluation does not include a comparison against end-to-end retrieval-augmented generation (RAG) systems that avoid hallucination at generation time rather than repairing it post hoc; such a comparison would clarify the relative merits of post-generation repair versus generation-time grounding and is left for future work.

Fourth, running the self-healing pipeline at document- or corpus-scale introduces cumulative API and LLM inference costs and latency (Section 4 reports per-citation and per-paragraph figures); scaling to large corpora, high submission volumes, or real-time editorial workflows would require batching, caching, or rate-limiting strategies that are outside the scope of the present evaluation.

Finally, the pipeline has not been systematically tested against adversarial or unconventional citation styles (e.g., non-standard formatting, mixed citation systems, or intentionally obfuscated references), and its behavior in such edge cases remains an open question.

5 Conclusion

This paper introduced a Self-healing LLM architecture for the automatic detection, verification, and repair of bibliographic hallucinations in AI-generated text. The proposed framework combines DOI extraction, external validation through metadata services, hallucination detection, and targeted text refinement within a modular MPAR (Monitor–Plan–Act–Review) loop.

When evaluated on a benchmark of 1,000 citations across four state-of-the-art language models, the verification pipeline achieved consistently strong performance, with accuracies between 93.4% and 96.7% and F1-scores up to 94.8%. These findings demonstrate that the proposed method generalizes effectively across multiple

LLM backends while maintaining robust detection capabilities under different citation corruption scenarios.

In addition to verification accuracy, the proposed sentence-based refinement strategy substantially improves efficiency. By restricting the repair process to the affected sentences rather than reprocessing entire documents, the system reduced token consumption, latency, and operational cost by approximately 68% without compromising output quality.

Overall, the results suggest that external verification combined with targeted self-repair can significantly improve the reliability of AI-generated scholarly content.

Declarations and statements

- **Funding:** The authors declare that no funds, grants, or other support were received during the preparation of this manuscript.
- **Data Availability:** The datasets generated during and/or analysed during the current study are available from the corresponding author on reasonable request.
- **Clinical Trial Number:** Not applicable.
- **Consent to Participate declaration:** not applicable.
- **Consent to Publish declaration:** not applicable.
- **Ethics declaration:** not applicable.

References

- [1] Linardon, J., et al.: Influence of topic familiarity and prompt specificity on citation fabrication in mental health research using large language models: Experimental study. *BMJ Mental Health* <https://doi.org/10.2196/80371>
- [2] Liu, N.F., et al.: Evaluating Verifiability in Generative Search Engines (2023). <https://doi.org/10.48550/arXiv.2304.09848>
- [3] Hendricks, G., Tkaczyk, D., Lin, J., Feeney, P.: Crossref: The sustainable source of community-owned scholarly metadata. *Quantitative Science Studies* **1**(1), 414–427 (2020) https://doi.org/10.1162/qss_a_00022
- [4] Forchino, M.V., Torres-Salinas, D.: The OpenAlex database in review: Evaluating its applications, capabilities, and limitations. *Journal of Informetrics* **20**(2), 101800 (2026) <https://doi.org/10.1016/j.joi.2026.101800>
- [5] Walters, W.H., Wilder, E.I.: Fabrication and errors in the bibliographic citations generated by ChatGPT. *Scientific Reports* **13**, 14045 (2023) <https://doi.org/10.1038/s41598-023-41032-5>
- [6] AllAboutAI: AI Hallucination Report 2026: Which AI Hallucinates the Most? (2026). <https://www.allaboutai.com/resources/ai-statistics/ai-hallucinations/>
- [7] Lewis, P., Perez, E., Piktus, A., Petroni, F., Karpukhin, V., Goyal, N., Küttler, H., Lewis, M., Yih, W.-t., Rocktäschel, T., Riedel, S., Kiela, D.: Retrieval-augmented

- generation for knowledge-intensive NLP tasks. *Advances in Neural Information Processing Systems* **33**, 9459–9474 (2021)
- [8] Gao, L., Dai, Z., Pasupat, P., Chen, A., Chaganty, A.T., Fan, Y., Zhao, V.Y., Lao, N., Lee, H., Juan, D.-C., Guu, K.: RARR: Researching and Revising What Language Models Say, Using Language Models (2023). <https://arxiv.org/abs/2210.08726>
 - [9] Gou, Z., Shao, Z., Gong, Y., Shen, Y., Yang, Y., Duan, N., Chen, W.: CRITIC: Large language models can self-correct with tool-interactive critiquing. In: *Proceedings of the International Conference on Learning Representations (ICLR)*, pp. 57734–57811 (2024)
 - [10] Dhuliawala, S., Komeili, M., Xu, J., Raileanu, R., Li, X., Celikyilmaz, A., Weston, J.: Chain-of-Verification Reduces Hallucination in Large Language Models (2023). <https://arxiv.org/abs/2309.11495>
 - [11] Tosi, D.: Comparing generative AI literature reviews versus human-led systematic literature reviews: A case study on big data research. *IEEE Access* **13**, 56210–56219 (2025) <https://doi.org/10.1109/ACCESS.2025.3554504>
 - [12] Metze, K., Morandin-Reis, R.C., Lorand-Metze, I., Florindo, J.B.: Bibliographic research with large language model ChatGPT-4: Instability, hallucinations and sometimes alerts. *Clinics* **79**(4), 1234–1245 (2024) <https://doi.org/10.1016/j.clinsp.2024.100409>
 - [13] Chelli, M., *et al.*: Hallucination rates and reference accuracy of ChatGPT and Bard for systematic reviews: Comparative analysis. *Journal of Medical Internet Research* **26**, 53164 (2024) <https://doi.org/10.2196/53164>
 - [14] Mugaanyi, J., Cai, L., Cheng, S., Lu, C., Huang, J.: Evaluation of large language model performance and reliability for citations and references in scholarly writing: Cross-disciplinary study. *Journal of Medical Internet Research* **26**, 52935 (2024) <https://doi.org/10.2196/52935>
 - [15] Athaluri, S.A., *et al.*: Exploring the boundaries of reality: Investigating the phenomenon of artificial intelligence hallucination in scientific writing through ChatGPT references. *Cureus* **15**(4), 37432 (2023) <https://doi.org/10.7759/cureus.37432>
 - [16] Topaz, M., Roguin, N., Gupta, P., Zhang, Z., Peltonen, L.-M.: Fabricated citations: An audit across 2.5 million biomedical papers. *The Lancet* **407**, 1779–1781 (2026) [https://doi.org/10.1016/S0140-6736\(26\)00603-3](https://doi.org/10.1016/S0140-6736(26)00603-3)
 - [17] Kazlaris, I., Antoniou, E., Diamantaras, K., Bratsas, C.: From illusion to insight: A taxonomic survey of hallucination mitigation techniques in LLMs. *Information* **6**(10), 260 (2025) <https://doi.org/10.3390/ai6100260>

- [18] Kipchumba, F., Kim, E., Min, S.: Geographic variation in LLM DOI fabrication: Cross-country analysis of citation accuracy across four large language models. *Publications* **13**(4), 49 (2025) <https://doi.org/10.3390/publications13040049>
- [19] Manakul, P., Liusie, A., Gales, M.: 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*, Singapore, pp. 9004–9017 (2023). <https://doi.org/10.18653/v1/2023.emnlp-main.557>
- [20] Shinn, N., Cassano, F., Gopinath, A., Narasimhan, K., Yao, S.: Reflexion: Language agents with verbal reinforcement learning. *Advances in Neural Information Processing Systems* **36**, 8634–8652 (2023)
- [21] Ji, Z., *et al.*: Survey of hallucination in natural language generation. *ACM Transactions on Information Systems* **55**(12), 1–38 (2023) <https://doi.org/10.1145/3571730>
- [22] Maheshwari, H., Tenneti, S., Nakkiran, P.: CiteFix: Enhancing RAG Accuracy Through Post-Processing Citation Correction (2025). <https://arxiv.org/abs/2504.15629>
- [23] Choi, Y.M., Guo, X., Fung, Y.R., Wang, Q.: CiteGuard: Faithful Citation Attribution for LLMs via Retrieval-Augmented Validation (2025). <https://arxiv.org/abs/2510.17853>
- [24] CrossRef: CrossRef REST API Documentation (2025). <https://api.crossref.org/>
- [25] OpenAlex: OpenAlex REST API Documentation. <https://developers.openalex.org>
- [26] Hughes, S., Bae, M., Li, M.: Vectara Hallucination Leaderboard (2023). <https://github.com/vectara/hallucination-leaderboard>