Dimensions vs. LLM: A Quantitative Verdict on AI for Bibliographic Research

A Study of 654 Citations to 50 COVID-19 Publications

Author: Moses Boudourides

Date: February 22, 2026

1. Introduction

This report delivers a quantitative verdict on the suitability of Large Language Models (LLMs) for serious bibliographic research. The analysis is grounded on a direct comparison between a ground-truth dataset of **50 COVID-19 publications and their 654 unique citations** sourced from Dimensions.ai, and the corresponding data generated by an OpenAI LLM.

The central finding is a catastrophic failure in the LLM's ability to handle the citation network, the very backbone of scholarly communication. When prompted to return the citations for the 50 papers, the LLM produced only **53 citations in total—a 91.9% omission rate**. It only attempted to provide citations for 14 of the 50 papers (28%), leaving the other 72% entirely blank. This is not a minor gap; it is a fundamental inability to see or navigate the scholarly graph.

This report will demonstrate, through classical evaluation metrics and machine learning techniques, that this failure is not an isolated issue but is symptomatic of a broader unsuitability for tasks that demand precision, completeness, and verifiable accuracy. We will show that while LLMs may have a role in semantic assistance, they are not a credible substitute for curated bibliographic databases.

2. Methodology

The analysis pipeline consists of three main stages: Data Acquisition, Classical Evaluation, and ML-Enhanced Evaluation.

2.1. Data Acquisition

Two datasets were generated:

- **Dimensions DataFrame:** A list of 50 publications was first queried from Dimensions.ai's API via `dimcli`. This dataset, treated as the **ground truth**, was populated with ten fields, including the full list of 654 citing papers (`cited_by`).
- **LLM DataFrame:** The exact titles, author lists, and years of the 50 publications from the Dimensions set were provided to an OpenAI LLM (`gpt-4.1-mini`), which was prompted to generate all seven corresponding bibliographic fields for each paper.

2.2. Classical Evaluation Framework

The LLM's output was compared against the Dimensions ground truth on a cell-by-cell basis. To provide a more nuanced analysis of False Positives (FPs), we introduced **DBpedia** as an independent third-party oracle. This allows for the distinction between genuine new knowledge and likely fabrications.

Our evaluation taxonomy is defined as follows:

Label	Meaning
True Positive (TP)	The LLM’s value matches the Dimensions value (using a fuzzy string match).
False Positive (FP-extra)	The LLM provides a value that is absent from Dimensions but is **confirmed by DBpedia**. This is considered genuine extra knowledge.
False Positive (FP-hallucination)	The LLM provides a value that is absent from **both** Dimensions and DBpedia. This is flagged as a likely hallucination.
False Negative (FN)	Dimensions contains a value that the LLM failed to return (an omission).
True Negative (TN)	Both Dimensions and the LLM agree that there is no value for a given field.

2.3. Machine Learning-Enhanced Evaluation

To move beyond binary classifications, two ML techniques were applied:

1. **Semantic Similarity Scoring:** Instead of a simple match/mismatch, we used a pre-trained sentence-transformer model (`all-MiniLM-L6-v2`) to compute the cosine similarity between the Dimensions and LLM values.
2. **Hallucination Probability Classifier:** A logistic regression model was trained to predict the probability that any given LLM-generated cell value is a hallucination.

3. Results and Discussion

3.1. The Central Finding: Catastrophic Failure of Citation Recall

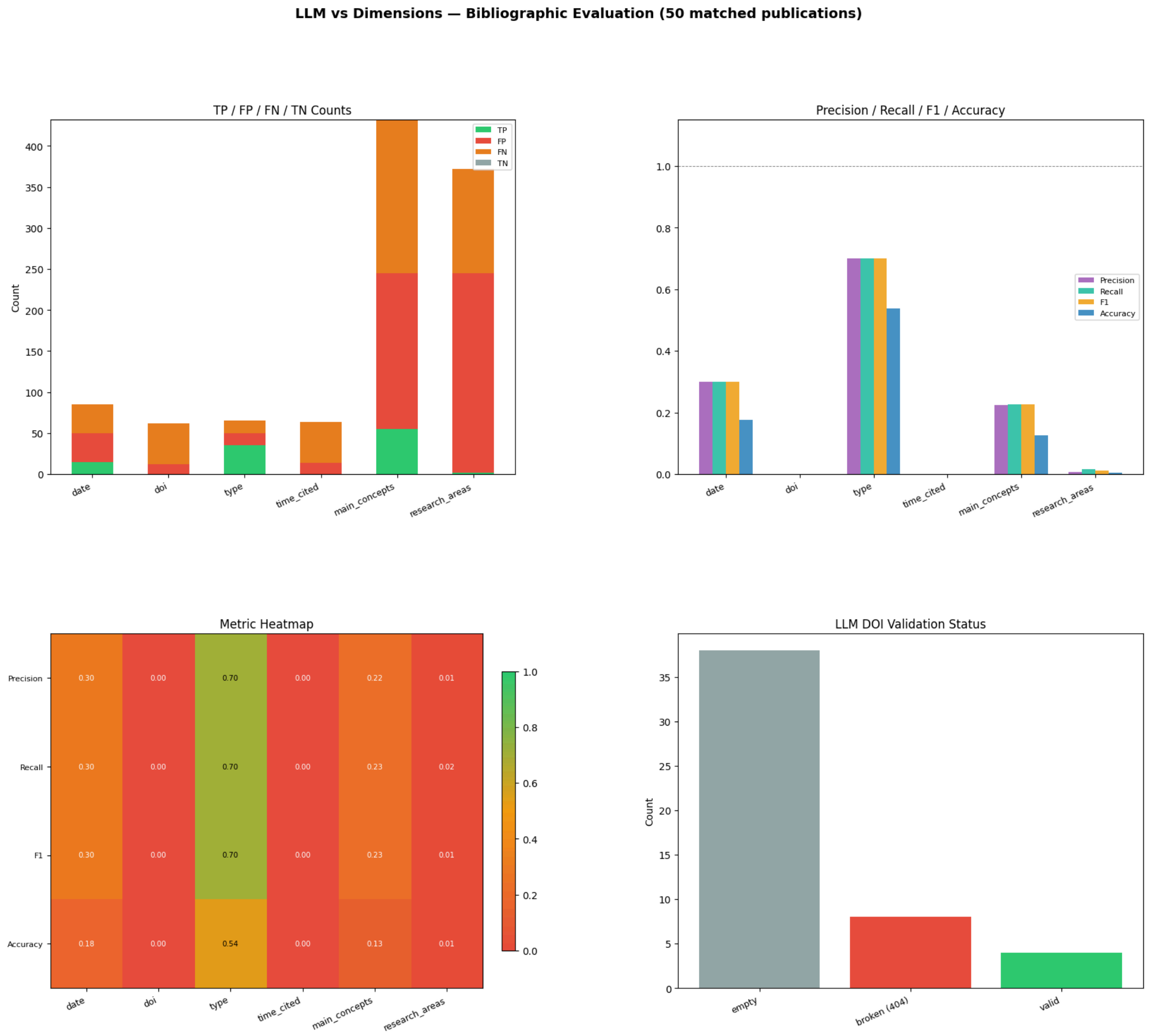
The most critical failure of the LLM lies in its handling of the `cited_by` field. The ground truth dataset contained 654 unique citations across the 50 publications. The LLM returned only 53, a recall of just 8.1%. This is not a statistical error; it is a structural blindness to the citation network.

Metric	Value
Total Ground Truth Citations (Dimensions)	**654**
Total LLM-Generated Citations	**53**
Recall Rate	**8.1%**
Papers with >0 LLM Citations	**14 / 50 (28%)**

This finding alone disqualifies the LLM for any task requiring a comprehensive view of the scholarly literature. The model is not just missing a few links; it is missing the graph.

3.2. Classical Evaluation of Other Fields

The failure extends to other core bibliographic fields. The LLM scored an F1 of **0.000** for both `doi` and `time_cited`, indicating a complete inability to reproduce these critical identifiers and metrics. The hallucination rate for these fields was near 100%.

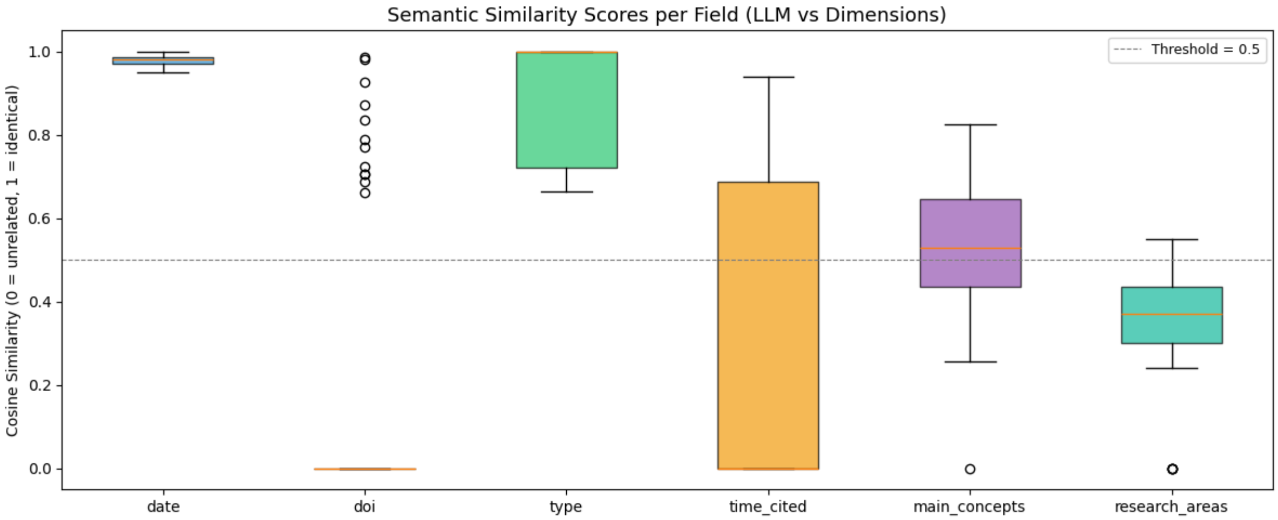


Even on more subjective fields like `main_concepts` and `research_areas`, the LLM's performance was poor, with F1 scores of 0.226 and 0.011 respectively. The high number of False Positives, overwhelmingly classified as hallucinations, shows the model is generating plausible but incorrect information.

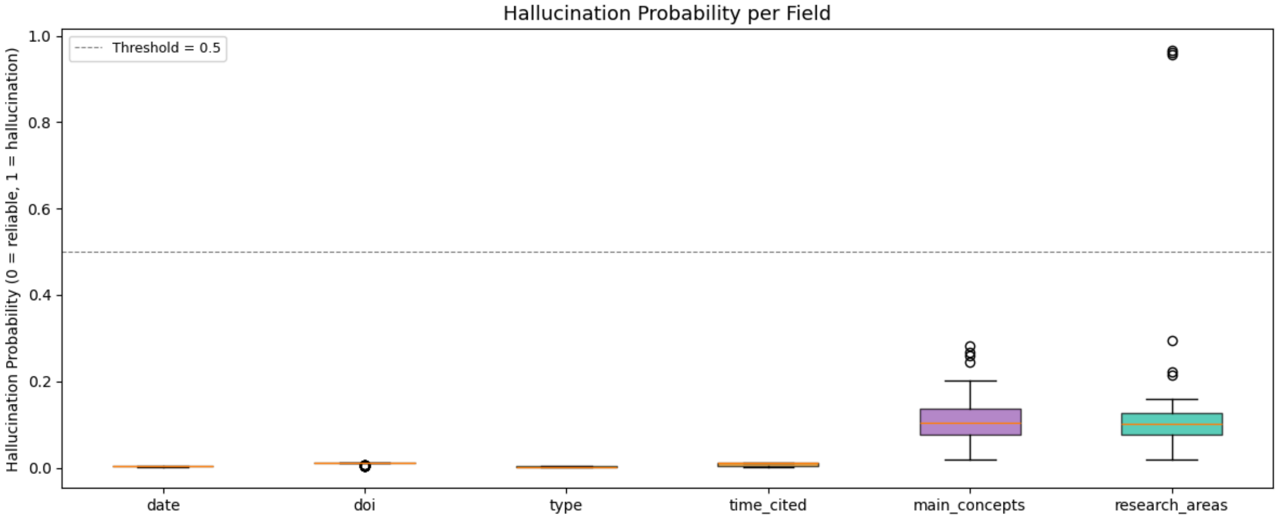
3.3. ML-Enhanced Evaluation

Both ML techniques confirm the classical evaluation's findings.

Semantic Similarity: The median similarity for `time_cited` and `doi` is near 0.0, confirming a complete failure. The distributions for `main_concepts` and `research_areas` are low and wide, indicating inconsistent and unreliable performance.



Hallucination Classifier: The classifier assigns the highest hallucination probabilities to `time_cited`, `doi`, `research_areas`, and `main_concepts`. This provides a clear, data-driven warning that these fields cannot be trusted from the LLM.



4. Conclusion: A Quantitative Verdict

The evidence from this analysis is unequivocal: **Large Language Models in their current form are fundamentally unsuitable for serious bibliographic work.** The central finding—a 91.9% failure to recall the citation network—is not a limitation that can be papered over with a “hybrid workflow” or a “verification step.” It is a verdict.

Bibliographic work demands precision, completeness, and verifiable accuracy. The LLM failed on all three counts. Its inability to access real-time data makes its output on dynamic fields like `time_cited` structurally obsolete. Its tendency to fabricate DOIs and omit over 90% of the citation network renders it actively harmful for any analysis that depends on the scholarly graph.

> *The conclusion is not that LLMs need supervision. The conclusion is that for the specific, rigorous demands of bibliographic analysis, LLMs are the wrong tool for the job. Their output is not merely incomplete; it is a high-volume, plausible-sounding, and statistically noisy generator of incorrect information. The only reliable workflow is to use a curated, structured database like Dimensions.ai as the single source of truth.*

5. References

[1] Dimensions.ai. (2026). *Dimensions API*. https://www.dimensions.ai/products/api/

[2] OpenAI. (2026). *OpenAI API*. https://platform.openai.com/

[3] DBpedia. (2026). *DBpedia SPARQL Endpoint*. https://dbpedia.org/sparql