

Article

Geographic Variation in LLM DOI Fabrication: Cross-Country Analysis of Citation Accuracy Across Four Large Language Models

Eungi Kim ^{*}, Frankline Kipchumba and Sein Min 

Department of Library and Information Science, Keimyung University, 1095 Dalgubeoldaero, Dalseo-Gu, Daegu 42601, Republic of Korea; 1114476@stu.kmu.ac.kr (F.K.); 1005985@stu.kmu.ac.kr (S.M.)

* Correspondence: egkim@kmu.ac.kr

Abstract

This study evaluates digital object identifier (DOI) hallucination in large language model (LLM)-generated scholarly citations, with a focus on systematic geographic disparities. To conduct this study, we systematically evaluated four LLMs (GPT-4o-mini, Claude-3-haiku, Gemini-2.0-flash-lite, and DeepSeek V3) using standardized information behavior prompts across ten countries with diverse income levels. The models generated 3451 citations, which we validated using the CrossRef API. The results showed that DOI hallucination follows systematic patterns influenced by model choice, geographic context, and publication recency. Hallucination rates exceeded 80% in lower-income countries and increased sharply for publications from the 2020s across all regions. Fabricated citations—citations that appear structurally complete but contain invalid DOIs—were especially prevalent in countries such as India and Bangladesh. Model-specific factors showed the strongest association with hallucination, followed by income level and publication period. These findings raise concerns about the epistemic reliability of LLM-generated scholarly references and underscore the need for region-aware training, real-time DOI validation, and robust verification protocols in academic contexts.

Keywords: Large Language Models (LLMs); DOI hallucination; geographic bias; citation accuracy; cross-country analysis



Academic Editor: Andrew Kirby

Received: 29 July 2025

Revised: 12 September 2025

Accepted: 16 September 2025

Published: 1 October 2025

Citation: Kim, E., Kipchumba, F., & Min, S. (2025). Geographic Variation in LLM DOI Fabrication: Cross-Country Analysis of Citation Accuracy Across Four Large Language Models.

Publications, 13(4), 49. <https://doi.org/10.3390/publications13040049>

Copyright: © 2025 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license (<https://creativecommons.org/licenses/by/4.0/>).

1. Introduction

Citations are fundamental to academic research, serving to validate claims and link new studies to established knowledge (Bornmann & Daniel, 2008; Hyland, 1999). However, generating accurate and reliable citations presents challenges, particularly with the increasing use of Large Language Models (LLMs) such as ChatGPT (Mugaanyi et al., 2024; Zhang & Zhao, 2025). Key concerns include citation bias, which favors certain publication years or geographic regions, and issues with citation accuracy, including invalid Digital Object Identifiers (DOIs) and hallucinated references. DOIs are critical in scholarly communication, ensuring that academic sources remain traceable and verifiable (Teixeira da Silva & Nazarovets, 2022). Tools like CrossRef provide essential services for verifying DOI accuracy and managing citation data (Daquino et al., 2020; Hendricks et al., 2020). Citation bias, where studies with statistically significant results are cited more often, has been well documented in medical and other research fields (Jannot et al., 2013). Likewise, hallucinated DOIs, in which LLMs generate incorrect or nonexistent references, have become a growing concern in AI-generated content (Frosolini et al., 2024).

To investigate these geographic patterns systematically, this study examines geographic patterns using information behavior research in the Library and Information Science (LIS) domain. Information behavior research—encompassing how people seek, use, share, and assess information needs—has documented scholarship across countries with varying economic development levels, providing the geographic breadth necessary for our analysis (Earhart et al., 2021). The conceptual framework maintains sufficient consistency across cultural contexts to enable controlled comparison while allowing for meaningful local variation. Moreover, information behavior research has adequate scholarly depth across our ten selected countries to support systematic citation requests without exhausting available literature. These characteristics make information behavior research within LIS appropriate for testing whether LLM citation generation patterns vary systematically by geographic context.

In this study, we build on prior research, such as Bridges (2024), which identified a significant rate of hallucinated references and DOIs when evaluating ChatGPT's performance in generating diagnostic citations. However, existing citation hallucination studies did not examine geographic variation, despite documented evidence of systematic geographic bias in other LLM outputs (Manvi et al., 2024; Moayeri et al., 2024). Training data imbalances predict geographic bias in LLMs (Kumari et al., 2024), but the specific issue of DOI fabrication across different countries has received limited attention despite its importance for citation verification. Thus, the objective of this study is to investigate geographic variation in DOI fabrication patterns across four LLMs, examining how citation accuracy varies across countries representing varying levels of economic development.

2. Literature Review

Research relevant to this study spans three areas: LLM citation fabrication and geographic bias, global publishing inequalities, and mitigation strategies. LLMs have been shown to systematically fabricate DOIs and exhibit geographic bias that disproportionately affects lower-income regions. Studies report that models generate “totally artificial” citations by creating plausible author names, journal titles, and DOI numbers that lead to non-functional links. For example, Cheng et al. (2025) found that 38% of ChatGPT-generated references contained incorrect or fabricated DOIs, with only 7% fully accurate. Bridges (2024) observed that ChatGPT-4 produced only 52 correct DOIs (31.5%) out of 145 clinical references. Jedrzejczak et al. (2024) noted that ChatGPT often “borrowed author names from one paper and merged them with the title of another,” generating entirely fictitious DOIs. These findings, across domains including medical education (Cheng et al., 2025), health sciences (Jedrzejczak et al., 2024), and chemistry education (Ruff et al., 2024), suggest that DOI hallucination is systematic rather than incidental.

Geographic patterns are similarly consistent. Manvi et al. (2024) demonstrated that LLMs display systematic bias against locations with lower socioeconomic conditions, finding Spearman correlations up to 0.70 for negative portrayals of regions such as Africa in domains like attractiveness, morality, and intelligence. Moayeri et al. (2024) found that error rates in LLM factual recall were 1.5 times higher for countries from Sub-Saharan Africa compared to North American countries, demonstrating pervasive geographic disparities across 20 different LLMs. Liu et al. (2025) found that geographic bias extends across foundation models (GPT-3.5, GPT-4o, Mistral, Claude), noting the counter-intuitive result that “less advanced models can generate more geographically diverse content than state-of-the-art ones,” implying that model complexity may exacerbate geographic inequity.

Global publishing infrastructure may also contribute to citation fabrication in under-represented regions. Chavarro et al. (2025) reported that although 47% of journals using Open Journal Systems originate from low-income countries, many are excluded from global

indexes due to barriers such as DOI registration through Crossref. [Gathama et al. \(2025\)](#) found that 45% of references in the Kenya Eye Health Journal lacked DOIs, which may help explain the high hallucination rates for Kenya-specific citations. [Ramesh et al. \(2025\)](#) showed that AI systems like WildfireGPT systematically exclude references without DOIs through validation filters, creating structural biases against research from regions lacking DOI infrastructure.

To address these issues, several mitigation strategies have been proposed. In a Delphi study involving 86 publishing professionals, [Ng et al. \(2025\)](#) found strong agreement on the importance of flagging fabricated DOIs, with 84.9% rating it as “important” for journal evaluation. [Ng et al. \(2024\)](#) reported that 91.8% of researchers and clinicians viewed fake DOI detection as essential for promoting transparency. On the technical front, [Kumari et al. \(2024\)](#) identified training data imbalance—specifically, “overrepresentation of data from certain parts of the world”—as a major cause of regional exclusion. They proposed Dynamic Region-Aware Fine-Tuning (DRAFT), which uses region-specific embeddings and dynamic sampling to improve geographic representation by 20% without reducing model accuracy.

Despite these contributions, no previous study has systematically examined how geographic location and national income level jointly influence LLM citation accuracy. This study addresses that gap by analyzing how DOI fabrication varies across countries with differing economic contexts, offering the first large-scale assessment of geographic bias in LLM-generated scholarly citations.

3. Materials and Methods

We employed a controlled experimental design to systematically investigate geographic bias and DOI hallucination patterns across four major LLMs. The overall research process is outlined in Figure 1. The following subsections describe each step of this methodology in detail.

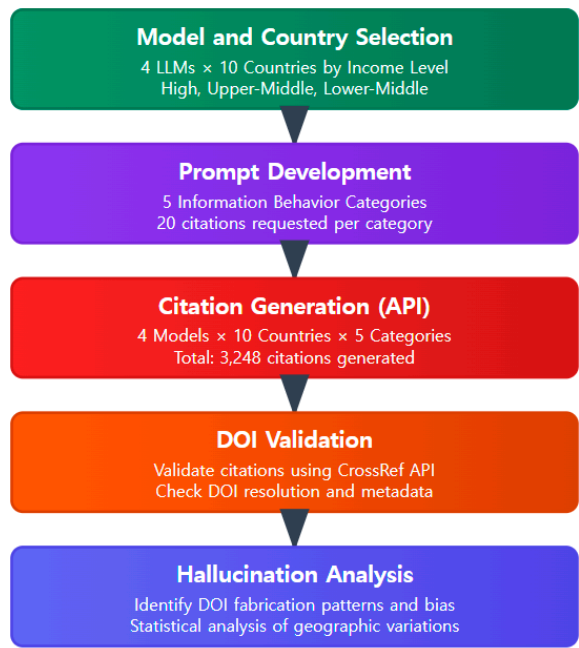


Figure 1. Research Methodology Process Flow.

3.1. LLM and Country Selection

Four LLMs were selected for this study: GPT-4o-mini, Claude-3-haiku, Gemini-2.0-flash-lite, and DeepSeek V3. All models were configured with identical parameters, using

a temperature of 0.2 to minimize randomness and a 2000-token limit to ensure consistent and complete citation generation. The models tested vary in scale and training approaches (e.g., GPT-4o-mini as a lightweight model vs. DeepSeek V3 as a larger model), which may contribute to performance differences beyond architectural design factors.

We selected ten countries using purposive sampling based on the World Bank's fiscal year 2024 income classification ([World Bank, 2024](#)) to represent varying levels of economic development:

- High-income countries (United States, United Kingdom, Germany, South Korea, Australia): characterized by well-established digital infrastructures and comprehensive scholarly databases
- Upper-middle-income countries (China, Brazil): representing regions with growing scholarly capacity and expanding research infrastructure
- Lower-middle-income countries (India, Kenya, Bangladesh): underrepresented in global research databases and characterized by developing scholarly infrastructure

3.2. Prompt Development and Information Behavior Categories

We used a standardized system prompt and five user prompt variations, all targeting different aspects of information behavior research. This domain was chosen for its conceptual clarity, cross-cultural relevance, and strong foundation in LIS scholarship. The system prompt instructed each model to act as an academic research assistant and produce 20 real citations with valid DOIs in a tab-delimited format. The five user prompts addressed: Information Seeking, Use, Sharing, Needs, and General Behavior. Each prompt included the target country and maintained consistent formatting requirements to ensure experimental comparability. Table 1 lists the prompt texts used for each category.

Table 1. Information Behavior Research Prompts by Category.

Category	Prompt Text
Information Seeking	"Provide 20 academic journal articles with DOIs that investigate information seeking behavior and search strategies in [country]. Format your response as a tab-delimited table with columns for Author Names, Title, Publication Year, Journal Name, and DOI."
Information Use	"Provide 20 academic journal articles with DOIs that examine how people evaluate, use, and apply information in [country]. Format your response as a tab-delimited table with columns for Author Names, Title, Publication Year, Journal Name, and DOI."
Information Sharing	"Provide 20 academic journal articles with DOIs that study information sharing, dissemination, and communication behaviors in [country]. Format your response as a tab-delimited table with columns for Author Names, Title, Publication Year, Journal Name, and DOI."
Information Needs	"Provide 20 academic journal articles with DOIs that investigate information needs assessment and identification in [country]. Format your response as a tab-delimited table with columns for Author Names, Title, Publication Year, Journal Name, and DOI."
Information Behavior	"Provide 20 academic journal articles with DOIs that study general information behavior patterns and practices in [country]. Format your response as a tab-delimited table with columns for Author Names, Title, Publication Year, Journal Name, and DOI."

3.3. DOI Validation and Geographic Relevance Assessment

Following LLM citation generation, a total of 3451 citations were collected across all models and countries. Each citation containing a DOI required validation to determine accuracy and geographic relevance. All DOIs were validated using the CrossRef REST API ([Hendricks et al., 2020](#)) with proper academic headers and normalization protocols to ensure compliance and standardization.

The validation process involved DOI normalization (removing common prefixes and validating 10.xxxx format), submitting each DOI using HTTP GET methods with institutional identification headers, and evaluating response status. DOIs returning successful responses (HTTP 200) with complete bibliographic metadata were classified as valid, while those generating error responses or missing essential metadata fields were classified as hallucinated. Robust error handling included HTTP status code management (200/404/other), automatic retry with exponential backoff (maximum 2 attempts), 15-s timeout protection, 0.5-s rate limiting delays, and incremental progress saving every 10 validations to ensure data integrity during extended validation runs.

A custom Python 3.12 script was developed to perform DOI validation and geographic relevance checks using the CrossRef API. The script includes routines for DOI normalization, API querying, metadata extraction, and country-specific content analysis based on author affiliations, journal locations, and research context indicators extracted from titles and abstracts. Systematic error handling was implemented to manage format errors, network failures, rate limits, and server responses. The complete script, together with the datasets, is available in the public repository referenced in the Data Availability Statement.

3.4. Statistical Analysis Framework

Primary outcome variables included DOI hallucination rate and geographic bias rate. The DOI hallucination rate measures the proportion of citations that contain DOIs which fail to resolve through the CrossRef API. Let N_{doi} denote the total number of citations that include a DOI, and let N_{invalid} represent the number of those DOIs that do not return a valid response. The hallucination rate is defined as:

$$\text{HR}_{\text{DOI}} = \frac{N_{\text{invalid}}}{N_{\text{DOI}}}$$

where HR_{DOI} is the DOI hallucination rate, N_{invalid} is the number of non-resolving DOIs, and N_{DOI} is the total number of citations containing a DOI. A DOI was considered hallucinated if it failed to return a successful resolution regardless of syntactic validity. The geographic bias rate was defined as the proportion of valid DOIs lacking explicit national relevance based on content analysis. Both measures provide consistent and interpretable metrics of citation-level reliability.

Citation completeness was assessed using a 5-point scale based on the presence of essential bibliographic elements. Each citation received one point for each of the following components: (1) author name(s), (2) article title, (3) publication year, (4) journal name, and (5) DOI. The completeness score ranged from 0 (no elements present) to 5 (all elements present), providing a standardized measure of surface-level citation formatting independent of factual accuracy.

We used Fisher's exact tests with Monte Carlo simulation (2000 iterations) to assess categorical associations, including DOI hallucination by income level and LLM, geographic bias patterns, and citation yield analysis. Citation yield analysis was conducted to examine whether models consistently generated the requested number of citations across experimental conditions. This method was selected over the chi-square test due to low expected frequencies and zero counts in several cross-tabulation cells, which violate chi-square assumptions and can lead to inaccurate p -values. The simulated Fisher test provides more reliable significance estimates under sparse data conditions, particularly important for citation yield analysis where the experimental design resulted in relatively small cell counts ($N = 195$ conditions across 10 countries, 4 models, and 5 prompt types). Furthermore, Kruskal–Wallis rank sum tests were used to examine temporal variation in hallucination rates across publication periods. Spearman's rank correlation assessed relationships be-

tween citation age and hallucination, accommodating non-linear trends and outliers. Effect sizes included Cramer's V for categorical comparisons and Spearman's ρ for correlation analyses. All tests were conducted in R with full documentation, version control, and structured export pipelines to ensure transparency and reproducibility.

4. Results

4.1. DOI Hallucination Patterns by Model and Income Level

Figure 2 shows DOI hallucination rates across three income levels for four LLMs. Overall, hallucination rates increase as income levels decrease. In High-Income countries, all models show lower rates, with Claude-3-haiku and DeepSeek V3 below 50%, while Gemini-2.0-flash-lite and GPT-4o-mini exceed this threshold but remain moderate. For Upper-Middle-Income countries, Gemini-2.0-flash-lite reaches approximately 73%, with Claude-3-haiku, DeepSeek V3, and GPT-4o-mini exhibiting moderate increases. In the Lower-Middle-Income category, hallucination rates remain high across all models, with Claude-3-haiku and GPT-4o-mini nearing 80–85%, while DeepSeek V3 shows an elevated rate and Gemini-2.0-flash-lite approaches approximately 87%. This convergence suggests reduced reliability in generating citations from lower-resource regions, likely reflecting the underrepresentation of scholarly outputs from less affluent countries in training data. These findings emphasize the need to assess LLM performance across diverse global contexts to promote equitable and accurate knowledge generation.

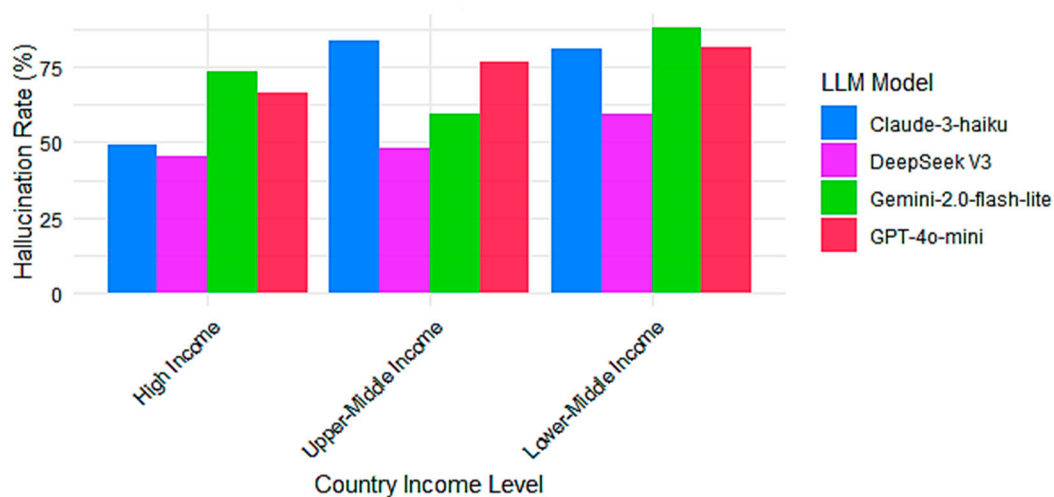


Figure 2. DOI Hallucination Rates by Country Income Level and LLM.

4.2. Geographic and Country-Specific Variations

Figure 3 displays DOI hallucination rates across countries and LLMs, revealing substantial variation across both national contexts and model differences. GPT-4o-mini and Claude-3-haiku exhibit generally elevated hallucination rates across most countries, with particularly high values in lower-income and non-Western regions such as India, Kenya, and Bangladesh. GPT-4o-mini shows consistently high hallucination rates even in high-income countries such as the United States and Germany, indicating a model-wide tendency to hallucinate regardless of national information infrastructure. Claude-3-haiku also shows elevated rates across several countries, though less extreme than Gemini. These patterns suggest that some models hallucinate broadly across contexts, rather than selectively based on geographic representation. In contrast, DeepSeek V3 and Gemini-2.0-flash-lite display greater variation in performance. DeepSeek V3 shows relatively lower hallucination rates in countries such as South Korea and Germany, demonstrating different performance patterns across geographic contexts. Gemini-2.0-flash-lite, however, demonstrates extreme behavior

in specific contexts, with hallucination rates nearing 100% for Bangladesh. Overall, the figure underscores that hallucination in LLM-generated citations is associated with both model design and country-specific information environments.

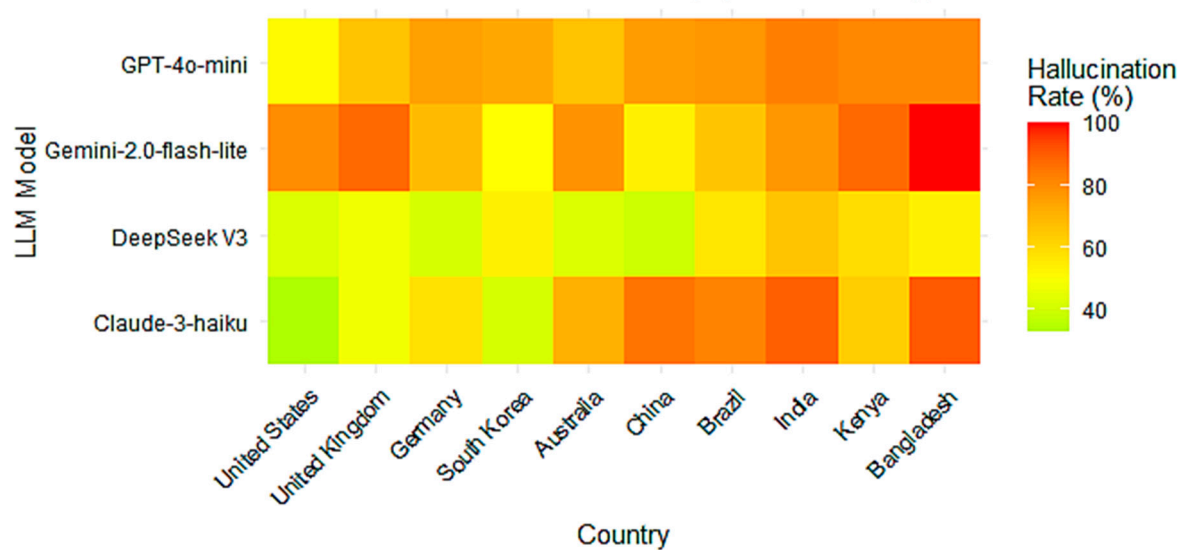


Figure 3. DOI Hallucination Rates by LLM and Country.

4.3. Hallucination Across Information Behavior Research Prompt Categories

Figure 4 demonstrates that income-level disparities in hallucination rates are consistent across all five information behavior categories, with lower-middle-income countries consistently experiencing the highest rates, followed by upper-middle-income, then high-income countries. This systematic replication provides evidence that our core findings are robust across different experimental conditions. The income-level gap varies by category, with “seeking” and “behavior” showing the most pronounced disparities (approximately 28–30 percentage point differences between high- and lower-middle-income countries), while “needs” shows the smallest gap (13 percentage points), possibly reflecting more universal approaches to information needs assessment. Notably, hallucination rates exceed 50% across all categories even in high-income contexts, indicating that geographic bias represents a fundamental challenge in LLM citation generation rather than a domain-specific phenomenon. This consistency across diverse prompt types demonstrates that systematic biases warrant attention in model development regardless of information behavior query formulation.

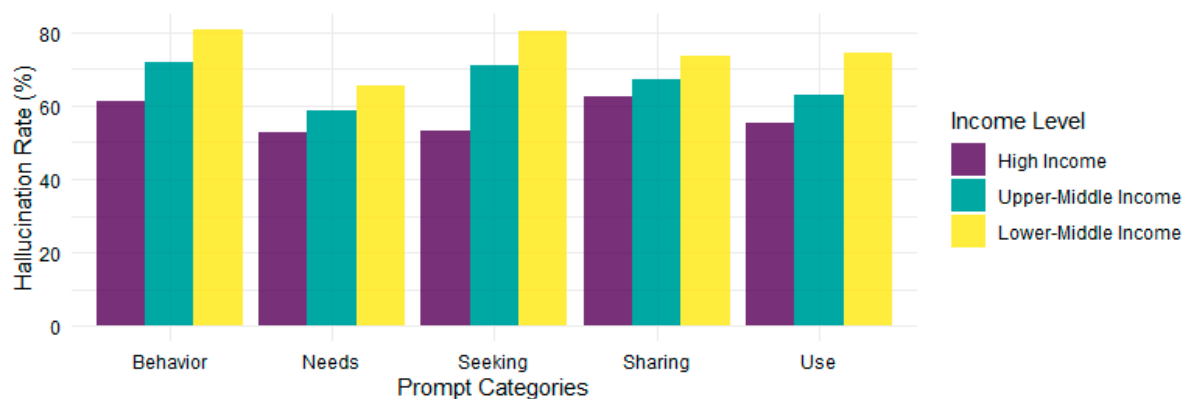


Figure 4. DOI Hallucination Rates Across Information Behavior Categories by Country Income Level.

4.4. Temporal Trends in DOI Hallucination

Figure 5 presents temporal trends in hallucination rates for LLM-generated citations, covering periods from the late 1990s to the 2020s. The data reveals distinct temporal patterns that converge toward uniformly high rates in the 2020s. High-income and upper-middle-income countries show relatively stable rates from the late 1990s through the 2010s (approximately 45–55%), followed by sharp increases to 65–75% in the 2020s. Lower-middle-income countries maintain consistently elevated rates throughout all periods, showing modest increases from around 70% to 80%. This convergence occurs primarily through dramatic increases in high- and upper-middle-income countries during the most recent period, catching up to the already-high rates that characterized lower-middle-income countries throughout all decades. Although we cannot determine the underlying mechanisms driving these patterns due to the proprietary nature of training datasets, the systematic variation by country income level and publication period suggests non-random factors influence model performance. Multiple explanations could account for these patterns, including but not limited to training data composition, model optimization differences, or other architectural factors that we cannot directly assess.

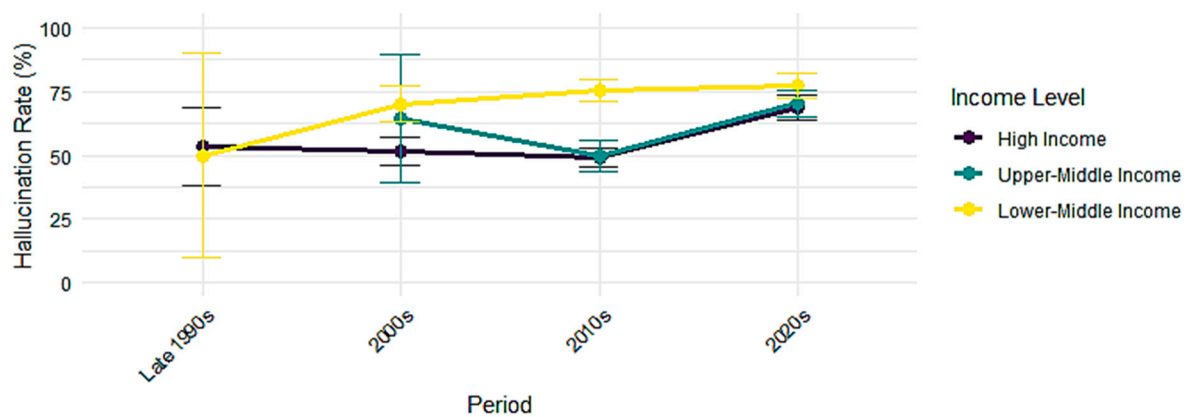


Figure 5. Temporal Trends in DOI Hallucination Rates by Country Income Level (1997–2020s).

4.5. Relationship Between Citation Completeness and DOI Hallucination Rates by Country

Figure 6 presents a counterintuitive relationship between citation completeness scores and DOI hallucination rates across the ten studied countries. Although completeness scores show limited variance (clustering between 4.85–5.0), certain countries like India and Bangladesh demonstrate a notable pattern of combining high completeness scores with elevated hallucination rates. Certain countries like India and Bangladesh demonstrate high completeness scores combined with elevated hallucination rates, contradicting the expectation that more complete-looking citations would be more accurate. However, other countries with similarly high completeness scores, such as Germany and South Korea, show relatively lower hallucination rates. India and Bangladesh demonstrate this pattern most clearly, showing both the highest completeness scores (near 5.0) and the highest hallucination rates (exceeding 80%). In contrast, countries like the United Kingdom and South Korea, despite maintaining high completeness scores, show relatively lower hallucination rates (around 48% and 50% respectively).

This counterintuitive relationship suggests that LLMs generate citations with all required bibliographic fields present but fabricate the DOIs. However, it should be noted that the accuracy of other bibliographic fields (author names, titles, journals, publication years) was not assessed in this study. The finding demonstrates that completeness—measured simply as the presence of author, title, year, journal, and DOI fields—should not be interpreted as an indicator of DOI accuracy, as models can generate structurally complete

citations with invalid DOIs. This has important implications for users who might assume that citations containing all expected components are more trustworthy than incomplete ones. The pattern is also particularly concerning for developing countries, where both high completeness scores and elevated hallucination rates indicate that LLM-generated citations may appear credible to users despite containing significant inaccuracies. The results underscore the critical importance of DOI validation beyond mere completeness assessment when evaluating LLM-generated academic citations.

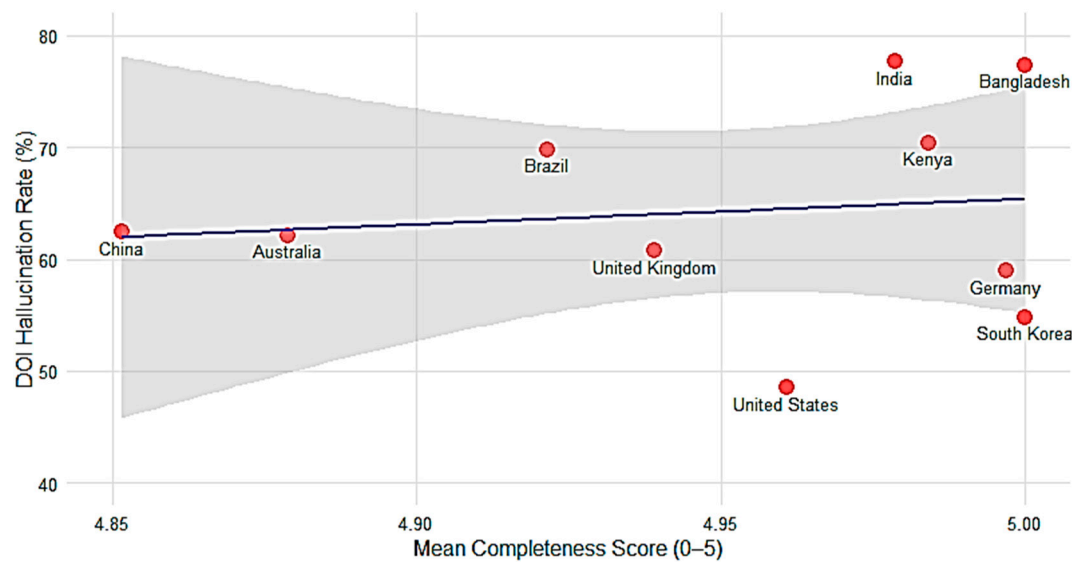


Figure 6. Relationship Between Citation Completeness and DOI Hallucination Rates Across Countries.

4.6. Statistical Analysis of DOI Validity

The statistical analysis revealed significant associations across all key variables, with most tests reporting *p*-values below 0.001 (Table 2). Model-specific factors showed the strongest association with DOI hallucination rates (Cramer’s *V* = 0.203), indicating substantial variation in citation accuracy across GPT-4o-mini, Claude-3-haiku, Gemini-2.0-flash-lite, and DeepSeek V3. Country income level was significantly associated with both DOI hallucination (*V* = 0.162) and geographic bias (*V* = 0.189), confirming that citation reliability is systematically influenced by national economic context. Temporal differences in hallucination patterns were also significant. The Kruskal–Wallis test showed that hallucination rates varied by the cited publication period (*H* = 161.74), suggesting that model performance is affected by the recency of the cited literature.

Table 2. Summary of Statistical Tests for DOI Hallucination and Geographic Bias.

Comparison	Statistical Test	N	Test Statistic	<i>p</i> Value	Effect Size	Significance
Income Level × Hallucination	Fisher’s Exact Test (simulated)	3248	Fisher’s exact	<0.001	<i>V</i> = 0.162	***
LLM × Hallucination	Fisher’s Exact Test (simulated)	3248	Fisher’s exact	<0.001	<i>V</i> = 0.203	***
Income Level × Geographic Bias	Fisher’s Exact Test (simulated)	1172	Fisher’s exact	<0.001	<i>V</i> = 0.189	***
Citation Yield × Country	Fisher’s Exact Test (simulated)	195	Fisher’s exact	1.000	<i>V</i> = 0.042	ns
Citation Yield × LLM	Fisher’s Exact Test (simulated)	195	Fisher’s exact	0.000	<i>V</i> = 0.987	***
Citation Yield × Prompt Type	Fisher’s Exact Test (simulated)	195	Fisher’s exact	1.000	<i>V</i> = 0.011	ns
Publication Period × Hallucination	Kruskal–Wallis	3087	<i>H</i> = 161.74	<0.001	η^2 = 0.052	***
Citation Age × Hallucination	Spearman Correlation	3087	ρ = −0.110	<0.001	ρ = −0.110	***

Note: *V* = Cramer’s *V*; ρ = Spearman’s correlation coefficient; η^2 = eta squared. *** *p* < 0.001; ns = not significant.

In addition to statistical significance, the effect size estimates clarify the practical relevance of these associations. Most associations were statistically significant given our large sample size (*N* > 1000). Effect sizes, however, varied considerably. The moderate values

observed for model and geographic variables ($V = 0.162$ to 0.203) suggest consistent but not dominant effects on hallucination patterns. The negative correlation between citation age and hallucination ($\rho = -0.110$) indicates that LLMs perform slightly more reliably when citing older works, which are likely better represented in training data. The correlation is weak, and the predictive value of citation age is limited. Fisher’s exact tests with Monte Carlo-simulated p -values were used in place of chi-square tests to address low expected frequencies in several cross-tabulation cells, ensuring robust significance estimates.

Citation yield analysis showed systematic bias in response completeness was model-specific rather than geography or content-related. Of the 4000 requested citations, only 3451 were received, representing a 13.7 percent shortfall. This shortfall did not vary by country ($p = 1.000$, $V = 0.042$) or prompt type ($p = 1.000$, $V = 0.011$), indicating that response volume was not biased by geographic or content-related factors. In contrast, LLMs showed an extremely large effect on citation yield ($p < 0.001$, $V = 0.987$), with some models consistently failing to return complete outputs. This finding indicates that differences in citation hallucination across countries cannot be attributed to variations in output volume alone, suggesting systematic rather than random patterns in model performance across geographic contexts.

4.7. Citation Yield Analysis Across LLMs

Citation yield analysis revealed systematic differences in models’ capacity to generate complete citation sets, as shown in Table 3. The analysis examined 195 experimental conditions (10 countries $\times$ 5 prompts $\times$ 4 models, with OpenAI having 45 conditions due to technical constraints) where each model was requested to produce 20 citations per condition.

Table 3. Citation Yield Summary by LLMs.

LLM	Conditions	Avg Citations per Condition	Response Rate (%)	Conditions with Full 20 Citations
Claude-3-haiku	50	20.3/20	101.5	33/50 (66%)
DeepSeek V3	50	20.1/20	100.3	46/50 (92%)
GPT-4o-mini	45	19.7/20	98.3	32/45 (71%)
Gemini-2.0-flash-lite	50	11.0/20	54.8	0/50 (0%)

Three models demonstrated consistently high performance across all experimental conditions. Claude-3-haiku, DeepSeek V3, and GPT-4o-mini successfully delivered near-complete citation sets, with response rates between 98–102%. Claude-3-haiku and DeepSeek V3 occasionally generated more than the requested 20 citations per condition, with some experimental runs producing 21–22 citations, resulting in response rates slightly above 100%. This over-delivery indicates robust citation generation capacity rather than measurement error.

In contrast, Gemini-2.0-flash-lite showed markedly different behavior, averaging only 11.0 citations per condition (54.8% response rate) and failing to achieve full citation sets in any of its 50 experimental conditions. This model produced fewer than 16 citations in 98% of cases, indicating systematic capacity limitations rather than random variation.

The systematic nature of these differences produced an extraordinarily large effect size (Cramer’s $V = 0.987$), indicating near-perfect predictability in model citation generation behavior. This finding demonstrates that citation yield patterns are highly consistent within each model but differ dramatically between models. For practitioners selecting LLM tools for citation generation tasks, these results provide valuable guidance about model reliability and expected output completeness, independent of citation accuracy considerations.

5. Discussion

This study demonstrates that DOI fabrication in LLM-generated citations exhibits systematic patterns that vary across model-specific factors, country income levels, and geographic contexts. Among all tested variables, model-specific factors showed the strongest association with hallucination, followed by country income level. This model-specific variation extends beyond hallucination patterns to fundamental differences in citation generation capacity, with three models consistently delivering near-complete citation sets while Gemini-2.0-flash-lite systematically under-delivered citations across all experimental conditions. The significantly higher hallucination rates observed in lower-income countries, compared to high-income contexts, represent a pattern consistent with systematic performance differences across geographic contexts. Gemini-2.0-flash-lite's complete citation failure for Bangladesh, characterized by a near-100% hallucination rate, represents an extreme example of differential model performance across national contexts. These models consistently generate citations with complete formatting across all contexts, though accuracy varies significantly by geographic region.

Temporal analysis shows a key limitation in current LLMs. Hallucination rates rise markedly for citations dated in the 2020s across all income levels, a pattern consistent with challenges in generating citations for recent literature. A modest negative correlation between citation age and hallucination rates supports this interpretation. More notably, a positive correlation is observed between citation completeness and hallucination. Countries such as India and Bangladesh exhibit both the highest average completeness scores and the highest hallucination rates, suggesting that citations may appear structurally complete while containing fabricated DOIs. This pattern indicates that LLMs tend to prioritize surface-level features of citations, such as formatting and field inclusion, even in the absence of factual grounding. When completeness is treated as the main objective, models generate citations that look correct but rely on invalid identifiers. For instance, an LLM may generate a citation containing all required bibliographic fields while providing an invalid DOI. In such cases, completeness no longer reflects accuracy, which illustrates Goodhart's Law: once a measure becomes the target, it loses its validity as a measure (Elton, 2004).

The relationship between citation completeness and hallucination rates reveals the complexity of studying proprietary LLMs. Both high completeness and high hallucination in countries like India and Bangladesh represent a pattern where surface-level formatting appears complete while DOI accuracy remains problematic, but the opacity of commercial training datasets prevents direct assessment of such mechanisms. However, this pattern remains practically significant regardless of causal pathways: it demonstrates that surface-level citation features cannot be relied upon as indicators of accuracy. This finding challenges user assumptions about citation reliability and underscores the need for systematic DOI validation rather than reliance on formatting cues.

These findings have direct implications for global research equity. All four models show elevated hallucination rates across contexts, suggesting shared architectural tendencies that prioritize fluent generation over epistemic accuracy. The consistently high hallucination rates in lower-income countries, even when valid DOIs are generated, exacerbate disparities in scholarly visibility. As noted by Chavarro et al. (2025), limited DOI infrastructure in these regions compounds the problem, creating feedback loops that further marginalize non-Western scholarship. Addressing these challenges requires more than technical refinement. It demands region-aware training strategies, integration of real-time DOI validation, and a shift toward design principles that privilege reliability and contextual appropriateness over output fluency.

In sum, our findings suggest that LLM citation hallucination exhibits systematic patterns shaped by model design and training data limitations rather than random error.

The intersection of high completeness scores with elevated hallucination rates in lower-income countries reveals a fundamental tension between surface plausibility and factual accuracy. Rather than acknowledging uncertainty or gaps in their training data, LLMs consistently choose to fabricate authoritative-sounding citations, with this tendency most pronounced where scholarly infrastructure is least available for verification. This pattern suggests that the current generation of LLMs may be particularly unreliable as research tools in the very contexts where they could provide the greatest benefit—regions with limited access to comprehensive academic databases and verification resources.

6. Limitations

This study has several important limitations that constrain the interpretation and generalizability of the findings. First, causal mechanisms for observed geographic patterns cannot be established due to proprietary training data constraints, restricting conclusions to systematic associations rather than explanations. Second, the DOI validation procedure focused only on resolution success and geographic relevance based on metadata, without verifying whether valid DOIs accurately matched the cited content. This likely results in a substantial underestimation of citation fabrication. Third, the geographic scope is a critical limitation: low-income countries were entirely excluded, and only three lower-middle-income countries were included compared to seven upper-middle- and high-income countries. This sampling bias toward better-resourced contexts limits generalizability to regions where bias effects are likely most pronounced. Fourth, the findings reflect specific model versions tested during a fixed period; citation behavior may differ under later updates or alternative configurations. Fifth, results are limited to information behavior research within Library and Information Science, and patterns may not extend to other academic domains. Finally, hallucination was operationalized solely as DOI resolution failure, which may overlook other forms of fabrication such as incorrect authorship, publication years, or mismatched bibliographic details.

Collectively, these limitations indicate that the study likely provides a conservative estimate of geographic bias in LLM citation generation. The exclusion of low-income countries and the narrow focus on DOI validity suggest that the true extent of disparities may be greater than documented here.

7. Conclusions

This study provides evidence that LLM-generated citations show systematic patterns of higher hallucination rates in lower-income countries, supporting concerns about potential algorithmic bias in scholarly communication tools. Hallucination rates consistently exceed those observed in high-income contexts, indicating systematic rather than random performance differences across geographic settings. Model-specific factors showed the strongest association with hallucination, revealing significant variation in reliability across implementations. The frequent generation of fabricated citations in underrepresented regions, often accompanied by high completeness, reflects systematic tendencies to prioritize surface-level formatting over factual accuracy. Such patterns risk reinforcing global inequities in scholarly communication by producing unreliable references precisely in contexts where verifiable academic information is least accessible.

Future research should broaden sampling to include low-income countries, apply more comprehensive validation beyond DOI resolution to capture errors in authorship, publication years, and bibliographic details, and track how hallucination patterns shift with subsequent model updates. Comparative studies in disciplines outside Library and Information Science are also needed to assess the generalizability of the findings across academic domains.

Author Contributions: Conceptualization, E.K.; Data curation, E.K. and F.K.; Formal analysis, E.K. and F.K.; Funding acquisition, E.K.; Investigation, E.K.; Methodology, E.K., F.K. and S.M.; Project administration, E.K.; Resources, E.K.; Software, E.K.; Supervision, E.K.; Validation, E.K., F.K. and S.M.; Visualization, E.K.; Writing—original draft, E.K.; Writing—review & editing, E.K. All authors have read and agreed to the published version of the manuscript.

Funding: This research was supported by the Bisa Research Grant of Keimyung University in 2024 (Project No. 20240535).

Data Availability Statement: The datasets and scripts used in this study are openly available in a public GitHub repository: https://github.com/egkim68/citation_bias_geography (accessed 8 September 2025). The repository contains raw citation datasets generated from all four LLMs, including citations with DOIs, together with the scripts used for citation generation (LLM API) and DOI verification (CrossRef API).

Conflicts of Interest: The authors declare no conflicts of interest.

References

- Bornmann, L., & Daniel, H. D. (2008). What do citation counts measure? A review of studies on citing behavior. *Journal of Documentation*, 64(1), 45–80. [CrossRef]
- Bridges, J. M. (2024). Computerized diagnostic decision support systems—A comparative performance study of Isabel Pro vs. ChatGPT4. *Diagnosis*, 11(3), 250–258. [CrossRef]
- Chavarro, D., Alperin, J. P., & Willinsky, J. (2025). A caminho da indexação universal: OpenAlex e open journal systems. *SciELO preprints*. [CrossRef]
- Cheng, A., Calhoun, A., & Reedy, G. (2025). Artificial intelligence-assisted academic writing: Recommendations for ethical use. *Advances in Simulation*, 10, 22. [CrossRef] [PubMed]
- Daquino, M., Peroni, S., Shotton, D., Colavizza, G., Ghavimi, B., Lauscher, A., & Zumstein, P. (2020). The OpenCitations data model. In *International semantic web conference* (pp. 447–463). Springer International Publishing.
- Earhart, A. E., Risam, R., & Bruno, M. (2021). Citational politics: Quantifying the influence of gender on citation in digital scholarship in the humanities. *Digital Scholarship in the Humanities*, 36(3), 581–594. [CrossRef]
- Elton, L. (2004). Goodhart's law and performance indicators in higher education. *Evaluation & Research in Education*, 18(1–2), 120–128. [CrossRef]
- Frosolini, A., Catarzi, L., Benedetti, S., Latini, L., Chisci, G., Franz, L., Gennaro, P., & Gabriele, G. (2024). The role of Large Language Models (LLMs) in providing triage for maxillofacial trauma cases: A preliminary study. *Diagnostics*, 14(8), 839. [CrossRef]
- Gathama, N., Mwaurah, N. W., Miriti, A., & Mwangi, N. (2025). Accuracy of references in the kenya eye health journal volume 1 number 1. *Kenya Eye Health Journal*, 2(1), 12–15.
- Hendricks, G., Tkaczyk, D., Lin, J., & Feeney, P. (2020). Crossref: The sustainable source of community-owned scholarly metadata. *Quantitative Science Studies*, 1(1), 414–427. [CrossRef]
- Hyland, K. (1999). Academic attribution: Citation and the construction of disciplinary knowledge. *Applied Linguistics*, 20(3), 341–367. [CrossRef]
- Jannot, A.-S., Agoritsas, T., Gayet-Ageron, A., & Perneger, T. V. (2013). Citation bias favoring statistically significant studies was present in medical research. *Journal of Clinical Epidemiology*, 66(3), 296–301. [CrossRef] [PubMed]
- Jedrzejczak, W. W., Skarzynski, P. H., Raj-Koziak, D., Sanfins, M. D., Hatzopoulos, S., & Kochanek, K. (2024). ChatGPT for tinnitus information and support: Response accuracy and retest after three and six months. *Brain Sciences*, 14(5), 465. [CrossRef]
- Kumari, K. S., Shree, S. U., & Calarany, C. (2024, August 16–19). *Dynamic region-aware fine-tuning: Enhancing geographic inclusivity in large language models*. 2024 IEEE 16th International Conference on Advanced Communication Technology (pp. 1–8), Enshi, China.
- Liu, Z., Janowicz, K., Majic, I., Shi, M., Fortacz, A., Karimi, M., Mai, G., & Currier, K. (2025). Operationalizing geographic diversity for the evaluation of AI-generated content. *Transactions in GIS*, 29(3), e70057. [CrossRef]
- Manvi, R., Khanna, S., Burke, M., Lobell, D., & Ermon, S. (2024). Large language models are geographically biased. *arXiv*, arXiv:2402.02680. [CrossRef]
- Moayeri, M., Tabassi, E., & Feizi, S. (2024, June 3–6). *WorldBench: Quantifying geographic disparities in LLM factual recall*. The 2024 ACM Conference on Fairness, Accountability, and Transparency (pp. 1211–1228), Rio de Janeiro, Brazil. [CrossRef]
- Mugaanyi, J., Cai, L., Cheng, S., Lu, C., & Huang, J. (2024). Evaluation of large language model performance and reliability for citations and references in scholarly writing: Cross-disciplinary study. *Journal of Medical Internet Research*, 26, e52935. [CrossRef]

- Ng, J. Y., Liu, H., Masood, M., Farin, R., Messih, M., Perez, A., Aalbersberg, I. J., Alperin, J., Bryson, G. L., Chen, Q., Ehrlich, A., Iorio, A., Meester, W. J. N., Willinsky, J., Grudniewicz, A., Cobo, E., Cranston, I., Cress, P. E., Gunn, J., . . . Moher, D. (2025). Publisher preferences for a journal transparency tool: A modified three-round Delphi study. *F1000Research*, *13*, 915. [CrossRef]
- Ng, J. Y., Liu, H., Masood, M., Kochhar, J., Moher, D., Ehrlich, A., Iorio, A., & Cobey, K. D. (2024). A mixed-methods survey and focus group study to understand researcher and clinician preferences for a Journal Transparency Tool. *Scientific Reports*, *14*, 26626. [CrossRef]
- Ramesh, M., Sun, Z., Li, Y., Zhang, L., Annam, S. K., Fang, H., & Tong, D. (2025). Assessing WildfireGPT: A comparative analysis of AI models for quantitative wildfire spread prediction. *Natural Hazards*, *121*, 13117–13130. [CrossRef]
- Ruff, E. F., Franz, J. L., & West, J. K. (2024). Using ChatGPT for method development and green chemistry education in upper-level laboratory courses. *Journal of Chemical Education*, *101*(8), 3224–3232. [CrossRef]
- Teixeira da Silva, J. A., & Nazarovets, S. (2022). Publication history: A double-DOI-based method for storing and/or monitoring information about published and corrected academic literature. *Journal of Scholarly Publishing*, *53*(2), 85–108. [CrossRef]
- World Bank. (2024). *World Bank Country and Lending Groups*. World bank data help desk. Available online: <https://datahelpdesk.worldbank.org/knowledgebase/articles/906519-world-bank-country-and-lending-groups> (accessed on 8 September 2025).
- Zhang, M., & Zhao, T. (2025). Citation accuracy challenges posed by large language models. *JMIR Medical Education*, *11*, e72998. [CrossRef] [PubMed]

Disclaimer/Publisher’s Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.