The Illusion of Validity: Structural Hallucinations in LLM-Generated Scholarly Citations

Anonymous ACL submission

Abstract

Large language models frequently hallucinate scholarly citations, yet existing evaluations often rely on binary existence checks, such as whether a DOI resolves or a title appears in a bibliographic database. We argue that such checks underestimate structural citation corruption. We introduce a structural evaluation framework that decomposes generated citations into five metadata fields—title, authors, year, venue, and DOI—and validates each field against CrossRef and Semantic Scholar records. We apply the framework to citations generated by five LLMs across English and Bangla prompts (n=1,210). We identify **chimeric hallucination**, a failure mode in which metadata fields from multiple real papers are combined into a single structurally inconsistent citation that may still pass binary validation. Our taxonomy classifies citations as VALID, CHIMERIC, PARTIAL CORRUPTION, or FABRICATED. Under our evaluation setting, 41% of citations are chimeric, while only 6% are fully structurally valid. Results remain stable across threshold settings, and Bangla prompts show a modestly higher observed chimeric rate. These findings suggest that citation validity should be evaluated as cross-field structural consistency rather than binary existence alone.

1 Introduction

Consider the following LLM-generated citation:

> **Title:** "Neural Machine Translation for Low-Resource Languages: A Survey" ← Paper A
> **Authors:** Rahman, M., Islam, T., & Das, A. ← Paper B
> **DOI:** 10.18653/v1/2020.acl-main.485 ← Paper C

This citation passes every standard hallucination check. The DOI resolves. The title exists in CrossRef. A conventional binary classifier would label it VALID. Yet it is structurally corrupt: the title, authors, and DOI each originate from *different real papers*, stitched together into a plausible but fictitious reference. This phenomenon of reference fabrication represents a major challenge in LLM-based scholarly writing (Alkaissi and McFarlane, 2023; Walters and Wilder, 2023).

While automated verification pipelines have advanced citation checking (Agrawal et al., 2024; Liu et al., 2023), they predominantly produce a binary *existence* signal: does the paper exist in a bibliographic database? This binary framing collapses rich structural metadata. A citation is not a single fact but a *tuple of claims*: the title, authors, venue, year, and DOI must all resolve to the same underlying work. When these fields are independently valid but mutually inconsistent, existence-based checks fail silently.

We address this gap by proposing a *structural* evaluation framework that verifies metadata fields independently and audits their cross-field consistency. This field-level decomposition conceptually extends atomic fact verification (Min et al., 2023) to structured bibliographic metadata. It reveals systematic corruptions that bypass conventional checks. We evaluate our framework across five state-of-the-art LLMs, examining 1,210 citations generated under both English and Bangla scholarly prompt contexts (n=1,210). This setup lets us audit structural corruption dynamics in both high- and low-resource language settings (e.g., Bangla; Bhattacharjee et al., 2023).

Our contributions are:

1. A structural evaluation framework that validates individual metadata fields against multiple databases and audits cross-field consistency.

2. A four-category taxonomy identifying chimeric hallucinations, supported by deterministic classification rules and a human validation protocol ($n = 50$, Appendix F).

3. An empirical evaluation across five LLMs showing that, under our evaluation setting,

only 6% of generated citations are structurally valid, with results robust to threshold sweeps.

2 Related Work

Citation Hallucination and Verification. Prior verification systems primarily frame citation hallucination as a binary existence check: resolving DOIs or matching titles against online indexes (Agrawal et al., 2024; Liu et al., 2023). CiteAudit (Yuan et al., 2026) advances this by incorporating semantic LLM-as-a-judge pipelines for context validation, but incurs significant API costs and latency. Crucially, CiteAudit verifies whether a cited paper *supports* the citing claim—a content-level check—whereas our framework operates at the *structural metadata level*, verifying whether the citation tuple itself is internally consistent, independent of citing context. A citation can be semantically appropriate yet structurally chimeric (e.g., correct title but wrong authors). We present a lightweight, deterministic, rule-based auditor that flags cross-field corruption without LLM intervention, scaling to thousands of citations in seconds.

Factuality Evaluation. FActScore (Min et al., 2023) decomposes text into atomic facts verified independently; SelfCheckGPT (Manakul et al., 2023) detects hallucinations without external knowledge. Our work extends atomic verification to structured bibliographic metadata, where fields are not independent atoms but *jointly constrained*: title, authors, year, venue, and DOI must resolve to the same underlying work. Our framework exploits this joint constraint through cross-field consistency auditing.

Multilingual Hallucination. Recent benchmarks show that hallucination rates increase in low-resource languages (Ahuja et al., 2023). Surveys of hallucination in NLG (Ji et al., 2023; Zhang et al., 2025) comprehensively categorize failure modes but do not address structured scholarly metadata. To our knowledge, no prior work has conducted a field-level structural analysis of citation hallucination in a low-resource language setting; our Bangla analysis provides an initial exploratory data point.

3 Methodology

3.1 Prompt Design and Data Collection

We design 100 scholarly prompts in English and Bangla (Appendix G), spanning 10 NLP topics and 5 prompt types (e.g., literature reviews and surveys). Each prompt requests citations with structured metadata fields: title, authors, year, venue, and DOI.

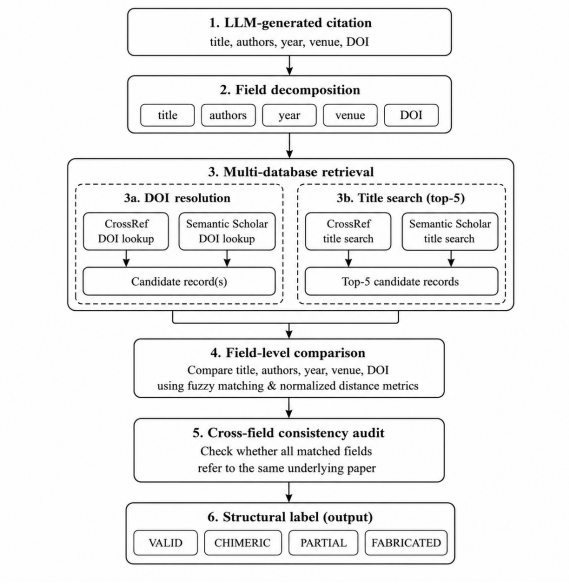

Figure 1: Overview of the proposed multi-database structural validation framework. Each generated citation is decomposed into metadata fields and validated through candidate retrieval, field-level matching, and cross-field consistency auditing before assigning a structural label.

We evaluate five representative LLMs: Meta-Llama-3.3-70B-Instruct, Google-Gemma-3-12B-IT, Meta-Llama-4-Scout-17B-16E-Instruct, GPT-OSS-120B, and DeepSeek-V3.2. All inference was performed through hosted SambaNova Cloud APIs.[1] We use deterministic decoding ($T = 0.0$) under a closed-book setting for reproducibility. Across models, citation yield varies due to generation and parsing failures. We retain all successfully parsed citations, yielding a final dataset of n=1,210 citations (Llama-3.3: 740, Gemma-3: 190, Llama-4: 180, GPT-OSS: 70, DeepSeek-V3.2: 30), followed by standard field normalization (Appendix E).

3.2 Multi-Database Structural Validation

Unlike binary approaches that rely only on DOI resolution or title existence, our framework evaluates citations as structured metadata tuples. Figure 1 summarizes the validation pipeline.

For each citation, we independently validate title, authors, year, venue, and DOI using CrossRef (Hendricks et al., 2020) and Semantic Scholar (Kinney

[1]Models were queried under consistent system prompts via SambaNova Cloud APIs (SambaNova Systems, 2026).

Category	Definition
VALID	All fields match one real paper.
CHIMERIC	Fields are drawn from ≥2 real papers, including DOI–title drift.
PARTIAL	One real paper is identified, but one or more fields are incorrect for that paper.
FABRICATED	No matching paper is found by DOI resolution or title search.

Table 1: Structural hallucination taxonomy. CHIMERIC subsumes cases where the DOI points to Paper A while the title matches Paper B, as these represent cross-field inconsistency by definition.

Category	n	%	95% CI
VALID	69	6%	[4.5%, 7.2%]
CHIMERIC	493	41%	[38.0%, 43.5%]
PARTIAL	240	20%	[17.7%, 22.2%]
FABRICATED	408	34%	[31.1%, 36.4%]
Binary existence-valid		66%	
Structurally valid		6%	

Table 2: Distribution of structural hallucination categories with Wilson 95% confidence intervals (n=1,210). Binary existence-based validation substantially overestimates citation validity relative to field-level structural consistency checks.

et al., 2023). The pipeline consists of four stages: (1) **DOI resolution**, retrieving candidate records from both databases; (2) **title search**, retaining the top-5 title matches; (3) **field-level comparison**, matching citation fields using fuzzy and normalized distance metrics (Table 3); and (4) **cross-field consistency auditing**, determining whether matched fields correspond to the same underlying paper. Citations matching multiple papers are labeled structurally inconsistent.

3.3 Hallucination Taxonomy

Based on the validation results, we classify each citation into one of four mutually exclusive, collectively exhaustive categories (Table 1).

3.4 Classification Decision Rules

Fine-grained labels (e.g., AUTHORPERMUTATION, DOIDRIFT) are merged into Table 1. We apply deterministic rules in sequence. If the DOI resolves, all fields are compared to the DOI record: full matches are VALID, title-only matches are PARTIAL, and DOI–title mismatches are CHIMERIC. Cases where the title similarity is below 60% but authors or venue match the DOI paper are also labeled CHIMERIC. To avoid false positives from version skew or venue variants, citations with a matching preprint title or ≥90% author overlap are downgraded to VALID/PARTIAL. If the DOI does not resolve, we search by title: full matches are VALID, author mismatches are PARTIAL, and remaining cases are FABRICATED. Truncated author lists are treated as PARTIAL unless the authors come from a second unrelated paper.

3.5 Human Validation

To validate the automated classifier, we sample n=50 citations stratified across the four taxonomy categories (approximately 12–13 per class) from the full dataset. Two NLP-trained annotators independently label each citation by reviewing retrieved bibliographic metadata from CrossRef and Semantic Scholar. Disagreements are resolved through adjudication to produce a final gold label. We report Cohen's κ (Cohen, 1960) alongside macro-averaged precision and recall between automated and adjudicated labels (§4.5). Additional annotation details are provided in Appendix F.

4 Results

4.1 Overall Distribution

Table 2 summarizes results across 1,210 citations. Chimeric hallucination is the largest category (41%, 95% Wilson CI: [38.0%, 43.5%] (Wilson, 1927)), followed by fabricated citations (34%, CI: [31.1%, 36.4%]); only 6% (CI: [4.5%, 7.2%]) are structurally valid.

Binary DOI/title checks mark 66% of citations as valid (CI: [63.6%, 68.9%]), but structural validation reduces this to 6%. Thus, binary checks may substantially overestimate citation validity in similar closed-book generation settings.

4.2 Threshold Sensitivity

To address potential concerns about threshold sensitivity, we sweep three key parameters: title match threshold (75–95%), author Jaccard threshold (0.3–0.8), and chimeric dissimilarity threshold (40–70%). Across all configurations, chimeric hallucination remains the dominant or second-most prevalent category, and the structural validity rate never exceeds 10%. Full results are reported in Appendix B.

4.3 Chimeric Hallucination Analysis

Since chimeric hallucinations constitute the largest share of citation corruptions (41%, Table 2), we analyze their manifestation. They are difficult to

detect through binary validation, as they combine real metadata fields from different sources (e.g. DOI resolves to Paper A, title matches Paper B, authors match Paper C). Per-model breakdowns and a visual distribution heatmap are reported in Appendix C (Table 4, Figure 3). Model-specific validation rates vary with formatting yield; notably, DeepSeek-V3.2 (n=30) has a wide 95% Wilson CI of [33.2%, 66.8%], limiting interpretability.

4.4 Language Comparison

As an exploratory analysis, we compare citation distributions across English (n=960) and Bangla (n=250) prompts. Bangla prompts yield a modestly higher chimeric rate (45% vs. 40%). A χ^2 test of independence (χ^2=8.16, p=0.0421) suggests a possible association, though unbalanced group sizes and absence of multiple-comparison correction limit this to a preliminary observation. Full breakdowns appear in Table 5 (Appendix D).

4.5 Human Validation Results

We evaluate our automated classifier on a stratified gold set ($n = 50$). The classifier achieves high agreement with human annotators (Cohen's κ=0.92, Table 6 in Appendix F), reaching 96% precision and 96% recall, confirming the automated pipeline's reliability. The full confusion matrix and error analysis are presented in Table 7 (Appendix F).

5 Discussion

Implications for Evaluation. Our findings suggest that citation validity is fundamentally a structured consistency problem rather than a binary retrieval problem. Evaluation methods based solely on DOI resolution or title matching may systematically overestimate citation reliability by overlooking cross-field inconsistencies.

Limitations. Our framework relies on CrossRef and Semantic Scholar, which provide broad coverage of CS/NLP publications; however, unindexed or region-specific works may still be incorrectly classified as FABRICATED. To reduce false CHIMERIC assignments arising from version skew (e.g., preprint vs. proceedings), we incorporate title-similarity and author-overlap deduplication rules (§3.4). We further restrict our evaluation to closed-book, deterministic decoding (temperature=0.0) to establish reproducible baselines; retrieval-augmented and higher-temperature generation settings remain important directions for future work. Our multilingual analysis concerns prompt language rather than generated citation language, which remains predominantly English across both settings.

Reproducibility. To support reproducibility, we plan to release the structural validation code, anonymized citation dataset, prompt templates, and analysis scripts upon acceptance, with explicit license information. External services and databases, including CrossRef, Semantic Scholar, and SambaNova Cloud, were used for research evaluation under their respective terms; released artifacts are intended for research and reproducibility use.

6 Conclusion.

We presented a structural evaluation framework for citation hallucination that moves beyond binary existence-based validation. Across our evaluated setting, structurally inconsistent yet superficially plausible citations—particularly CHIMERIC cases combining metadata from multiple real papers—occur frequently across models and prompts. Our findings suggest that binary validation substantially overestimates citation correctness by overlooking cross-field inconsistencies. We hope this work encourages future evaluation frameworks to treat citation validity as a problem of structural consistency rather than existence alone.

Ethical Considerations

This work evaluates LLM APIs and validates generated citations against public bibliographic databases, including CrossRef and Semantic Scholar. The collected data consists of model-generated citation metadata and may include public author names, but contains no private user data, sensitive personal information, or offensive content. Released data will be anonymized where appropriate and limited to research reproducibility. Our findings highlight risks for academic integrity when trusting LLM-generated scholarly references.

References

Ayush Agrawal, Mirac Suzgun, Lester Mackey, and Adam Kalai. 2024. Do language models know when they're hallucinating references? In *Findings of the Association for Computational Linguistics: EACL 2024*, pages 912–928.

Kabir Ahuja, Harshita Diddee, Rishav Hada, Millicent Ochieng, Krithika Ramesh, Prachi Jain, Akshay Nambi, Tanuja Ganu, Sameer Segal, Mohamed Ahmed, and 1 others. 2023. Mega: Multilingual evaluation of generative ai. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 4232–4267.

Hussam Alkaissi and Samy I McFarlane. 2023. Artificial hallucinations in chatgpt: implications in scientific writing. *Cureus*, 15(2).

Max Bachmann. 2020. Rapidfuzz: Rapid fuzzy string matching. Software library.

Abhik Bhattacharjee, Tahmid Hasan, Wasi Ahmad, and Rifat Shahriyar. 2023. Banglanlg and banglat5: Benchmarks and resources for evaluating low-resource natural language generation in bangla. In *Findings of the Association for Computational Linguistics: EACL 2023*, pages 726–735.

Jacob Cohen. 1960. A coefficient of agreement for nominal scales. *Educational and psychological measurement*, 20(1):37–46.

Ginny Hendricks, Dominika Tkaczyk, Jennifer Lin, and Patricia Feeney. 2020. Crossref: The sustainable source of community-owned scholarly metadata. *Quantitative Science Studies*, 1(1):414–427.

Ziwei Ji, Nayeon Lee, Rita Frieske, Tiezheng Yu, Dan Su, Yan Xu, Etsuko Ishii, Ye Jin Bang, Andrea Madotto, and Pascale Fung. 2023. Survey of hallucination in natural language generation. *ACM computing surveys*, 55(12):1–38.

Rodney Kinney, Chloe Anastasiades, Russell Authur, Iz Beltagy, Jonathan Bragg, Alexandra Buraczynski, Isabel Cachola, Stefan Candra, Yoganand Chandrasekhar, Arman Cohan, and 1 others. 2023. The semantic scholar open data platform. *arXiv preprint arXiv:2301.10140*.

Nelson F Liu, Tianyi Zhang, and Percy Liang. 2023. Evaluating verifiability in generative search engines. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 7001–7025.

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

Sewon Min, Kalpesh Krishna, Xinxi Lyu, Mike Lewis, Wen-tau Yih, Pang Koh, Mohit Iyyer, Luke Zettlemoyer, and Hannaneh Hajishirzi. 2023. Factscore: Fine-grained atomic evaluation of factual precision in long form text generation. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 12076–12100.

SambaNova Systems. 2026. Sambacloud models overview and specifications. Online documentation. Accessed 2026-05-24.

William H Walters and Esther Isabelle Wilder. 2023. Fabrication and errors in the bibliographic citations generated by chatgpt. *Scientific Reports*, 13(1):14045.

Edwin B Wilson. 1927. Probable inference, the law of succession, and statistical inference. *Journal of the American Statistical Association*, 22(158):209–212.

Zhengqing Yuan, Kaiwen Shi, Zheyuan Zhang, Lichao Sun, Nitesh V Chawla, and Yanfang Ye. 2026. Citeaudit: You cited it, but did you read it? a benchmark for verifying scientific references in the llm era. *arXiv preprint arXiv:2602.23452*.

Yue Zhang, Yafu Li, Leyang Cui, Deng Cai, Lemao Liu, Tingchen Fu, Xinting Huang, Enbo Zhao, Yu Zhang, Yulong Chen, and 1 others. 2025. Siren's song in the ai ocean: A survey on hallucination in large language models. *Computational Linguistics*, 51(4):1373–1418.

A Validation Thresholds

Table 3 reports the fuzzy matching thresholds used in the structural validation pipeline. Thresholds were selected based on preliminary analysis and validated via sensitivity analysis (§4.2).

Field	Method	Threshold
Title	Token-sort ratio	$\geq$85%
Authors	Jaccard (last names)	$\geq$0.6
Venue	Abbreviation-aware fuzzy	$\geq$75%
Year	Absolute difference	$\leq$1
DOI	Exact match (normalized)	100%
Chimeric	Title dissimilarity	$<$60%

Table 3: Validation thresholds for field-level comparison.

B Threshold Sensitivity Analysis

Figure 2 shows category proportions under varying threshold settings. Chimeric hallucination remains the dominant category across all tested configurations.

C Per-Model Breakdown

Table 4 reports the hallucination category distribution for each of the five models. Figure 3 visualizes the distribution across all models.

D Language Comparison Details

Table 5 breaks down the citation hallucination categories across English and Bangla prompts.We observe a modestly higher proportion of CHIMERIC citations under Bangla prompting conditions. However, this analysis should be interpreted as exploratory rather than causal, as the prompt-language conditions exhibit unequal citation yields

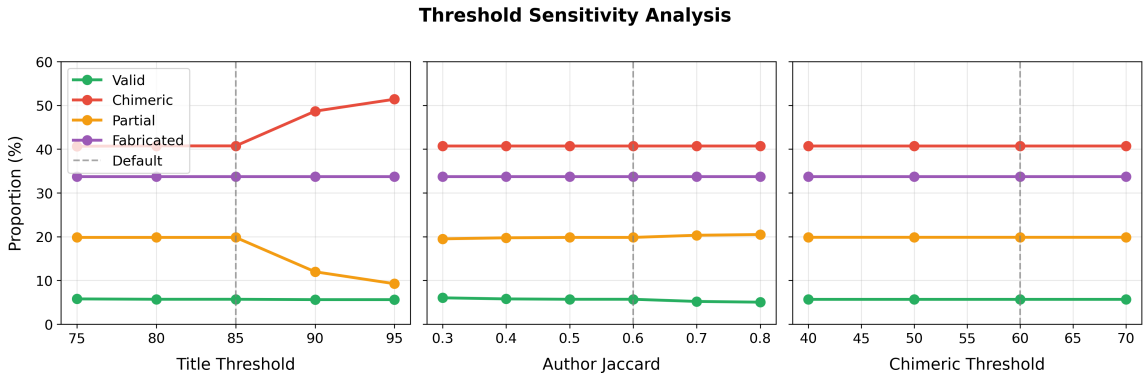

Figure 2: Threshold sensitivity analysis. Category proportions remain stable across parameter sweeps, confirming robustness of the chimeric hallucination finding.

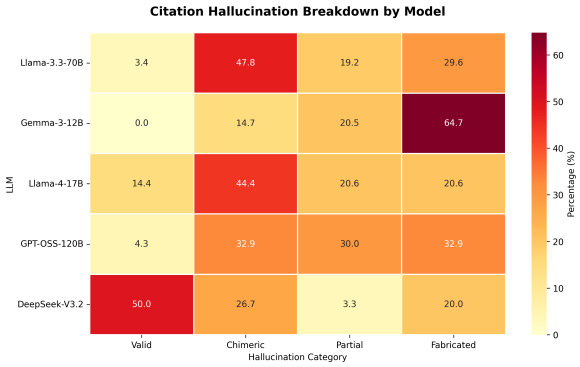

Figure 3: Hallucination category distribution across five LLMs (merged 4-category taxonomy). Chimeric hallucination is prevalent across all models.

Model	n	Val.	Val. 95% CI	Chim.	Part.	Fabr.
Llama-3.3-70B	740	3%	[2.3%, 4.9%]	48%	19%	30%
Gemma-3-12B	190	0%	[0.0%, 2.0%]	15%	21%	65%
Llama-4-17B	180	14%	[10.1%, 20.3%]	44%	21%	21%
GPT-OSS-120B	70	4%	[1.5%, 11.9%]	33%	30%	33%
DeepSeek-V3.2	30	50%	[33.2%, 66.8%]	27%	3%	20%

Table 4: Per-model hallucination distribution (%). Unequal sample sizes (n) reflect natural variation in citation yield and parseability under identical prompting conditions. Wilson 95% confidence intervals highlight the instability of low-yield settings (e.g., DeepSeek-V3.2 at $n = 30$).

and the evaluated bibliographic databases predominantly index English-language metadata.

E Field Normalization

Titles: lowercased, punctuation stripped, whitespace collapsed; RapidFuzz (Bachmann, 2020) token-sort ratio ≥85% for a match. **Authors:** last names extracted from "Last, First" or final token; Jaccard overlap ≥0.6 (truncated lists on the same paper often still match). **Venues:** abbreviation table (EMNLP, ACL, NeurIPS, etc.) expanded before fuzzy match ≥75%. **Years:** ±1 tolerance. **DOIs:** lowercased, URL prefixes removed. Bangla

Lang.	n	Val.	Chim.	Part.	Fabr.
English	960	6%	40%	19%	35%
Bangla	250	4%	45%	23%	28%

Table 5: Category distribution by prompt-language condition (%). Generated citation metadata remained predominantly English across both settings.

prompts do not alter normalization; generated metadata is compared as returned by the model (typically English).

F Human Annotation Details

Annotation Protocol. Annotators independently labeled a stratified sample of LLM-generated citations using the full annotation guide provided in the supplementary material. The annotators were NLP-trained volunteers recruited through the authors' academic network; no crowdsourcing platform or paid external participants were used. They were informed that their annotations would be used for research validation, and no annotator-identifying information is included in the released data. For each citation, they inspected CrossRef and Semantic Scholar search results and assigned one of four labels: VALID if all fields matched one paper, CHIMERIC if fields came from multiple papers, PARTIAL if one identifiable paper contained incorrect fields, and FABRICATED if no matching paper was found. Disagreements were adjudicated to construct the final gold dataset.

Metrics & Sample Size. Table 6 reports validation metrics on the stratified gold set ($n = 50$). Inter-annotator agreement is high ($\kappa = 0.92$), with representation across all taxonomy categories. Our automated classifier achieves 96% macro precision and 96% macro recall against adjudicated labels. The two observed misclassifications arise from

category-boundary edge cases (author-format variation; non-indexed venue), consistent with known database coverage limitations.

Metric	Value
Sample size (stratified)	50
Cohen's κ (inter-annotator)	0.92
Macro precision (auto vs. gold)	96%
Macro recall (auto vs. gold)	96%

Table 6: Human validation metrics showing strong inter-annotator agreement and automated classifier performance on the gold set.

Error Analysis & Confusion Matrix. Table 7 presents the confusion matrix of the automated classifier against the adjudicated human gold labels. The classifier demonstrates high robustness. The minor misclassifications are: (1) one 'VALID' citation was classified as 'PARTIAL' because of a minor author formatting difference (middle initials omitted in LLM output), and (2) one 'PARTIAL' citation was classified as 'FABRICATED' due to a non-indexed conference venue variation.

Gold (Human) \ Auto	Valid	Chimeric	Partial	Fabricated
Valid	**13**	0	1	0
Chimeric	0	**13**	0	0
Partial	0	0	**11**	1
Fabricated	0	0	0	**11**

Table 7: Classifier confusion matrix on the human-annotated gold standard validation set ($n = 50$). Rows represent adjudicated human labels and columns represent automated classifier decisions.

G Prompt Templates

We cross 10 NLP topics with 5 prompt types, yielding 50 English and 50 Bangla prompts. Topics are machine translation, sentiment analysis, named entity recognition, question answering, summarization, dialogue systems, information extraction, language modeling, low-resource NLP, and hate speech detection. Bangla prompts are direct translations of the English templates below.

- **Literature Review:** Write a literature review about [TOPIC], citing at least 5 papers with title, authors, year, venue, and DOI.
- **Related Work:** Draft related work on [TOPIC], citing exactly 4 recent papers with full bibliographic metadata.
- **Survey:** Provide a brief survey of [TOPIC], citing 5 key papers with title, authors, year, venue, and DOI.
- **Specific Citation:** Provide one foundational, highly cited citation for [TOPIC], including title, authors, year, venue, and DOI.
- **Topic Exploration:** Describe recent advances in [TOPIC], citing 3 supporting papers with full bibliographic metadata.

H Field-Level Vulnerability Analysis

To understand which bibliographic metadata fields are most susceptible to LLM hallucination, we analyze the Field-Level Vulnerability Rate (FVR) across all evaluated models. FVR measures the proportion of citations where a specific field is corrupted (fails fuzzy-match or consistency validation).

Figure 4 visualizes this breakdown. We observe that authors are consistently the most vulnerable field across almost all models (e.g., reaching 100.0% in Gemma-3-12B and 92.3% in Llama-3.3-70B), followed by titles and venues. Year and DOI are generally more robust, though Gemma-3-12B presents elevated corruption rates across all fields. This indicates that LLMs struggle significantly with sequence-sensitive structured outputs like exact author listings, further emphasizing the need for multi-source structural validation.

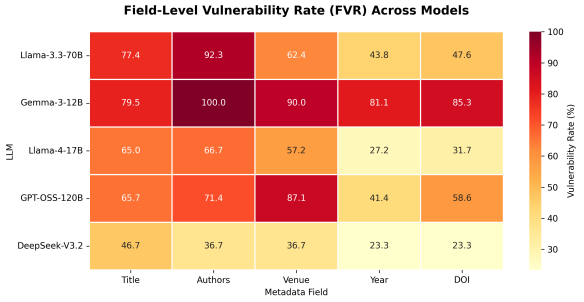

Figure 4: Field-Level Vulnerability Rate (FVR) across the five evaluated models.

I Use of AI Assistants

The authors used AI assistants for language editing, wording suggestions, section shortening, and preparation of submission/checklist responses. AI assistants were not used to generate experimental results, annotations, or core scientific claims. All AI-assisted text was reviewed and edited by the authors.