Publication Date: 02 September 2026
Archs Sci. (2026) Volume 75, Issue 5 Pages 45-54, Paper ID 2025505.
https://doi.org/10.68304/as/75505

Diagnosing Ceiling Effects in Synthetic Citation Verification

Georgi E. Karadzhov
Institute of Mathematics and Informatics, Bulgarian Academy of Sciences, 1113 Sofia, Bulgaria
Corresponding author: Georgi E. Karadzhov (e-mail: georgikaradzhov@gmail.com).

Abstract Perfect performance on synthetic citations can arise from benchmark construction rather than a uniquely effective verification method. This methodological study diagnoses that possibility using a reproducible audit of 776 Crossref records organized into 388 disjoint source–donor pairs and 7760 controlled instances. Seven baselines share candidate retrieval, and a sufficient separation condition explains how mild acceptable edits and whole-field inconsistencies create a ceiling. Four rules classify all 1620 primary test instances correctly, but each is perfect in only 16 of 30 alternative pair partitions; their minimum F1 is 0.9963. Moreover, 953 of 1,000 random positive weight vectors achieve perfect primary test performance. A frozen-threshold stress test retrieves the correct source for all 162 queries at every title-deletion budget, yet all four rules reject every query at the eight-character budget. Their identical decisions follow from the one-dimensional feature variation in this experiment. Removing the near-title exclusion does not remove the primary ceiling and exposes ambiguity in a restored close-title pair. A separate reanalysis of 104 independently labeled references illustrates error tradeoffs in published tool decisions without constituting external validation of our implementations. The findings support reporting calibration sensitivity, retrieval coverage, and perturbation-specific behavior alongside aggregate accuracy. The accompanying computational supplement provides an auditable protocol for doing so.

Index Terms citation integrity, bibliographic metadata, record linkage, data leakage, benchmark, reproducibility

I. Introduction

Academic references connect a claim to the evidence a reader can inspect. When a reference is fabricated or its fields identify different publications, that connection can fail even though the citation appears complete. Research on language-model hallucination provides the broader context [1], and direct evaluations of generated bibliographies have documented both invented works and errors in references to real publications [2]. Similar mistakes can arise through manual editing or reference-manager merges. Their detection should therefore depend on bibliographic evidence rather than an inference about how a text was written.

A digital object identifier (DOI) is useful evidence, but recognizing an identifier and verifying a citation are different operations. If a citation carries the DOI of work A and the title of work B, a successful lookup of A does not establish that the citation is acceptable. Conversely, failure to find a citation in one database does not establish that its source is fictitious. Crossref provides extensive publisher-deposited metadata [3]; agreement with a Crossref record remains a narrower claim than complete bibliographic or substantive correctness.

This problem is naturally related to record linkage: normalization, candidate selection, field comparisons, and decision rules must work together [4]. Existing citation-verification research already considers these components. CITECHECK combines retrieval with metadata checking [5], and GHOSTCITE investigates citation validity across generated and published material [6]. Thus, the distinction between identifier existence and metadata identity is established prior work, not a novelty claim of the present study.

The focus here is the evaluation protocol. Controlled benchmarks are useful because the intended transformation is known. They can nevertheless produce misleading conclusions if a "corruption" changes nothing, identical inputs receive different labels, or donor records appear in a different evaluation partition from the citations constructed from them. In addition, a benchmark in which every acceptable reference is guaranteed to be in the candidate index can conceal retrieval failures that arise in practice.

We examine how benchmark construction produces a ceiling, whether that ceiling survives recalibration and changes in field weights, and how source recovery changes when titles are degraded or the correct record is unavailable. The contribution is a reproducible methodological audit comprising a sufficient separation condition, seven transparent baselines, dependency-preserving evaluation, and controlled sensitivity experiments. A secondary analysis of published

tool decisions supplies external context while retaining the original authors' independent reference labels. The purpose is to diagnose what a benchmark measures, not to introduce a superior citation-verification algorithm.

II. Related Work and Scope

A. Reference Verification and Practical Error Sources

Walters and Wilder distinguish fabricated references from bibliographic inaccuracies in references to existing works [2]. Those categories motivate separate questions about existence and field agreement. More recent systems, including CITECHECK and GHOSTCITE, extend citation verification through retrieval and structured checking [5], [6]. They address broader settings than the fixed-index experiment reported here.

Badalova and Mayr compare available tools and document practical difficulties involving extraction, incomplete metadata, and inconsistent verification outcomes [7]. Their public dataset contains independently produced reference-level labels and standardized tool decisions [8]. We reanalyze those recorded decisions separately. The present structured-input rules are not evaluated on its free-text references; a common extraction and retrieval protocol would be needed for a direct comparison.

B. Dependencies and Evaluation Units

Multiple variants derived from one publication are related observations. A corruption that borrows metadata from a second publication introduces another dependency. Keeping only the original source together therefore does not necessarily keep the construction process within one partition. General guidance on structured evaluation emphasizes matching the split to the dependence structure and the intended generalization claim [9]. Here, disjoint work pairs provide an explicit unit for both partitioning and resampling. This controls the known construction dependencies; it does not imply that all topical, author, or venue similarities have been removed.

C. Operational Definition

The input is an already structured citation

$$c = (d, t, A, y, v, e), \tag{1}$$

where d is an optional DOI, t a title, A a list of family-name strings, y a year, v a venue, and e an explicit author-list truncation indicator. Parsing PDFs, inferring surname boundaries, and extracting citations from prose are outside the evaluated task.

An instance is labeled *acceptable* ($z = 1$) if it is constructed from a source record using one of the five permitted transformations in Table 2. It is labeled *inconsistent* ($z = 0$) if a specified identifying field is deliberately replaced or shifted. The acceptable class includes one defined transcription error; it is therefore an identity-oriented acceptance policy, not a requirement for character-for-character metadata accuracy. Labels are generated by construction, not by independent human annotation. They should not be reinterpreted as judgments about every possible citation style or publication version.

III. Data and Benchmark Construction

A. Frozen Metadata and Eligibility

The starting collection contains 800 records selected from four Crossref keyword queries: "artificial intelligence machine learning," "climate change sustainability," "public health epidemiology," and "materials science nanotechnology." The initial query requested 260 relevance-ranked journal-article records per query with a publication-date filter from 2019 to 2025. The first 200 records per query meeting the original field-completeness and title-length rules were retained. The supplied response arrays and original 800-record snapshot are preserved with checksums in the supplementary project. The calculations use these frozen files; repeating the live queries may return a different ranking.

The revised eligibility rules require a DOI, title, deposited family names, venue, and recorded year in 2019–2025. Years follow the original priority order: print publication, online publication, issued date, and creation date. Because these date fields can differ, the year is a benchmark reference value rather than a universally correct publication year. The explicit year check removes a record that passed the original query filter but had a selected print year of 2026.

Titles containing the terms "editorial," "call for papers," "editor announcement," "foreword," "erratum," "corrigendum," or "retraction" are excluded. This mechanical screen does not constitute manual confirmation that every retained item is a substantive research article. A title must also contain an alphabetic word of at least six characters to support the declared typo transformation.

Exact duplicate DOIs and normalized titles are removed in original identifier order. To reduce ambiguous near duplicates, a record is also excluded if its normalized title has character similarity of at least 0.95 to an earlier retained title. This is a conservative exclusion rule, not a claim that every similar title identifies the same work. It can remove distinct publications and potentially simplify title matching. An ablation removes this cutoff while retaining exact-title deduplication, allowing its effect to be measured rather than assumed. Original deposited family-name strings are restored from the saved responses, preserving accents and compound surnames instead of inventing initials.

After screening, deduplication, and pairing, 776 records remain; 24 of the original 800 are excluded. Table 1 reports the actual composition. These are keyword-query strata, not representative samples of four disciplines. A journal title that matches a query can dominate its results, so equal initial query quotas do not imply balanced topical or venue coverage.

B. Pairing and Partitioning

Within each stratum, all candidate pairs with different normalized author sets are ranked by decreasing title similarity.

Table 1. Composition after screening and pairing. The final column is the share of each stratum contributed by its largest venue.

Query stratum	Records	Venues	Largest venue	Share
Artificial Intelligence	196	58	International Journal of Artificial Intelligence and Machine Learning	7.1%
Climate Sustainability	196	6	Sustainability and Climate Change	81.6%
Materials Science	194	80	Nanotechnology	23.7%
Public Health	190	63	Journal of Interventional Epidemiology and Public Health	32.6%

Table 2. Ten instances per source. All fields not specified remain unchanged. "Paired" refers to the single donor in the source's disjoint pair.

Acceptable variants	Construction
Exact metadata	Original title, DOI, family names, year, and venue.
Formatting	Case changes preserving normalization, with title punctuation separators.
Short representation	First family name, explicit truncation flag, and venue initials.
DOI omitted	Empty DOI and normalized title.
Title typo	Empty DOI and one-character deletion in the title.
Inconsistent variants	Construction
DOI swap	Replace DOI with the paired DOI.
Author swap	Replace the complete family-name list with the paired list.
Year shift	Add three to the recorded year.
Title substitution	Replace title with the paired title.
Composite splice	Use paired DOI, authors, year, and venue; retain source title.

A deterministic greedy procedure selects an edge only when neither record has already been paired; ties are broken by source identifiers. Unpaired records are excluded. Both directions of a pair are used: each work acts as source once and as donor once. The resulting title similarities have median 0.518, interquartile range 0.447–0.590, and maximum 0.868. Thus, the paired substitutions are not uniformly close-title adversarial examples.

The resulting 388 pairs are shuffled within each stratum with seed 20260921. The first $\lfloor 0.6m \rfloor$ pairs in a stratum form training data, the next $\lfloor 0.2m \rfloor$ form validation data, and the remainder form the test data, where m is that stratum's number of pairs. All variants and both source–donor directions stay together. Consequently, no record supplies labels or corruption fields across these partitions. Pairing is computed before assignment and uses only metadata; no labels, predictions, or selected thresholds influence it. The split is a within-snapshot, within-stratum evaluation rather than a temporal or unseen-domain test.

The reference index contains all 776 retained records, including held-out works. This is intentional: the primary task is verification against a known reference collection. Training labels and generated variants are partitioned, but the reference index is not restricted to training works. The separate omission experiment tests the effect of violating reference coverage.

C. Transformations and Integrity Assertions

Every source generates five acceptable and five inconsistent instances (Table 2), giving 7760 instances in total. Formatting variants capitalize a character only if its normalized form is unchanged. This prevents noninvertible Unicode case changes, such as uppercasing a Turkish dotless i, from altering a surname. An executable assertion checks field agreement for every formatting variant. Full author lists are retained in exact representations. The shortened-author variant explicitly supplies the first family name with a truncation flag, corresponding to an "et al." representation. Its venue is the sequence of initial letters of the deposited venue name; this narrow abbreviation rule is declared rather than inferred from an unrestricted synonym dictionary.

The title-typo variant deletes the middle character of the longest alphabetic title word with at least six characters, breaking ties by the earliest word. Its DOI is omitted so that retrieval must tolerate this edit. This transformation represents a limited transcription error and does not establish robustness to arbitrary title paraphrases, translated titles, or meaning-changing edits.

For each negative instance, the executable audit checks that the full structured citation differs from its unchanged source representation. It also checks that identical structured inputs never have opposite labels, and that every source and donor belongs to the same pair and partition. The final data pass all three checks. Repeated acceptable representations are retained when different permitted transformations happen to produce the same fields; this known repetition is contained within the resampling unit.

IV. Verification Rules

A. Candidate Retrieval

A populated DOI is looked up exactly in the frozen index. No title fallback is used to override an explicit DOI. When the DOI is absent, the candidate with the highest normalized title similarity is selected from the index; ties favor the smaller source identifier. Candidate selection does not use the instance label, source identifier, or evaluation partition as input evidence.

Lookup here is *dictionary membership*, not a live DOI-resolution experiment. An additional set of 776 identifiers is formed by appending a fixed suffix to source DOIs. Their

absence from the index is reported separately as an unresolved status. They are not included as known fabricated references in the binary benchmark, because no external nonexistence determination was made.

B. Normalization and Field Scores

Normalization decodes HTML entities, removes markup, applies Unicode compatibility decomposition (NFKD), case folding, and a second NFKD pass, removes combining marks, replaces nonalphanumeric characters with spaces, and collapses whitespace. Both sides of a field comparison use the same operation. Accent folding is a pragmatic matching choice and can conflate different names; the study does not validate multilingual identity resolution.

Title similarity s_t and ordinary venue similarity use Python's `SequenceMatcher` with `autojunk=False`. Its score is

$$s(u, v) = \frac{2M(u, v)}{|u| + |v|}, \tag{2}$$

where M is the total length of the matching blocks selected by that algorithm after normalization. This score is not a Levenshtein distance or a semantic similarity measure. A venue that matches the explicitly defined initials of the candidate venue receives $s_v = 1$.

Author agreement compares *sets of normalized family-name strings*:

$$s_a = \frac{|N(A) \cap N(A_r)|}{|N(A) \cup N(A_r)|}. \tag{3}$$

For an explicitly truncated list of length k, A_r is first restricted to its first k family names. Thus, a compound surname is consistently represented as one element on both sides. The set comparison ignores order and cannot distinguish people with identical family names. No credit is claimed for disambiguating such authors. Year agreement is $s_y = \mathbb{1}(y = y_r)$; alternative online and print years are not modeled.

C. Baselines and Calibration

All metadata-based rules use the same candidate and four features $x = (s_t, s_a, s_y, s_v)$. The seven baselines are:

1) *DOI membership*: accept exactly when the supplied DOI occurs in the index.
2) *Normalized exact*: accept when all four field scores equal one, including the declared truncation and venue-initial rules.
3) *Title only*: threshold s_t.
4) *Equal weights*: threshold the mean of all four features.
5) *Weighted agreement* (METACHECK): threshold
$$S = 0.46s_t + 0.24s_a + 0.18s_y + 0.12s_v. \tag{4}$$
6) *Minimum field*: threshold $\min(s_t, s_a, s_y, s_v)$.
7) *Logistic regression*: fit a linear logistic model to training features, using L_2 regularization, $C = 1$, and the `lbfgs` solver.

The weighted coefficients are declared heuristic values retained for comparison; they were not estimated or justified as optimal. The minimum-field rule prevents a high score on one field from compensating for a low score on another. Logistic regression provides a learned linear comparison using scikit-learn [10]. Neither its regularization constant nor its feature set is tuned on test data.

For the five scored rules, the threshold maximizes acceptable-class F1 on the validation partition. Candidate thresholds are all observed validation scores plus one value above their maximum. Ties favor fewer false acceptances and then the higher threshold. DOI membership and normalized exact matching are fixed binary rules. Thresholds are frozen for test evaluation. The study is a retrospective benchmark-development analysis, not a prospectively preregistered experiment; code and selection rules are supplied to make its decisions inspectable.

V. A Sufficient Condition for a Ceiling

The construction permits a simple explanation of perfect discrimination. Consider a nonnegative weighted score $S_w(x) = \sum_j w_j x_j$, with $\sum_j w_j = 1$. Suppose that acceptable examples have $s_a = s_y = s_v = 1$ and $s_t \geq 1 - \epsilon$. Suppose further that title-based inconsistencies have title deficit at least $\delta_t > 0$, author substitutions have author deficit at least $\delta_a > 0$, and year substitutions have $s_y = 0$. These assumptions describe the declared transformations after successful candidate retrieval; they do not characterize arbitrary natural references.

Every acceptable example then satisfies $S_w \geq 1 - w_t\epsilon$. The inconsistent examples satisfy

$$S_w \leq 1 - \beta_w, \qquad \beta_w = \min\{w_t\delta_t, w_a\delta_a, w_y\}. \tag{5}$$

For DOI swaps, title substitutions, and composite splices, the bound follows from their candidate-relative title deficit; any additional disagreements only reduce the score. The other two negative families give the author and year bounds directly. Consequently, if

$$w_t\epsilon < \beta_w, \tag{6}$$

any threshold in $(1 - \beta_w,\ 1 - w_t\epsilon]$ separates the two classes. This is a finite-sample sufficient condition, not a population error bound. A validation-selected threshold must still fall inside the test interval; a positive test gap alone does not guarantee successful threshold transfer.

Eq. (6) identifies two mechanisms worth testing: mild acceptable perturbations keep ϵ small, while whole-field substitutions keep negative deficits large. It also shows why many weight vectors may work equally well. A minimum-field rule has the analogous sufficient condition $\epsilon < \min\{\delta_t, \delta_a, 1\}$. These observations motivate the perturbation and weight analyses below; they are elementary properties of the score, not claims of a new general record-linkage theorem.

A. Equivalent Decisions for Title-Only Edits

When a retrieved candidate matches the authors, year, and venue, its feature vector is $(s_t, 1, 1, 1)$. Along this line,

$$S_w = 1 - w_t + w_t s_t, \qquad S_{\min} = s_t. \tag{7}$$

For $w_t > 0$, weighted and minimum-field scores are strictly increasing functions of the same scalar. Logistic regression is also strictly increasing along this line when its fitted title coefficient is positive. If all these rules perfectly separate the validation examples, our higher-threshold tie rule selects the score of the least similar acceptable validation title. Their decisions on further title-only edits then coincide: all accept exactly when s_t reaches that common title cutoff. Equal performance in this setting consequently need not reflect independent robustness across four different decision mechanisms. The deletion experiment tests this implication while checking candidate retrieval separately.

VI. Evaluation Protocol

The positive class is *acceptable*. Precision measures the proportion of accepted instances with label one; recall measures the proportion of acceptable instances accepted. We report their harmonic mean F1, accuracy, false-acceptance rate among inconsistent instances, and area under the receiver-operating-characteristic curve (AUROC). Average precision is included in the machine-readable supplement. Ranking measures and thresholded decisions answer different questions [11]; class prevalence also affects the interpretation of precision [12].

The held-out partition contains 1620 instances from 162 sources and 81 pairs. Confidence intervals use 2,000 stratified cluster-bootstrap replicates, sampling complete test pairs with replacement within each query stratum. All 20 instances belonging to a pair enter together. Percentile intervals summarize variation across these pair resamples [13]. The same draws are used for paired differences between methods. These are conditional intervals for the frozen sampling and construction procedure. They do not account for uncertain source metadata, alternative harvesting strategies, or repeated model and threshold selection.

Two diagnostic experiments supplement the primary test. In *corruption-family transfer*, title substitution and composite splicing are excluded from training and validation. Thresholds are selected again, and logistic regression is refitted; testing uses all acceptable variants and the two withheld negative families from held-out pairs. DOI-swap examples remain in calibration and can also create candidate-relative title mismatches, so this is a family-level holdout rather than removal of all evidence about title conflicts. The class mixture changes, so false acceptance of the withheld negatives is reported explicitly rather than interpreting the resulting F1 as directly comparable with the balanced primary test.

In *reference omission*, the correct source record is removed from the index separately for each held-out DOI-free instance. The weighted rule retains its primary threshold, and retrieval searches the remaining records. An acceptance then constitutes an incorrect candidate match; a rejection indicates that the available candidate is insufficient. This test assesses behavior under an intentionally incomplete index, not the prevalence of missing records in Crossref.

A. Controlled Sensitivity Analyses

Four additional analyses investigate the origin and stability of the ceiling. First, 30 stratified pair partitions use seeds 20261001–20261030, keeping the original proportions. Each run refits logistic regression and reselects all scored thresholds on its own validation partition. The metadata, pair construction, and features remain fixed. The resulting range describes partition sensitivity, not 30 independent datasets or a confidence interval.

Second, 1,000 weight vectors are sampled from $\mathrm{Dirichlet}(1, 1, 1, 1)$ using seed 20261031. This distribution is uniform on the four-component simplex. Each vector receives a validation-selected threshold on the primary split; test performance is recorded without selecting or recommending a best vector. This analysis asks whether successful weighting occupies a broad region rather than validating the original heuristic coefficients.

Third, DOI-free queries are generated for the 162 held-out source works using deletion budgets of 1, 2, 4, 8, and 16 alphabetic characters from the normalized title. Each source uses a deterministically shuffled list of positions, seeded by 20261101 plus its identifier; successive budgets remove nested sets of positions, retaining at least one alphabetic character. Other fields remain unchanged. The index, trained model, and primary thresholds are frozen. We report correct retrieval, acceptance of the correct source, and acceptance of an incorrect candidate separately. These are source-origin recovery tests: deletion can change meaning, so the experiment does not declare every degraded string an acceptable natural citation. Each budget contains one query per held-out source; its different budgets are dependent observations.

Fourth, the near-title exclusion cutoff of 0.95 is removed while all other eligibility rules remain fixed. Pairing, splitting with the primary seed, fitting, and calibration are repeated, followed by the same 30 alternative partition seeds. Because eligibility changes the paired cohort, this is a pipeline sensitivity analysis rather than a paired comparison of identical test instances. Exact duplicate titles remain excluded.

All additions were specified during retrospective revision after inspecting the original benchmark. Their grids and seeds are fully reported. The stress results do not feed back into the primary thresholds or coefficients.

B. External Context from Published Tool Decisions

We separately recompute confusion counts from version 1.0.0 of the Badalova–Mayr dataset: 104 references from three documents, including 71 verified and 33 problematic references [8]. Their manual labels and five tools' published decisions are retained unchanged. To match our positive-class

convention, "verified" is positive and "not_flagged" predicts positive. No tools are rerun, no new annotation is claimed, and these observations are not pooled with the synthetic benchmark. This secondary analysis supplies context, not external validation of our implementations.

VII. Results

A. Primary Benchmark

Table 3 presents the held-out results. Equal weighting, weighted agreement, the minimum-field rule, and logistic regression each accept all 810 acceptable instances and reject all 810 inconsistent instances. There is no observed performance advantage for the weighted rule, whose validation-selected threshold is 0.989. Normalized exact matching obtains F1 0.889 because it rejects all 162 deliberately tolerated typos. DOI membership accepts every inconsistent instance and rejects the two DOI-free acceptable variants, giving F1 0.462. Title-only matching cannot detect author or year substitutions and obtains F1 0.833. Figure 1 shows the corresponding score distribution and method comparisons.

All pair-bootstrap intervals are degenerate, not just those of the perfect classifiers. Every test pair contributes the same confusion counts for a given method, so resampling pairs cannot change its F1. The intervals therefore describe a structural property of this benchmark and cannot quantify the probability of errors on unseen natural citations. The paired F1 differences between the four strongest methods are identically zero across these resamples.

B. Transformation-Specific Behavior

Table 4 reports the fraction classified correctly within each transformation. For acceptable variants, this is acceptance; for inconsistent variants, it is rejection. Exact matching is expected to penalize the deliberately tolerated title typo, while a title-only rule cannot detect an author or year substitution when the title remains unchanged. These are consequences of the information supplied to each rule, not evidence of learned reasoning about publications.

High performance after applying the declared normalization rules must also be interpreted in light of the construction process. Most acceptable fields are copied directly from the reference record, and negative fields are deliberately altered. The relative severity of tolerated title edits and substituted metadata should therefore be examined directly; the ablation below separately tests the near-duplicate exclusion. Accordingly, the benchmark is suitable for detecting implementation mistakes and testing specified failure modes, but its aggregate scores alone are insufficient for ranking comprehensive citation-verification systems.

C. Transfer to Withheld Corruption Families

Table 5 reports the separate calibration exercise in which neither title-substitution family appears in training or validation. Equal weighting, weighted agreement, the minimum-field rule, and refitted logistic regression again classify all 1,134 test instances correctly. Title-only matching also succeeds on this narrower test because both withheld negative families alter the candidate-relative title. This is evidence of transfer only to these particular substitutions. It does not establish generalization to arbitrary unseen corruption mechanisms, and it does not resolve the ceiling effect.

D. Separation Margins

The ceiling can be inspected directly. For a scoring rule, define the descriptive test-set margin

$$\Delta = \min_{i:z_i=1} S_i - \max_{i:z_i=0} S_i. \tag{8}$$

A positive margin means that some threshold perfectly separates the two classes in that finite set; it is not used to select a test threshold. Table 6 reports these extrema. For the strongest rules, the restricted acceptable edits remain close to the source while the paired substitutions create a larger discrepancy. The observed separation explains why substantially different weighting choices produce the same decisions.

E. Retrieval and Reference Coverage

With the complete index, title retrieval selected the correct source for 162 of 162 DOI-omitted test instances and 162 of 162 title-typo test instances. After the correct source was removed separately for each of these 324 instances, the weighted rule accepted 0 substitute candidates. All 776 suffixed identifiers were absent from the fixed index. Their status is unresolved; no claim is made that they cannot resolve elsewhere.

The omitted-record experiment distinguishes a metadata mismatch from a finding that a work does not exist. The reference remains a known source of the benchmark even when it has deliberately been removed from the index. A practical system should retain an unresolved or insufficient-evidence outcome and offer a second retrieval source or human review, instead of treating the failure of one lookup as proof of fabrication.

F. Sensitivity to Splits, Weights, and Perturbations

The observed primary test features satisfy the sufficient condition with $\epsilon = 0.0154$, $\delta_t = 0.2115$, and $\delta_a = 0.6667$. For the weighted rule, the acceptable-score deficit is bounded by $0.46\epsilon = 0.0071$, whereas the inconsistent-score deficit is at least $\beta_w = 0.0973$. The resulting separation is therefore explained by the construction, without requiring optimal coefficients.

Table 7 shows that the ceiling is sensitive to which pairs calibrate the threshold. Each of the four strongest rules is perfect in 16 of 30 alternative partitions; F1 ranges from 0.9963 to 1.0000. The other runs produce only false rejections, with at most six per run. These results retain very high accuracy but show why a single perfect split and a degenerate

ARCHIVES DES SCIENCES
A Multidisciplinary Journal

Table 3. Primary test results. Positive class: acceptable citation. FAR is the fraction of inconsistent instances accepted. Intervals resample source–donor pairs.

Method	Precision	Recall	F1	95% F1 interval	FAR	AUROC
DOI membership	0.375	0.600	0.462	[0.462, 0.462]	1.000	0.300
Normalized exact	1.000	0.800	0.889	[0.889, 0.889]	0.000	0.900
Title only	0.714	1.000	0.833	[0.833, 0.833]	0.400	0.760
Equal weights	1.000	1.000	1.000	[1.000, 1.000]	0.000	1.000
Weighted	1.000	1.000	1.000	[1.000, 1.000]	0.000	1.000
Minimum field	1.000	1.000	1.000	[1.000, 1.000]	0.000	1.000
Logistic	1.000	1.000	1.000	[1.000, 1.000]	0.000	1.000

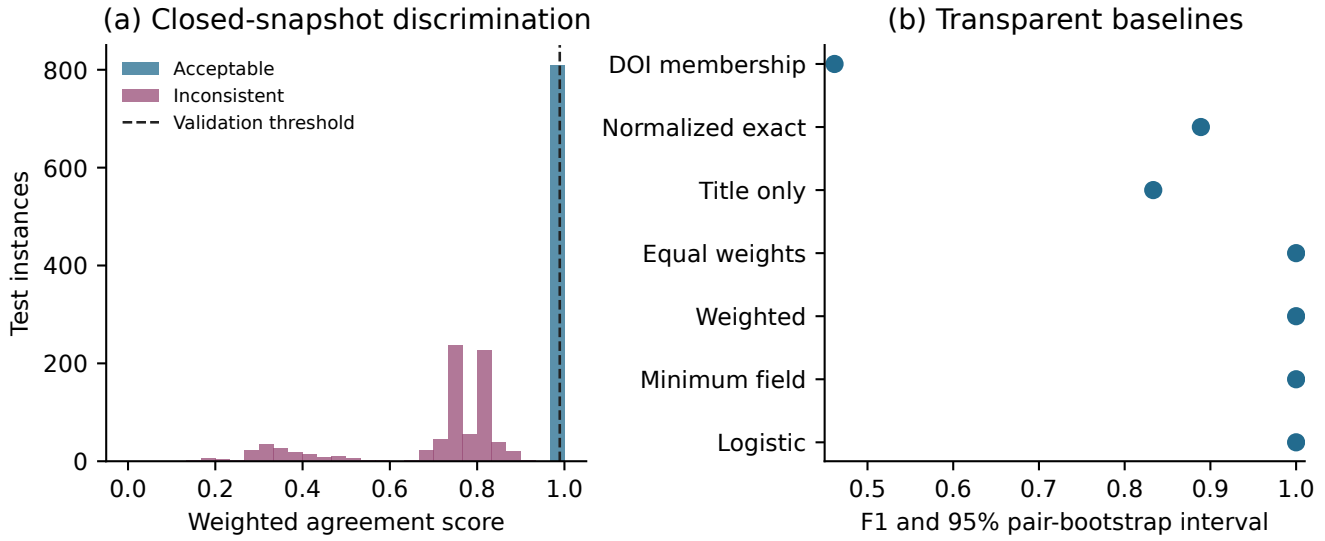

Figure 1. Primary test performance. (a) Weighted agreement scores for acceptable and inconsistent instances; the dashed line is selected using validation data. (b) F1 for seven transparent rules with 95% stratified pair-bootstrap intervals. Narrow or degenerate intervals describe this controlled test and do not establish certainty about real-world performance.

Table 4. Fraction correctly classified by transformation on held-out sources. Exact denotes normalized exact matching; Min. denotes the minimum-field rule.

Transformation	Exact	Weighted	Min.	Logistic
Exact metadata	1.000	1.000	1.000	1.000
Case/punctuation variant	1.000	1.000	1.000	1.000
Short author list/venue	1.000	1.000	1.000	1.000
DOI omitted	1.000	1.000	1.000	1.000
One-character title typo	0.000	1.000	1.000	1.000
DOI swap	1.000	1.000	1.000	1.000
Author swap	1.000	1.000	1.000	1.000
Year shift	1.000	1.000	1.000	1.000
Paired-title substitution	1.000	1.000	1.000	1.000
Composite splice	1.000	1.000	1.000	1.000

conditional bootstrap interval are incomplete descriptions of stability.

Of 1,000 random positive weight vectors, 953 (95.3%) are perfect on the primary test split. A positive test-score gap occurs for 959 vectors, illustrating that a separating interval and successful validation-to-test threshold transfer are different properties. The observed F1 range is 0.9091–1.0000. These results demonstrate that this benchmark supplies little evidence for preferring the declared weights over many alternatives.

Table 8 and Fig. 2 separate retrieval from acceptance. Correct-source retrieval remains 162/162 at every deletion budget. Nevertheless, each of the four rules accepts only 162, 153, 60, 0, and 0 correct sources at budgets 1, 2, 4, 8, and 16, respectively. No incorrect candidate is accepted. Their identical acceptance curves accord with the title-only equivalence argument: all four primary thresholds reduce to the same title cutoff, 0.9767, on this restricted feature line. The fitted logistic title coefficient is positive. This is calibration brittleness under a defined stressor, not evidence that the rejected degraded strings would all be acceptable to an editor.

Removing the near-title cutoff retains 776 records after pairing and again gives F1 1.000 for all four rules on the primary partition. The closest pair has title similarity 0.9905 and falls in validation. Under the same 30 alternative seeds, each rule is perfect in 15/30 runs, with F1 0.9951–1.0000 and at most 4 false acceptances per run relative to the construction labels. Thus, the cutoff alone does not explain the original test ceiling.

The restored close pair also exposes label ambiguity. Its two

ARCHIVES DES SCIENCES
A Multidisciplinary Journal

Table 5. Corruption-family transfer. Calibration excludes both paired-title substitution and composite splicing. The test mixture contains five acceptable and two withheld inconsistent instances per source. All methods use the same held-out pairs as the primary test.

Method	Threshold	F1	Acceptable recall	Withheld-negative acceptance
DOI membership	0.5000	0.600	0.600	1.000
Normalized exact	0.5000	0.889	0.800	0.000
Title only	0.9767	1.000	1.000	0.000
Equal weights	0.9942	1.000	1.000	0.000
Weighted	0.9893	1.000	1.000	0.000
Minimum field	0.9767	1.000	1.000	0.000
Logistic	0.9915	1.000	1.000	0.000

Table 6. Descriptive test-score extrema. A positive margin indicates finite-sample separation; thresholds remain validation-selected.

Method	Min. acceptable	Max. inconsistent	Margin
DOI membership	0.0000	1.0000	-1.0000
Normalized exact	0.0000	0.0000	0.0000
Title only	0.9846	1.0000	-0.0154
Equal weights	0.9962	0.9471	0.0490
Weighted	0.9929	0.9027	0.0902
Minimum field	0.9846	0.7885	0.1962
Logistic	0.9638	0.5410	0.4228

Table 7. Sensitivity to 30 pair partitions. A perfect run has no false acceptance or rejection. The range is descriptive, not a confidence interval.

Method	Perfect	Min. F1	Median	Max. F1
Equal weights	16/30	0.9963	1.0000	1.0000
Weighted	16/30	0.9963	1.0000	1.0000
Minimum field	16/30	0.9963	1.0000	1.0000
Logistic	16/30	0.9963	1.0000	1.0000

DOI records share the recorded year, venue, and pagination and have nearly identical titles and author lists. They may represent inconsistent deposits of the same work; no independent identity adjudication was performed. Consequently, their constructed mismatch labels should not be treated as verified natural errors. The ablation measures pipeline sensitivity and demonstrates why closer-title examples require identity review before they become a trustworthy challenge set.

G. Published Decisions on Natural References

Table 9 reports the recomputed counts. Verified-class F1 ranges from 0.367 to 0.745; no recorded tool reproduces every manual label. Both missed problematic references and flags on verified references occur. These are historical outcomes on three selected documents, not a current tool ranking. Differences from our synthetic results cannot be attributed solely to dataset difficulty: the tools, inputs, retrieval procedures, and label policies also differ. Document-level counts are supplied for inspection.

VIII. Discussion

A. What the Benchmark Establishes

The experiment shows how a citation verifier can be evaluated without conflating source identity, metadata agreement, and database membership. It provides an auditable construction in which source–donor dependencies remain within the evaluation unit. It also makes normalization part of the method: author-list truncation is explicit, compound surnames receive symmetric treatment, and venue initials follow a specified rule.

The strongest result is methodological. Performance should be interpreted relative to the exact acceptance policy and the transformations available during calibration. A benchmark with copied acceptable metadata and deterministic corruptions can be highly separable. That property is useful for regression testing but provides limited evidence about ambiguous references encountered by editors. The simple baselines are therefore substantive comparators, and favorable results for a weighted score do not establish that its particular coefficients are necessary. The weight and partition experiments distinguish score separation from calibration stability, while the deletion experiment distinguishes retrieving a source from accepting it. This separation of failure stages is more informative than a single aggregate score.

B. Triage and External Validation

An operational workflow should distinguish at least three outcomes: metadata compatible with a retrieved candidate, metadata in conflict with that candidate, and insufficient evidence to decide. Selective-classification research formalizes the broader option to trade coverage for reliability [14]. The present experiment does not estimate a deployment-ready abstention policy or calibrate scores as probabilities.

External evaluation remains necessary. A suitable study would fix the extraction and retrieval procedure, use independently adjudicated natural citation errors, distinguish incorrect metadata from legitimate publication versions, and retain unresolved cases. The public reference-verification dataset of Badalova and Mayr supplies one starting point [8]. Such evaluation should include multiple databases and report candidate-retrieval coverage separately from verification conditional on successful retrieval. It should also compare relevant deployed tools under the same input and access conditions. Numerical results from unrelated papers should not be treated as a common leaderboard.

C. Limitations

Several limits materially constrain interpretation. The starting records come from a small relevance-ranked convenience sample, with substantial venue concentration. Mechanical

Table 8. Title-deletion stress test with frozen primary thresholds. Retrieval reports selection of the correct source; method columns report acceptance of that correct source. Each denominator is 162 queries.

Deletion budget	Correct retrieval	Equal weights	Weighted	Minimum field	Logistic
1	162/162	162/162	162/162	162/162	162/162
2	162/162	153/162	153/162	153/162	153/162
4	162/162	60/162	60/162	60/162	60/162
8	162/162	0/162	0/162	0/162	0/162
16	162/162	0/162	0/162	0/162	0/162

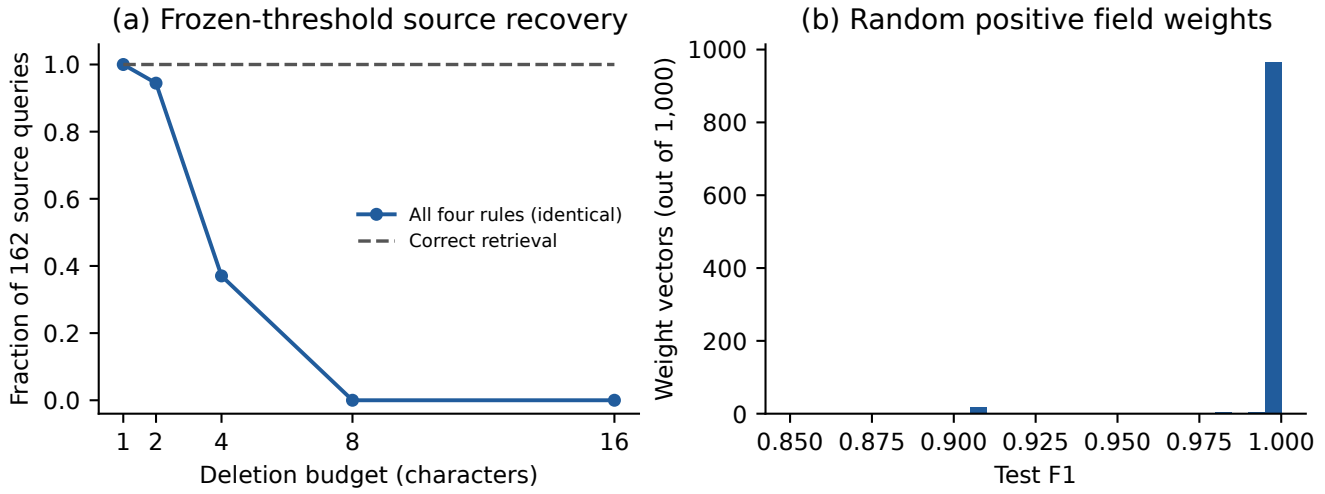

Figure 2. Diagnostics beyond the original ceiling. (a) Recovery and acceptance of source works under nested title deletions, with all primary thresholds frozen. (b) Test F1 for 1,000 positive weight vectors, each calibrated on validation data. Neither panel selects a preferred rule.

Table 9. Secondary analysis of recorded external tool decisions [8]. TP: verified and unflagged; FP: problematic and unflagged; FN: verified and flagged; TN: problematic and flagged. Positive class is verified.

Tool	TP	FP	FN	TN	F1
CheckIfExist	37	2	34	31	0.673
HalluCiteChecker	51	15	20	18	0.745
Hallucinator	43	4	28	29	0.729
HalRef	18	9	53	24	0.367
RefChecker	35	1	36	32	0.654

notice filtering does not verify article quality, peer review, or scientific validity. Source metadata can contain errors, and near-duplicate exclusions can remove distinct works. The fixed print-first year rule does not resolve online/issue-date differences.

Only the specified structured fields are evaluated. Namesake authors, author order, identifier aliases, journal synonyms beyond generated initials, multilingual references, and free-text extraction are not comprehensively tested. Acceptable titles differ through a restricted set of edits; the declared typo policy is not a semantic guarantee. Pairing eliminates known cross-partition construction dependencies while leaving shared subject matter and venues across partitions.

The balanced class mixture is artificial. Precision in an editorial workflow depends on the actual frequency and type of errors, which this study does not estimate. Pair-bootstrap intervals condition on the observed sample and fixed predictions, and a zero-width interval at a ceiling score cannot establish zero deployment error. The study does not evaluate live DOI resolution, full-text claim support, AI authorship detection, fabrication prevalence, or citation impact. No independent human adjudication of the constructed labels or external evaluation of the seven implementations is claimed. The secondary analysis inherits another study's labels, convenience sample, and recorded decisions; it cannot correct annotation uncertainty or establish current tool performance. The title perturbations are mechanical stressors rather than a model of naturally occurring error frequencies.

D. Reproducibility and Reuse

The supplementary project contains the original snapshot, saved Crossref response arrays, eligibility exclusions, pair assignments, generated instances, predictions, selected thresholds, software versions, and analysis code. Tables, figures, and numerical summaries are generated from the saved predictions and result files. The downloaded external CSV and its original documentation are retained under their CC BY 4.0 license, with attribution and checksums. The analysis copy converts the CSV from CP850 to UTF-8 without changing field values or labels. Reproduction of the reported analysis requires no network access. These provisions support inspection and reuse

in the spirit of FAIR data stewardship [15]; a permanent public repository identifier should accompany any final archival release.

IX. Conclusion

The CITEINTEGRITY audit explains why a synthetic citation benchmark can give several verification rules indistinguishable, perfect scores. Mild acceptable perturbations and large candidate-relative inconsistencies create a broad separating region, and 95.3% of sampled positive weight vectors attain the primary test ceiling. Nevertheless, alternative pair partitions expose threshold sensitivity, and stronger title degradation causes rejection even when the correct source is still retrieved. These findings distinguish benchmark separability from robust verification.

A useful evaluation should therefore report simple baselines, construction-label audits, complete source–donor dependencies, repeated calibration splits, and retrieval and acceptance as separate outcomes. Closely matching records require identity adjudication before they are labeled as different works. The accompanying code, predictions, and sensitivity analyses make these recommendations executable. Claims about deployment on natural references remain a separate empirical question requiring a common extraction and retrieval protocol and independent reference judgments.

Data and Code Availability

The complete computational supplement accompanies this manuscript as a LaTeX project archive. It includes all files required to regenerate the benchmark, sensitivity analyses, tables, figures, and reported metrics. The independently annotated dataset used for the secondary analysis is available from its original repository at https://doi.org/10.5281/zenodo.21457492.

Funding and Competing Interests

No external funding was reported for this study. No competing interests were declared.

Ethics and Use of Generative AI

The analysis uses public bibliographic metadata and synthetically modified citation records; it involves no participant recruitment, clinical intervention, or private participant data. OpenAI ChatGPT/Codex assisted with the manuscript drafting and revision. The author is responsible for verifying the scientific content and cited sources and for approving any final submission.

References

[1] Ji, Z., Lee, N., Frieske, R., Yu, T., Su, D., Xu, Y., Ishii, E., Bang, Y. J., Madotto, A., & Fung, P. (2023). Survey of hallucination in natural language generation. *ACM Computing Surveys, 55*(12), Article 248, 1–38.

[2] Walters, W. H., & Wilder, E. I. (2023). Fabrication and errors in the bibliographic citations generated by ChatGPT. *Scientific Reports, 13*, Article 14045.

[3] 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.

[4] Christen, P. (2012). *Data matching: Concepts and techniques for record linkage, entity resolution, and duplicate detection*. Springer.

[5] Khajavi, K., Sadeghi, S., Adhikari, R., & Tessier, A. (2026). *CiteCheck: Retrieval-grounded detection of LLM citation hallucinations in scientific text* [Preprint]. arXiv. https://arxiv.org/abs/2605.27700.

[6] Xu, Z., Qiu, Y., Sun, L., Miao, F., Wu, F., Li, X., Wang, X., Lu, H., Zhang, Z., Hu, Y., Li, J., Luo, J., Zhang, F., Luo, R., Liu, X., Li, Y., & Liu, J. (2026). *GhostCite: A large-scale analysis of citation validity in the age of large language models* [Preprint]. arXiv. https://arxiv.org/abs/2602.06718.

[7] Badalova, F., & Mayr, P. (2026). *Detecting hallucinated and suspicious citations: What current tools can and cannot do* [Preprint]. arXiv. https://arxiv.org/abs/2607.22693.

[8] Badalova, F., & Mayr, P. (2026). *Manual reference verification dataset for hallucinated and suspicious citation detection tools* (Version 1.0.0) [Data set]. Zenodo. https://zenodo.org/records/21457492.

[9] Roberts, D. R., Bahn, V., Ciuti, S., Boyce, M. S., Elith, J., Guillera-Arroita, G., Hauenstein, S., Lahoz-Monfort, J. J., Schröder, B., Thuiller, W., Warton, D. I., Wintle, B. A., Hartig, F., & Dormann, C. F. (2017). Cross-validation strategies for data with temporal, spatial, hierarchical, or phylogenetic structure. *Ecography, 40*(8), 913–929.

[10] Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., Vanderplas, J., Passos, A., Cournapeau, D., Brucher, M., Perrot, M., & Duchesnay, É. (2011). Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research, 12*, 2825–2830.

[11] Fawcett, T. (2006). An introduction to ROC analysis. *Pattern Recognition Letters, 27*(8), 861–874.

[12] Saito, T., & Rehmsmeier, M. (2015). The precision-recall plot is more informative than the ROC plot when evaluating binary classifiers on imbalanced datasets. *PLOS ONE, 10*(3), e0118432.

[13] Efron, B., & Tibshirani, R. J. (1994). *An introduction to the bootstrap*. Chapman & Hall/CRC.

[14] El-Yaniv, R., & Wiener, Y. (2010). On the foundations of noise-free selective classification. *Journal of Machine Learning Research, 11*, 1605–1641.

[15] Wilkinson, M. D., Dumontier, M., Aalbersberg, I. J., Appleton, G., Axton, M., Baak, A., Blomberg, N., Boiten, J.-W., Bonino da Silva Santos, L., Bourne, P. E., Bouwman, J., Brookes, A. J., Clark, T., Crosas, M., Dillo, I., Dumon, O., Edmunds, S., Evelo, C. T., Finkers, R., … Mons, B. (2016). The FAIR guiding principles for scientific data management and stewardship. *Scientific Data, 3*, Article 160018.