Research Square

Preprints are preliminary reports that have not undergone peer review.
They should not be considered conclusive, used to inform clinical practice,
or referenced by the media as validated information.

Auditable verification of bibliographic reference integrity under incomplete provider metadata: evidence traces, selective abstention, and structured author diagnostic boundaries

Ercan Erkalkan

ercan.erkalkan@marmara.edu.tr

Marmara University

Research Article

Keywords:

Posted Date: November 13th, 2026

DOI: https://doi.org/10.21203/rs.3.rs-11423167/v1

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

Additional Declarations: No competing interests reported.

Auditable verification of bibliographic reference integrity under incomplete provider metadata: evidence traces, selective abstention, and structured author diagnostic boundaries

Ercan Erkalkan[1*]

[1*]Department of Electronics and Automation, Vocational School of Technical Sciences, Marmara University, Istanbul, Türkiye.

Corresponding author(s). E-mail(s): ercan.erkalkan@marmara.edu.tr;
ORCID: 0000-0001-9259-7112;

Abstract

Bibliographic-integrity audits must distinguish incorrect references from incomplete provider evidence. A frozen selective verifier was evaluated on 3,712 reference strings from 80 source articles and an independently labelled, balanced natural-validation sample of 600 references (575 valid, 15 inconsistent, 10 unresolved; inter-annotator $\boldsymbol{\kappa = 0.6594}$). Under provider-degraded execution, 542/542 API-queried sample rows recorded at least one provider failure. Endpoint A covered 398/590 reference-standard-decisive cases (**67.46%**), with **TP/TN/FP/FN = 9/134/255/0** and **3.41%** precision; Endpoint B routed all 25 review-positive cases but required human review of 466/600 references. A post-hoc selective-routing probe reduced false positives from 255 to 40, at the cost of reducing decisive coverage from 398/590 to 179/590; human review workload did not decrease. In a separate 320-reference error-enriched set, surname-level author-score decomposition produced no incremental decision changes on four natural author-identity errors. In a DOI-omitted 90-reference cohort (30 conference papers, 30 chapters, 30 theses), the frozen method covered 58/88 decisive cases with five true positives, one false positive, and two false negatives; a within-cohort parser patch increased coverage to 63/88 without reducing false negatives. A bounded 14-pair live-provider comparison recorded six different selected DOIs, seven unchanged DOIs, and one case without a comparable DOI; direct DOI insertion into corrected strings prevents causal attribution of the selected-DOI changes. The evidence supports evidence-traced review routing as an auditable

workflow, while demonstrating the limits of definitive automatic inconsistency claims, provider availability, and within-cohort parser improvements.

Keywords: bibliographic integrity, scientometrics, reference verification, metadata quality, selective abstention, evidence provenance

MSC Classification: 68P20

JEL Classification: C80

Author ORCID: 0000-0001-9259-7112 (https://orcid.org/0000-0001-9259-7112)

1 Introduction

Bibliometric and scientometric analyses depend on the reliable resolution of cited references to scholarly records. Citation counts, citation networks, co-citation structures, bibliographic coupling, and many derived indicators are constructed from links between citing documents and cited works. Errors in reference strings or in associated resolution metadata can fragment citation links, create duplicate representations, prevent record identification, or attach a citation to the wrong target. Reference quality is therefore not merely a formatting concern; reference quality forms part of the data-quality foundation underpinning quantitative studies of science.

Reliability is constrained by provider-dependent scholarly metadata. Coverage, metadata completeness, identifier availability, and curation practices differ across bibliographic databases and open scholarly infrastructures. Culbert et al. (2025), for example, compared reference coverage and selected metadata in OpenAlex, Web of Science, and Scopus and documented the dependence of coverage conclusions on the corpus definition and data source. Such differences matter directly for reference verification because a reference can be correct yet unsupported by one provider, recoverable from another provider, or associated with incomplete or conflicting metadata. A verification method therefore requires addressing two distinct questions: whether the reference is bibliographically consistent, and whether the available evidence is sufficient to determine bibliographic consistency.

Recent work has made substantial progress on both cited-reference matching and broader reference-integrity assurance. Aria et al. (2026) proposed a multi-phase reference-matching algorithm for bibliometric analysis designed to normalize heterogeneous reference strings, uses exact and fuzzy matching, clusters variants, reconciles metadata, and selects canonical representations. Yin et al. (2027) formulate reference integrity as a document-level problem combining multi-source canonical correction with citation–reference topology consistency. Reference matching and document-level correction address related but distinct research problems. Successful matching or document-level correction does not automatically resolve the evaluation problem created by incomplete provider evidence: a reference can resolve to a plausible target while containing an incorrect field, whereas failure to retrieve a candidate does not establish bibliographic inconsistency. The present study therefore focuses

on human-adjudicated field-level integrity, explicit evidence insufficiency, coverage-aware abstention, diagnostic boundary analysis, and transparent preservation of provider-degraded outcomes.

The distinction between matching and integrity verification is especially important when evidence is incomplete. A binary verifier forcing every reference into a valid or invalid decision can convert provider absence or syntax variation into false error claims. A method intended for scholarly-data auditing should instead preserve uncertainty and explicitly abstain when evidence is insufficient. Under an evidence-limited decision contract, abstention is not an implementation failure; abstention constitutes a scientifically meaningful outcome exposing the boundary between observed evidence and unsupported inference. Coverage, provider availability, false-positive diagnostic decomposition, and selective risk therefore become core components of the evaluation contract rather than secondary operational statistics.

Metadata inconsistencies constitute a legitimate object of scientometric analysis. Yeung (2026), for example, examined a recurring typographical error across scholarly platforms and demonstrated the feasibility of investigating cross-platform metadata consistency as a measurable property of the scientific record. More broadly, bibliographic-data cleaning and error assessment influence the reproducibility of downstream analyses. Reference-verification reports should identify the warning, conflicting bibliographic field, supporting evidence source, and decision constraints arising from missing or degraded provider evidence.

Bibliographic reference verification is formulated as a selective, evidence-traced data-quality problem for scientometric workflows. The method separates candidate retrieval from field-level integrity assessment. Title, year, venue, identifier, and structured author evidence are evaluated independently before being combined. To test whether fine-grained author modeling improves natural error detection, author evidence is decomposed into lead-author agreement, order-aware surname-sequence similarity, and surname-set overlap, and the empirical diagnostic boundary is evaluated through both same-evidence ablations and oracle candidate injection. Prespecified hard-conflict rules complement the soft evidence score, and cases without adequate support are routed to abstention or human review.

The empirical evaluation spans six complementary datasets with strict role separation: (i) a 320-case source-record-disjoint controlled benchmark, (ii) a 150-case identifier-free stress benchmark, (iii) an eight-journal, 80-article external reference corpus of 3,712 fidelity-passed references, (iv) a prediction-independent balanced stratified natural-validation cohort of 600 references (75 per journal across all 80 source articles), (v) a 320-reference error-enriched diagnostic cohort designed to stress-test rare author and metadata conflicts, and (vi) a prospective 90-reference DOI-omitted cohort (30 conference papers, 30 book chapters, and 30 theses across 44 source articles) designed to test generalization on non-DOI reference classes. Because the 600-reference natural cohort intentionally balances journals and articles rather than sampling in proportion to article reference-list length, raw cohort proportions are complemented by finite-frame Horvitz–Thompson post-stratification calibration to the 3,712-reference frame.

Six prespecified research objectives organize the analysis:

- **RQ1.** Quantify decisive verification coverage for naturally occurring references in multi-journal bibliographies and DOI-omitted conference, chapter, and thesis strata under incomplete, provider-dependent evidence.
- **RQ2.** Characterize evidence-coverage limitations during the preserved HTTP 429 provider-degraded execution and assess interpretive constraints arising from provider stratification and availability.
- **RQ3.** Assess the incremental detection performance of lead-author, author-order, and surname-set decomposition relative to surname-set matching under frozen retrieval and oracle candidate injection.
- **RQ4.** Quantify the trade-off among decisive coverage, false-positive rule triggers, selective error, and human-review routing.
- **RQ5.** Compare controlled benchmarks, identifier-free stress sets, a DOI-omitted cohort, and the balanced 600-reference natural-validation cohort; identify possible sources of cross-cohort precision differences.
- **RQ6.** Characterize prespecified linkage and representation consequences for 15 human-confirmed inconsistent references using the frozen pre-correction snapshot and paired fixed-linker comparisons.

Bibliographic-integrity verification is examined as an evidence-limited selective decision problem rather than as reference matching or autonomous correction. The contribution consists of an auditable evaluation contract separating candidate retrieval from field-level consistency, records the rule-level trace supporting each disposition, quantifies the exact false-positive taxonomy and parser-syntax interactions on natural references, and evaluates both negative ablation findings (on structured author decomposition) and positive selective-abstention gains (on DOI-omitted conference, chapter, and thesis citations). Unlike the reference-variant consolidation evaluated by Aria et al. (2026) or document-level correction and citation-topology alignment evaluated by Yin et al. (2027), the evaluated framework is positioned strictly as an auditable review-routing and diagnostic triage architecture.

2 Related work

2.1 Reference matching, normalisation, and cited-reference consolidation

Cited-reference matching is a central preprocessing problem in bibliometrics because the same work can appear under multiple string representations. Variation in author rendering, journal abbreviations, punctuation, page ranges, year fields, and identifier presence can fragment citation links and create multiple apparent targets for one cited work. Matching systems therefore typically combine normalisation, blocking, exact identifiers, lexical similarity, and reconciliation rules.

Aria et al. (2026) provide a primary comparator for the reference-matching and canonicalization boundary, addressing heterogeneous reference representations through normalisation, exact matching, candidate restriction, fuzzy comparison, clustering, and metadata reconciliation. Yin et al. (2027) provide a close comparator

at the broader reference-integrity boundary by combining multi-source evidence-grounded correction with document-level citation–reference topology validation across 5,000 manuscripts. The present study does not claim novelty for matching, multi-source correction, or document-level citation alignment. The narrower contribution is a prespecified, coverage-aware evaluation of field-level bibliographic integrity under incomplete provider evidence, using human-adjudicated reference standards, explicit abstention, same-evidence ablation, rule-level false-positive decomposition, and preserved negative and provider-degraded executions.

The conceptual boundary is critical. Reference matching asks whether two or more representations should be assigned to the same cited entity. Integrity verification asks whether the fields printed in a particular reference are consistent with authoritative evidence about the target entity and whether the available evidence is sufficient to support a bibliographic-integrity decision. A matching algorithm can successfully link a corrupted reference to the intended work, whereas an integrity verifier must expose the corrupted field. Conversely, a reference unmatched under a specific provider configuration must not automatically be labelled inconsistent. The present framework therefore treats candidate retrieval as an upstream evidence-acquisition stage and reserves final disposition for a separate field-level verification and abstention layer.

2.2 Bibliographic database coverage and metadata quality

Bibliometric databases differ in scope, curation, source coverage, field completeness, and update behaviour. Culbert et al. (2025) compared OpenAlex with Web of Science and Scopus and documented comparable source-reference coverage within a restricted shared corpus alongside heterogeneous metadata completeness. Delgado-Quirós and Ortega (2025) documented substantial differences in citation counts and reference inclusion across seven free-access scholarly databases. Forchino and Torres-Salinas (2026) reviewed OpenAlex applications and identified persistent metadata-quality limitations, while Rong et al. (2026) demonstrated source-level discrepancies between Web of Science and Crossref affecting citation-analysis quality. The cited comparisons demonstrate the dependence of data-quality conclusions depend on both corpus definition and the specific database snapshot used for evaluation.

For bibliographic verification, provider heterogeneity creates a selective-evidence problem. An absent record or missing field may reflect true non-existence, indexing boundaries (such as national theses outside international indexes), transient API failure, rate limiting, or query mismatch. A verification framework must preserve provider status and make evidence availability observable at row level. In the present study's primary 3,712-reference execution, provider exceptions (specifically HTTP 429 rate-limit responses from OpenAlex and Semantic Scholar) affected 3,357/3,712 corpus rows (90.44%) and 542/542 (100.0%) of the natural-cohort rows passing input parsing (with the remaining 58/600 rows failing at the pre-query `parse_insufficient` stage). The original execution is preserved as a transparent provider-degraded stress snapshot and complemented by (i) post-hoc provider-evidence stratification (Section 6.10), (ii) a paired 129-query provider re-execution on the confirmed natural inconsistencies (Section 6.9), and (iii) an authority-anchored multi-provider evaluation on the prospective 90-reference non-DOI cohort (Section 6.12).

2.3 Metadata inconsistency and integrity assessment

Bibliographic data-quality research has examined malformed identifiers, metadata discrepancies, document-type errors, author-record problems, and inconsistencies between database records and publisher sources. Tani et al. (2013) provide a foundational metadata-quality perspective from digital-library infrastructures, emphasizing the multidimensional nature of metadata quality during creation and aggregation. Nowakowska (2025) places data cleaning, deduplication, field consistency, and cross-source integration within a broader preprocessing framework for bibliometric analysis and emphasizes the continuing necessity of manual intervention at several stages. Maisano et al. (2025) use a semi-automated discrepancy-screening workflow followed by manual confirmation to assess document-type classification errors in Scopus and Web of Science. The cited methodological studies reinforce the role of auditable preprocessing and human-confirmed error assessment in bibliometric data quality.

Yeung (2026) traced a recurring typographical error across platforms to examine metadata consistency as an observable property of scholarly infrastructure. The present framework extends the metadata-consistency perspective to selective verification of individual reference instances, generating a structured evidence trace recording the retrieval route, candidate record, component-level similarity scores, triggered conflict rules, provider failures, and final disposition, consistent with the reproducibility principles emphasized by El Majjaoui et al. (2026).

2.4 Citation verification and automated checking

Recent citation-verification tools illustrate several complementary retrieval and adjudication designs. CiteCheck combines scholarly retrieval with structured verification (Khajavi et al. 2026). CheckIfExist performs multi-source reference validation against scholarly databases (Abbonato 2026). CiteTracer uses a taxonomy-aligned multi-agent workflow with deterministic field matching and escalation of ambiguous cases (Li et al. 2026). RefChecker provides large-scale multi-source verification and was used in the Phantom References study of peer-reviewed proceedings (Russinovich et al. 2026). Because the cited systems differ in evidence stacks, decision logic, and evaluation settings, RefChecker (`academic-refchecker 3.0.190`) is evaluated here on the identical 3,712 frozen reference strings and scored on the 600-reference Human Gold V2 cohort under explicitly documented status-mapping rules (Section 4.8) as a same-input operational comparator rather than a same-provider ablation baseline.

2.5 Uncertainty, abstention, and evaluation boundaries

In selective classification theory, Chow (1970) formalized the recognition error–reject trade-off, and El-Yaniv and Wiener (2010) developed classification under a reject option, trading coverage against conditional risk on the covered subset (see also Geifman and El-Yaniv 2019). The proposed evaluation adopts coverage–risk terminology using a deterministic bibliographic evidence contract rather than a learned neural confidence score. Crucially, the evaluation distinguishes two operational endpoints: Endpoint A (definitive inconsistency detection on covered decisive

records, where both no-candidate `unverified_reference` items and soft/venue-conflict `metadata_mismatch` items are withheld from decisive error claims) and Endpoint B (review-required routing across all records, where inconsistent and unresolved items are treated as review-positive).

3 Data and study design

3.1 Controlled source-record-disjoint benchmark

The controlled benchmark contains 40 source records identified by Digital Object Identifiers (DOIs) and 320 generated cases. The split is performed by source-record family to prevent variants of the same bibliographic record from crossing the development/test boundary. Ten source records define 80 development cases and 30 source records define 240 held-out cases. The fair bibliographic subset contains 180 held-out cases after exclusion of the supplied document-context cases. The controlled benchmark provides auditable case construction and rule sanity checks; the benchmark is not used to estimate the prevalence or distribution of naturally occurring bibliographic errors.

3.2 Identifier-free stress set

A separate 150-case challenge set is generated from the 30 held-out source families. Each family contributes identifier-free valid variants and controlled venue, year, and lead-author substitutions. The set tests behaviour when exact DOI lookup is unavailable and exposes failure modes hidden in DOI-anchored evaluation. Because all variants are protocol-generated, the set serves as a controlled DOI-free stress test rather than natural validation.

3.3 Eight-journal external reference corpus

The primary external corpus contains 3,712 publisher-rendered reference strings extracted from 80 locked primary articles across eight journals (J01–J08). Extraction fidelity was evaluated independently of bibliographic validity, and only fidelity-passed rows entered the external corpus to prevent transcription artifacts from being mislabelled as bibliographic inconsistencies.

The eight-journal frame was prespecified before article-level selection to span information science/scientometrics (*Quantitative Science Studies*, *Scientometrics*, *Frontiers in Research Metrics and Analytics*), computing/AI (*Frontiers in Artificial Intelligence*), Turkish academic publishing contexts (*Bilişim Teknolojileri Dergisi*, *Turkish Journal of Electrical Engineering and Computer Sciences*), and large multidisciplinary transfer venues (*PLOS ONE*, *Scientific Reports*). Eligibility was restricted to calendar-2025 original research, methods, or technical articles with a stable official identifier and a complete official reference list. Within each journal, eligible records were deterministically ordered by the cryptographic hash digest

$$\text{SHA-256}(\texttt{seed} \parallel \texttt{journal_ID} \parallel \texttt{normalized_identifier})$$

using the frozen seed `IPM-RWNEC-2025-v1`, where ‖ denotes string concatenation; the first 10 technically eligible records formed the primary sample and ranks 11–15 formed the reserve pool. Supplementary Tables S1–S3 in Online Resource 1 and machine-readable registries in Online Resources 2 and 3 document the complete frame.

3.4 Balanced stratified natural-validation cohort

The primary natural evaluation uses a deterministic, prediction-independent, article-balanced stratified sample of 600 references from the 3,712-reference corpus: exactly 75 references from each of the eight journals, with all 80 source articles represented (five articles contributing eight references and five contributing seven per journal). Selection used a fixed cryptographic salt without access to system predictions, prior labels, provider outcomes, or conflict flags.

Independent dual evaluation was conducted in separate blinded workbooks assigned to two external doctoral-level researchers experienced in scholarly publishing, bibliometrics, and bibliographic metadata verification (`Annotator A` / `REV-META-2026-A` and `Annotator B` / `REV-CITE-2026-B`), operating outside the author's department with documented independence and conflict-of-interest attestations. Each workbook contained the same 600 reference instances in a distinct deterministic row order without system predictions, similarity scores, retrieval candidates, diagnostic traces, or the other evaluator's decisions. Each row was assigned a primary status (`valid`, `inconsistent`, or `unresolved_or_ambiguous`) and nine boolean field-error flags (`wrong_doi`, `wrong_title`, `wrong_author_identity`, `wrong_author_order`, `wrong_year`, `wrong_venue`, `wrong_volume_issue_pages`, `duplicate_reference`, and `other_bibliographic_inconsistency`), together with primary authority URLs and evidence notes. The locked 600-reference Human Gold V2 standard is strictly evaluation-only under an evaluation firewall prohibiting any threshold, weight, or rule tuning.

3.5 Sampling calibration and scope of frequency statements

To distinguish the balanced validation-cohort composition from the composition of the complete 3,712-reference frame, a finite-population Horvitz–Thompson / post-stratification calibration was computed using the known article reference-list sizes N_a ($a = 1, \ldots, 80$), the article validation sample sizes $n_a \in \{7, 8\}$, and the sampled count $y_a(s)$ of references with status s in article a:

$$p_{\mathrm{cal}}(s) = \frac{1}{3712} \sum_{a=1}^{80} N_a \frac{y_a(s)}{n_a}.$$

Because within-stratum selection via cryptographic SHA-256(`seed` ‖ `ID`) ranking induces a deterministic pseudo-random permutation functioning operationally like simple random sampling within strata (SRS within articles) without formal randomisation draws, $p_{\mathrm{cal}}(s)$ corrects for unequal article reference-list sizes (N_a ranging from 16 to 101) and is reported conservatively as a descriptive finite-corpus calibration

without a design-based confidence interval. All primary classifier performance metrics remain computed on the unweighted 600-reference cohort, and all coded A/B workbooks, lock digests, and coordinator reconciliation records are preserved in the versioned reproducibility archive.

3.6 Annotation, evidence hierarchy, locking, and adjudication

The annotation unit was one publisher-rendered reference-string instance from one source-article reference list. Evaluation targeted substantive bibliographic consistency rather than house formatting style: punctuation, capitalization, standard ISO-4 journal abbreviations, and harmless surface variations were not treated as errors. An `inconsistent` label required at least one positive field-level error flag corroborated by an authoritative external source.

Evidence was verified according to a five-tier primary hierarchy: (1) official publisher article or book landing page, (2) DOI registration record, (3) Crossref metadata record, (4) official conference proceedings volume or table of contents, and (5) official national or institutional thesis repository (such as the Turkish Council of Higher Education National Thesis Center, YÖK Ulusal Tez Merkezi, or ProQuest Dissertations & Theses), followed by other primary bibliographic catalogs. Secondary scholarly aggregators could assist discovery, but absence from Crossref, OpenAlex, Semantic Scholar, or any single index was strictly prohibited as proof of inconsistency (fail-closed non-retrieval rule).

Annotator A completed and locked all 600 rows at `2026-09-30 07:37 UTC`, and Annotator B completed and locked all 600 rows at `2026-09-30 09:20 UTC`. Pre-adjudication agreement was computed immediately upon the second lock, yielding 588/600 (98.00%) exact primary-status agreement ($\kappa = 0.659381$) and leaving only 12 disagreement rows. Because both evaluators had already embedded the primary publisher, DOI, and repository URLs alongside field-level verification rationales inside the separate evaluator workbooks during the multi-day independent annotation campaign, the adjudication review only needed to inspect and reconcile the 12 pre-documented rows against the five-tier authority hierarchy, enabling the adjudicated consensus record to be verified and locked 40 minutes later at `2026-09-30 10:00 UTC`.

Subsequent post-lock provenance quality assurance preserved the independent A/B files untouched while adding negative-search documentation for seven consensus-unresolved rows, identifying one deterministic within-article duplicate (`J07-P04-R063`, an exact repeat of `J07-P04-R036` in the 3,712-reference frame, versioned into Gold V2 as `inconsistent` with `duplicate_reference = true`), and recording a dated provenance corrigendum for two Turkish thesis records in article `J01-P08` (`J01-P08-R023`, verified via YÖK Thesis No. 846267 as a 2024 Hasan Kalyoncu University doctoral dissertation; and `J01-P08-R024`, verified via YÖK Thesis No. 478770 as a 2017 SANKO University master's thesis, retaining the locked `inconsistent` status due to truncated/missing thesis venue metadata in the published article). Table 1 summarizes the frozen annotation and adjudication contract.

Table 1 Frozen human-annotation and adjudication contract for the 600-reference cohort

Stage / role	Frozen procedure	Scientific boundary
Sampling and unit	600 prediction-independent references; 75 per journal across all 80 articles; one published reference-string instance per row	No system prediction, provider status, or conflict flag used for sampling.
Annotator A role	Independent blinded bibliographic-consistency labeling (`REV-META-2026-A`); 600 rows; locked 2026-09-30 07:37 UTC	Separate blinded workbook and deterministic row permutation.
Annotator B role	Independent blinded bibliographic-consistency labeling (`REV-CITE-2026-B`); 600 rows; locked 2026-09-30 09:20 UTC	Identical 600 references in a distinct deterministic row order.
Evaluator qualifications	Two independent doctoral-level researchers in information/document management and computing, unaffiliated with Marmara University	Signed independence and conflict-of-interest attestations archived in restricted governance register.
Blinding	No system predictions, scores, candidates, diagnostic traces, or cross-annotator labels exposed	Coordinator reconciliation workspace restricted until both locks completed.
Primary labels	`valid` / `inconsistent` / `unresolved_or_ambiguous`	Unresolved status is strictly distinct from confirmed inconsistency.
Field diagnosis	Nine boolean field-error flags recorded for every row	`inconsistent` requires ≥ 1 positive flag; multiple flags permitted.
Evidence hierarchy	Publisher page → DOI registration → Crossref → official proceedings → official thesis repository → primary catalog	Secondary indexes serve as discovery aids only.
Non-retrieval rule	Absence from any single provider is not proof of inconsistency	Fail-closed `unresolved_or_ambiguous` status retained when authority evidence is incomplete.
Agreement & adjudication	Computed after both locks (588/600 [98.00%], $\kappa = 0.6594$); 12 disagreement rows reconciled and locked at 2026-09-30 10:00 UTC	Pre-logged primary authority URLs enabled rapid reconciliation of the 12 discordant rows.
Evaluation firewall	Locked Human Gold V2 (575 valid, 15 inconsistent, 10 unresolved) is evaluation-only	Zero threshold, weight, conflict-rule, or ablation selection permitted on Gold V2.
Post-lock provenance QA	1 exact within-article duplicate (`J07-P04-R063`) versioned into Gold V2; YÖK thesis IDs 846267 and 478770 documented	Original A/B locks and initial adjudication preserved as immutable provenance.

3.7 Error-enriched diagnostic cohort

Because naturally occurring author-identity and author-order errors were sparse in the 600-reference balanced cohort, a separate 320-reference error-enriched diagnostic cohort (40 references per journal: 240 conflict-heavy author-signal candidates plus 80 clean controls) was constructed after the natural cohort had been locked. The diagnostic cohort is used exclusively for rare-error diagnostic sensitivity, structured-author ablation, and oracle candidate injection analysis; the diagnostic cohort is never pooled with the 600-reference cohort for prevalence estimation or corpus calibration.

3.8 Provider availability and primary execution conditions

The primary external execution queried the Crossref REST API (Crossref 2020), OpenAlex (Priem et al. 2022), and Semantic Scholar (Ammar et al. 2018) across

the 3,712-reference corpus. Provider HTTP exceptions (predominantly HTTP 429 rate-limit responses from OpenAlex and Semantic Scholar) were preserved without imputation as a realistic unauthenticated/rate-limited stress snapshot. RQ2 is evaluated without overwriting the primary snapshot by combining post-hoc provider-evidence stratification across all 600 natural-cohort rows (Section 6.10) with two empirical multi-provider availability analyses: (i) a paired 129-request live provider re-query across the 14 replaceable inconsistent natural-cohort references (Section 6.9), and (ii) a full multi-provider authority-anchored evaluation across the prospective 90-reference DOI-omitted cohort (Section 6.12).

3.9 Prospective DOI-omitted citation evaluation cohort ($n = 90$)

To evaluate the generalizability of the locked selective verification framework to reference classes routinely omitting a printed DOI (`citation_doi_present = no`), a prospective 90-reference stratified cohort (30 conference papers, 30 book chapters, and 30 master's/doctoral theses across 44 unique source articles) was constructed from a pre-sealed three-journal DergiPark frame (*Gazi University Journal of Science Part C*, *Turkish Journal of Education*, and *Academia Eğitim Araştırmaları Dergisi*; 17 issues, 78 locked articles; frame SHA-256: `63c4d477...`, pre-collection seal SHA-256: `76b14237...`). Across 454 candidate reference positions, 370 complete reference strings were verified against the 3,712-reference primary corpus (with 84 HTML-absent out-of-pool positions permanently closed and 13 overlapping or malformed entries excluded in `EXCLUSION_LEDGER_REAL.csv`), leaving 257 eligible non-DOI candidates (60 conference papers, 69 book chapters, and 128 theses). Applying the prespecified within-stratum SHA-256 ordering (`R35-P2-019_PROSPECTIVE_ORDER_20261010_V1_FROZEN_BEFORE_POOL`) selected the frozen 90-reference cohort (`GATE1_SELECTED_90.csv`, SHA-256: `a0fdc80d...`).

Primary bibliographic authority verification across IEEE Xplore, Crossref, ACM Digital Library, ACL Anthology, NeurIPS Proceedings, SpringerLink, Routledge/Taylor & Francis, Wiley, Oxford University Press, ProQuest Dissertations & Theses Global, institutional DSpace repositories, and YÖK Ulusal Tez Merkezi confirmed the existence of all 90 cited works as authentic scholarly items (with 0/90 secondary or ResearchGate links in the primary authority register) and identified publisher/Crossref canonical DOIs omitted from the printed citations for 38/90 works (42.22%: 20/30 conference papers, 18/30 book chapters, and 0/30 theses). Under two-stage protocol-governed dual assessment (`REV-META-2026-A` tolerance-aware pass vs. `REV-CITE-2026-B` strict surface-form pass, locked prior to verifier scoring), exact three-class agreement was 86/90 (95.56%; 95% Wilson CI: 89.12–98.26%), with three-class Cohen's $\kappa = 0.8238$, binary review-positive $\kappa = 0.8209$, and PABAK $= 0.9111$. Reconciliation established a locked Gold standard of 77 `valid` (72 clean exact matches plus 5 single-character/lexical surface variants), 11 `inconsistent`, and 2 `unresolved_or_ambiguous` references (88 decisive rows and 13 review-positive rows). Table 2 summarizes all six study datasets and the respective evaluation boundaries; the experimental performance on the 90-reference cohort is reported in Section 6.12.

Table 2 Study datasets and evaluation roles across the six experimental cohorts

Dataset	Size and construction	Primary role	Interpretive boundary
Controlled benchmark	320 generated cases from 40 DOI-anchored source records; source-record-disjoint split (80 dev / 240 test; 180 fair bibliographic subset)	Parser, conflict-rule, and implementation sanity check	Protocol-generated perturbations; not used to estimate natural error frequencies.
Identifier-free stress set	150 cases from 30 held-out source families with controlled venue, year, and lead-author substitutions	Stress testing when persistent identifiers are absent	Protocol-generated stress test; not a natural-frequency sample.
External reference corpus	3,712 publisher-rendered, extraction-fidelity-passed references from 80 articles across 8 journals (J01–J08)	Sampling frame and operational stress execution	Primary unauthenticated API run experienced HTTP 429 rate limits on 90.44% of rows.
Balanced natural validation cohort	600 prediction-independent references (75/journal; 7–8/article across all 80 articles; 575 valid, 15 inconsistent, 10 unresolved)	Primary human-adjudicated validation, V0–V8 ablation, and RQ6 consequence audit	Cohort-specific raw rates complemented by Horvitz–Thompson finite-frame calibration.
Error-enriched diagnostic cohort	320 references (240 author-signal conflict candidates + 80 clean controls; 40/journal; 303 valid, 12 inconsistent, 5 unresolved)	Rare-error diagnosis, structured-author ablation, and oracle candidate injection	Error-enriched stress sample; excluded from natural prevalence and corpus calibration.
Prospective DOI-omitted cohort	90 non-DOI references (30 conference papers, 30 book chapters, 30 theses across 44 articles in 3 journals; 77 valid, 11 inconsistent, 2 unresolved)	Generalization test on non-DOI strata and paired evaluation of `V1_FROZEN` vs. `V1_PATCHED`	Prospectively sealed non-DOI stratified cohort; evaluated with full primary-authority verification.

4 Verification method

The stages and independent evaluation boundaries of the verification architecture are illustrated in Fig. 1.

4.1 Input and parsing

The fundamental input is one raw reference string. Conservative Unicode normalisation (NFKD), accent stripping, Turkish dotless `ı→ i’ mapping, and DOI extraction are applied first. In the frozen primary parser (V1 / V1_FROZEN), title, author, venue, and year fields are extracted using citation-structure priorities: (1) quoted title segments ("..." or “...”)`, (2) APA-style parenthesized four-digit years `((19|20)\d{2})` following the author segment, (3) bare author-year patterns `. YYYY.`, and (4) a positional fallback. As diagnosed in Section 6.5, when a journal uses an unquoted Vancouver or numbered house style with the publication year appearing at the end of the citation after the journal title (e.g., in *PLOS ONE*, *Scientific Reports*, or *TÜBİTAK* engineering styles), splitting at the late year places the article title into the author segment and leaves only the trailing volume/page tokens as the parsed title.

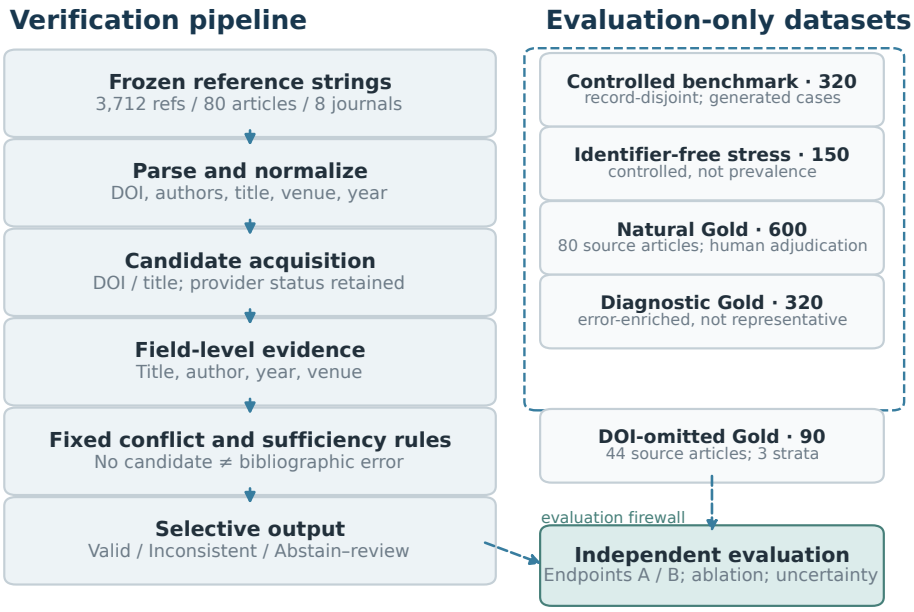


Fig. 1 Evidence-traced verification pipeline and evaluation-only study datasets. The human-annotated natural, diagnostic, and DOI-omitted cohorts are separate from candidate acquisition, rule evaluation, and selective decision routing. Controlled and identifier-free stress tests are not used to estimate naturally occurring bibliographic-error frequencies.

To evaluate parser boundary sensitivity on non-DOI humanities, education, and engineering citations without altering any downstream retrieval thresholds, weights, or conflict rules, a strictly isolated candidate parser patch is additionally specified (`V1_PATCHED`, tested prospectively on the 90-reference DOI-omitted cohort in Section 6.12). Relative to `V1_FROZEN`, `V1_PATCHED` introduces exactly two deterministic regex normalizations during input pre-parsing: (i) recognizing APA conference date expressions of the form `((19|20)\d{2},\s*[A-Za-z]+(?:\s+\d{1,2})?)` (e.g., `(2022, July)` or `(2021, November 14-16)`) as the publication-year delimiter, preventing loss of the article title before the conference name, and (ii) restricting quote-delimiter splitting to outer typographic quotes, preventing internal apostrophes or contractions inside thesis titles (e.g., `"I can read accurately but can't understand..."`) do not prematurely truncate the extracted title string.

4.2 Candidate retrieval

Candidate acquisition is separated from verification. Exact DOI lookup is attempted first when an identifier is present. If exact DOI evidence is unavailable, the highest-scoring fused title candidate across queried providers is accepted only when the lexical

title similarity satisfies

$$s_{\text{title}}(r, c) = 0.55\, s_{\text{seq}}(r, c) + 0.45\, s_{\text{Jaccard}}(r, c) \geq 0.42;$$

otherwise no candidate is selected and the reference is assigned `unverified_reference` with `abstained = true` (withheld from Endpoint A and routed to Endpoint B review). Conversely, an exact DOI candidate with retrieved title similarity below 0.72 triggers an immediate `doi_title_mismatch` hard conflict.

4.3 Field-level evidence

For a retrieved candidate c, field-level similarity scores $s_k \in [0, 1]$ are computed independently for $k \in \{\text{title}, \text{author}, \text{year}, \text{venue}\}$. The primary soft evidence score is the linear combination

$$C(r, c) = \sum_{k \in \{\text{title,author,year,venue}\}} w_k s_k, \qquad \sum_k w_k = 1, \tag{1}$$

with frozen weights $w_{\text{title}} = 0.35$, $w_{\text{author}} = 0.30$, $w_{\text{year}} = 0.20$, and $w_{\text{venue}} = 0.15$ (identifier evidence receives weight 0.00 in $C(r, c)$ because DOI agreement already governs candidate routing and the identifier–title gate). Year similarity is $s_{\text{year}} = 1.0$ for $|\Delta\text{year}| = 0$, 0.5 for $|\Delta\text{year}| = 1$, and 0.0 otherwise.

In the frozen implementation, an unparsed or provider-missing component receives $s_k = 0.0$ without renormalising the remaining weights. From a selective-classification perspective, assigning $s_k = 0.0$ to missing metadata creates a tension with the principle prohibiting the interpretation of absent evidence as evidence of inconsistency (for example, a valid book or report lacking a journal venue automatically loses $w_{\text{venue}} = 0.15$). Importantly, however, the frozen architecture mitigates the missing-evidence penalty through the Endpoint A disposition mapping (Section 4.5): rows failing *only* the soft threshold $C(r, c) < 0.65$ without triggering the identifier–title or strong-title author/year hard-conflict gates receive the diagnostic label `metadata_mismatch`, mapped to `coverage_status = review_only` (withheld from Endpoint A decisive inconsistency predictions and routed solely to Endpoint B human review). The exact effect is further quantified of weight renormalisation and missing-field abstention in Section 6.7.

4.4 Structured author diagnostics

To make author-list discrepancies inspectable at component level, structured author similarity is decomposed into three terms:

$$A_{\text{author}} = 0.45\, A_{\text{lead}} + 0.35\, A_{\text{order}} + 0.20\, A_{\text{set}}, \tag{2}$$

where $A_{\text{lead}} \in \{0, 1\}$ indicates exact agreement of the normalized first-author surname (preserving lowercase surname particles such as *van der*, *de la*, and *el*), $A_{\text{order}} \in [0, 1]$ is the Ratcliff/Obershelp sequence similarity of the ordered surname lists, and $A_{\text{set}} \in$

$[0, 1]$ is the surname-set Jaccard overlap. While the structured-author decomposition is designed to prevent long co-author lists from masking first-author substitutions or order inversions, the empirical utility of structured-author decomposition on natural errors is treated as a falsifiable hypothesis (RQ3) tested via same-evidence ablations (V1–V6) and oracle candidate injection.

4.5 Conflict rules and selective disposition

Deterministic conflict rules are evaluated in strict order before the soft-score threshold ($\tau = 0.65$), producing both a fine-grained diagnostic label and a two-endpoint selective disposition:

1. **No-candidate abstention (`unverified_reference`):** If no candidate meets the retrieval contract, the system sets `abstained = true` and `coverage_status = abstained` (Endpoint A: withheld; Endpoint B: review-required = 1).
2. **Identifier–title hard conflict (`doi_title_mismatch`):** If an exact DOI candidate has $s_{\text{title}} < 0.72$, the system assigns `doi_title_mismatch` with `coverage_status = covered` (Endpoint A = 1; Endpoint B = 1).
3. **Strong-title author/year hard-conflict gate (`author_year_mismatch`):** If $s_{\text{title}} \geq 0.72$ and any of four sub-rules fires—(a) `year_hard_conflict` ($|\Delta\text{year}| > 0$), (b) `author_hard_conflict` ($A_{\text{author}} < 0.55$), (c) `partial_author_year_conflict` ($s_{\text{year}} < 1.0 \wedge A_{\text{author}} < 0.95$), or (d) `lead_author_hard_conflict` (lead surnames differ $\wedge s_{\text{year}} = 1.0 \wedge s_{\text{venue}} \geq 0.70$)—the system assigns `author_year_mismatch` with `coverage_status = covered` (Endpoint A = 1; Endpoint B = 1).
4. **Venue hard conflict and post-gate soft threshold (`metadata_mismatch` vs. `valid`):** If $s_{\text{title}} \geq 0.72$, $A_{\text{author}} \geq 0.95$, $s_{\text{year}} = 1.0$, and a non-empty submitted venue has $s_{\text{venue}} < 0.70$, the system assigns `metadata_mismatch`. Otherwise, if $C(r, c) < 0.65$, the verifier assigns `author_year_mismatch` when $s_{\text{title}} \geq 0.72 \wedge (A_{\text{author}} < 0.95 \vee s_{\text{year}} < 1.0)$ and `metadata_mismatch` otherwise; if $C(r, c) \geq 0.65$, the verifier assigns `valid` (`coverage_status = covered`, Endpoint A = 0, Endpoint B = 0). Crucially, under the frozen Endpoint A/B mapping contract, `metadata_mismatch` is assigned `coverage_status = review_only` (withheld from Endpoint A's decisive inconsistency predictions and routed to Endpoint B human review).

4.6 Frozen reproducibility contract

Table 3 records the complete parameter and decision-rule contract for the primary `proposed_hybrid` (`V1`) configuration, including the exact Endpoint A/B disposition mapping and the monotonically decreasing logistic review-priority score $s_{\text{review}} = 1/(1 + \exp(10 \times (C(r, c) - \tau)))$.

4.7 Evidence trace and auditability

Every evaluated reference retains a machine-readable audit trace recording the raw string, parsed fields, retrieval route, provider error count, selected candidate metadata,

Table 3 Frozen parameter and decision-rule contract for the primary `proposed_hybrid` (`V1`) configuration

Component	Frozen setting	Operational rule and Endpoint A/B mapping
Text normalisation	Unicode NFKD; strip accents; casefold; map Turkish dotless 'ı→ i'; strip `doi.org/doi:` prefixes; collapse whitespace	Applied prior to lexical title/venue comparison and surname extraction.
Title / venue similarity	$0.55 \times$ sequence similarity $+ 0.45 \times$ token Jaccard	Lexical similarity $s_{\text{title}}, s_{\text{venue}} \in [0, 1]$.
Candidate routing	Exact DOI lookup first; otherwise highest fused title candidate	Title-route candidate accepted only if $s_{\text{title}} \geq 0.42$.
Retrieval title threshold	$s_{\text{title}} \geq 0.42$	Below 0.42 $\implies$ no candidate (`unverified_reference`).
Soft evidence weights	$w_{\text{id}} = 0.00$, $w_{\text{title}} = 0.35$, $w_{\text{author}} = 0.30$, $w_{\text{year}} = 0.20$, $w_{\text{venue}} = 0.15$	Identifier governs routing and hard-conflict gates; not double-counted in $C(r, c)$.
Missing soft evidence	Missing component $s_k = 0.0$; no weight renormalisation in `V1`	Tested against weight-renormalised and abstention-transfer probes in Section 6.7.
Year evidence	$\lvert\Delta\text{year}\rvert = 0 \implies 1.0$; $\lvert\Delta\text{year}\rvert = 1 \implies 0.5$; else 0.0	Soft 1-year tolerance is overridden when the strong-title hard year gate fires.
Structured author score	$A_{\text{author}} = 0.45\, A_{\text{lead}} + 0.35\, A_{\text{order}} + 0.20\, A_{\text{set}}$	Surname-level extraction preserves compound surname particles but does not test given-name initials.
Identifier–title hard conflict	Exact DOI hit $\wedge\ s_{\text{title}} < 0.72$	Immediate `doi_title_mismatch`; Endpoint A = 1 (`covered`), Endpoint B = 1.
Strong-title author/year gate	$s_{\text{title}} \geq 0.72\ \wedge\ \big(\lvert\Delta\text{year}\rvert > 0\ \vee\ A_{\text{author}} < 0.55\ \vee\ (s_{\text{year}} < 1.0 \wedge A_{\text{author}} < 0.95)\ \vee$ (lead surname mismatch $\wedge\ s_{\text{year}} = 1.0 \wedge s_{\text{venue}} \geq 0.70$)$\big)$	Immediate `author_year_mismatch`; Endpoint A = 1 (`covered`), Endpoint B = 1.
Venue hard conflict	Non-empty venue $\wedge\ s_{\text{title}} \geq 0.72 \wedge A_{\text{author}} \geq 0.95 \wedge s_{\text{year}} = 1.0 \wedge s_{\text{venue}} < 0.70$	Yields `metadata_mismatch`; withheld from Endpoint A (`review_only`), Endpoint B = 1.
Post-hard-rule soft threshold	$C(r, c) < 0.65 \implies$ `metadata_mismatch` (or `author_year_mismatch` if $s_{\text{title}} \geq 0.72$); $C(r, c) \geq 0.65 \implies$ `valid`	`metadata_mismatch` maps to `review_only` (Endpoint B = 1); `valid` maps to `covered` (Endpoint A = 0, Endpoint B = 0).
No-candidate disposition	`unverified_reference`; `abstained = true`	Endpoint A withheld (`abstained`); Endpoint B routes to review (= 1).
Heuristic review score	$s_{\text{review}} = 1/\big(1 + \exp(10 \times (C(r, c) - \tau))\big)$, with $\tau = 0.65$	Monotonically increases as $C(r, c)$ falls below τ; used for queue ranking, not calibrated probability.

component scores ($s_{\text{title}}, A_{\text{author}}, s_{\text{year}}, s_{\text{venue}}$), triggered conflict rule, coverage status (`covered`, `review_only`, or `abstained`), and Endpoint A/B decisions.

4.8 Comparison boundary and operational comparator

To provide external operational context on the identical reference strings, all 80 articles (3,712 references) were processed with `academic-refchecker 3.0.190` in core metadata-verification mode with LLM and enrichment paths disabled (`P0-12CJ_R14`). Prior to scoring on the 600-reference Human Gold V2 cohort, RefChecker's six native output statuses were mapped deterministically to Endpoint A and Endpoint B semantics:

- **Endpoint A (definitive inconsistency detection):** RefChecker native `clean` maps to 0 (`covered` valid); native `error` (and `warning`, when present) maps to 1 (`covered` inconsistent); and native `unverified`, `provider_error`, and `extraction_error` map to `abstained` (withheld from Endpoint A).
- **Endpoint B (review-required routing):** Native `clean` maps to 0 (no manual review required), whereas all non-clean statuses (`error`, `warning`, `unverified`, `provider_error`, and `extraction_error`) map to 1 (routed to human review).

Across the 3,712-reference corpus, RefChecker emitted 496 `clean`, 1,722 `error`, 1,494 `unverified`, and 0 `warning`, `provider_error`, or `extraction_error` rows (and in the 600-reference cohort: 71 `clean`, 283 `error`, 246 `unverified`, and 0 `warning`). Because zero rows received a native `warning` status, RefChecker's false-positive rate in Endpoint A reflects substantive native `error` flags (e.g., year, author, or DOI mismatches across the native RefChecker provider stack) rather than mild formatting warnings.

4.9 Use of generative AI tools

During manuscript preparation, the author used OpenAI ChatGPT to assist with language editing, structural organization, and documentation of code-based quality-assurance checks. All empirical data, bibliographic authority verifications, and statistical computations were generated from the versioned datasets and deterministic Python scripts. Following tool-assisted editing, the author reviewed and edited the content as needed and takes full responsibility for the integrity and accuracy of the manuscript.

5 Evaluation protocol and statistical analysis

5.1 Primary balanced natural-validation-cohort metrics and clustering rationale

After the human labels and adjudication decisions were locked, the 600-reference balanced natural-validation cohort was evaluated using coverage, conditional precision, conditional recall, specificity, balanced accuracy, Matthews correlation coefficient (MCC), selective error, and review-queue workload reduction, with F1 retained as a secondary descriptive metric.

Uncertainty estimation for all natural-cohort, diagnostic-cohort, and prospective-cohort metrics uses 2,000 paired article-clustered bootstrap resamples (`seed = 20261001`) with the source article (`article_id`: $M = 80$ clusters in the natural and diagnostic cohorts; $M = 44$ clusters in the 90-reference prospective cohort) as the resampling unit. Choosing the source article rather than the individual reference as the resampling unit is statistically essential because bibliographic errors exhibit extreme intra-article clustering driven by author-specific citation practices and journal copy-editing workflows: as shown in Section 6.3, 13 of the 15 confirmed inconsistent references (86.7%) in the 600-reference cohort originate from just two journals (`J01`: 6; `J04`: 7), and 8 of the 15 inconsistencies (53.3%) are concentrated within just two

individual source articles (J01-P08: 4 errors; J04-P06: 4 errors). Treating references as independent observations would severely underestimate sampling variance.

5.2 Structured-author ablation and parser sensitivity variants

The prespecified same-evidence ablation family on the 600-reference natural cohort and 320-reference diagnostic cohort compares the full method against eight variants holding the retrieved candidate snapshot fixed:

- **V0 – Metadata baseline:** Retrieval-aware title, year, and venue metadata scoring without structured author decomposition or author hard-conflict gates.
- **V1 – Full prespecified method:** Complete structured-author scoring ($0.45\,A_{\text{lead}} + 0.35\,A_{\text{order}} + 0.20\,A_{\text{set}}$), deterministic hard-conflict gates, and selective abstention (V1_FROZEN on the 90-reference cohort).
- **V2 – No lead-author score component:** Removes A_{lead} and renormalises author weights (0.00/0.60/0.40).
- **V3 – No author-order component:** Removes A_{order} and renormalises author weights (0.60/0.00/0.40).
- **V4 – No surname-set component:** Removes A_{set} and renormalises author weights (0.55/0.45/0.00).
- **V5 – Undecomposed surname-set aggregate:** Replaces Equation (2) with surname-set Jaccard overlap (A_{set}) alone.
- **V6 – No lead-author hard-conflict rule:** Retains Equation (2) but disables the standalone lead-author hard-conflict trigger.
- **V7 – Forced decision without abstention:** Forces a binary valid/inconsistent decision across all rows by converting abstentions and review-only dispositions into binary predictions.
- **V8 – Retrieval-only disposition:** Classifies any reference with a retrieved candidate as valid and abstains only when no candidate is retrieved.

On the prospective 90-reference DOI-omitted cohort, V1_PATCHED is additionally evaluated (the candidate parser patch defined in Section 4.1) paired against V1_FROZEN, V0, and V7.

5.3 Weight, threshold, and missing-component renormalization sensitivity

Prespecified robustness probes vary the author-weight vector (equal, lead-heavy, order-heavy, set-heavy, no-lead, no-order, and no-set), the overall metadata weights (title-heavy, author-heavy, year-heavy, equal-field, and venue-redistributed), and the primary decision thresholds (retrieval-title = 0.42, identifier–title conflict = 0.72, author conflict = 0.55, venue conflict = 0.70, and final evidence = 0.65). In addition, to address the methodological question of whether missing parsed components ($s_k = 0.0$) should be renormalised or routed to abstention, two supplementary post-hoc diagnostic probes are evaluated on the 600-reference cohort: (i) renormalising $C(r, c)$ over non-zero available components $A(r, c) = \{k : s_k > 0\}$, and (ii) routing

single-cue or syntax-ambiguous hard conflicts (where unquoted late-year reference syntax yields $s_{\text{title}} < 0.20$ despite exact DOI and year agreement, or where $s_{\text{title}} \geq 0.72$ and $|\Delta\text{year}| \leq 1$ accompany $A_{\text{author}} < 0.55$) to `review_only` (withheld from Endpoint A) rather than definitive Endpoint A inconsistency.

5.4 Selective risk, coverage, and hard-gate dominance

Selective evaluation records the relationship between decision coverage and selective error risk on the covered subset. Importantly, no conventional Area Under the Risk–Coverage Curve (AURC) summary is claimed across the soft-threshold grid around $\tau = 0.65$. The structural explanation lies in the frozen decision order: deterministic hard-conflict gates (`doi_title_mismatch` and the strong-title `author_year_mismatch` gate) execute *before* the soft-score threshold $\tau = 0.65$, and 100% of the 264 Endpoint A inconsistency predictions (9 TP + 255 FP) in the 600-reference cohort are triggered by the upstream hard gates, while post-gate soft-threshold failures ($C(r, c) < \tau$) map to `metadata_mismatch` (`review_only`, withheld from Endpoint A). Consequently, varying τ within a narrow grid while holding the hard-conflict gates fixed does not generate a monotonic risk–coverage trajectory on Endpoint A.

5.5 Statistical analysis and effective multiplicity

For paired ablation comparisons, each of the $B = 2,000$ bootstrap iterations draws a with-replacement sample of 80 source-article clusters and applies the identical article sequence to `V1` and every comparator V_j, storing the paired difference $\Delta_b = \text{BA}_b(\text{V1}) - \text{BA}_b(V_j)$. The exploratory two-sided resampling sign-tail p-value is computed as

$$p_i = \min\left\{1,\ 2 \times \min\left[\frac{n(\Delta \geq 0) + 1}{B + 1},\ \frac{n(\Delta \leq 0) + 1}{B + 1}\right]\right\}. \tag{3}$$

Including zero-difference bootstrap replicates ($\Delta = 0$) in both $n(\Delta \geq 0)$ and $n(\Delta \leq 0)$ makes Equation (3) a strictly conservative upward finite-sample adjustment: when two variants produce identical row-level predictions on the cohort ($\Delta_b \equiv 0$ for all $b = 1, \ldots, 2000$), both tail fractions equal $(2000 + 1)/2001 = 1$ and $p_i = 1.000$.

Holm–Bonferroni step-down multiplicity correction, $p_{\text{holm},(i)} = \max_{k \leq i} \min\{1, (m - k + 1)p_{(k)}\}$, is applied across the prespecified seven-comparison primary family (V1 vs. V0, V2, V3, V4, V5, V6, V7, with $m = 7$; V8 is excluded as a retrieval-only reference point). Because V3 and V6 produce zero row-level differences relative to V1 on the 600-reference cohort ($\Delta \equiv 0$), the effective number of non-degenerate hypotheses is $m_{\text{eff}} = 5$; both the conservative $m = 7$ Holm-adjusted p-values and the $m_{\text{eff}} = 5$ values are reported in Section 6.7.

5.6 Provider-availability sensitivity and post-hoc stratification

To describe the association between provider rate-limiting (HTTP 429) and observed coverage without altering the primary provider-degraded snapshot (`PO-12CC`), the 600

natural-validation rows are first stratified into three mutually exclusive observed evidence states: (i) no recorded provider error with `parse_insufficient` input and no candidate ($n = 58$), (ii) provider-failure exposed with a candidate recovered ($n = 487$), and (iii) provider-failure exposed with no candidate recovered ($n = 55$). A separate 129-request live-provider comparison is then described on 14 corrected/original reference pairs (`R35-P2-014` / `R35-P2-021`) and the DOI-omitted 90-reference cohort; neither is a matched provider-restoration experiment on all 600 reference instances.

5.7 Scientometric consequence analysis

To answer RQ6, linkage and representation consequences are evaluated for the 15 human-confirmed inconsistent references in the 600-reference cohort at two complementary levels. First, under the strict pre-rerun lower-bound audit, cases are classified using only the frozen pre-correction snapshot and authoritative target records into (i) *directly demonstrated consequence* (5/15), (ii) *confirmed unchanged bounded identifier/link* (1/15), and (iii) *pre-rerun indeterminate* (9/15). Second, to eliminate the 9-case indeterminacy, the coordinator-approved corrections for all 14 replaceable inconsistent references are processed through the frozen citation-linking and verification pipeline alongside the 1 counterfactual duplicate-consolidation operation (`R35-P2-014` / `R35-P2-016`), yielding a complete 15/15 before-and-after target-linkage and verification comparison (Section 6.11). No journal impact factor, h-index, or global citation-count extrapolation is claimed.

6 Results

6.1 Controlled and stress-set benchmark performance

On the source-record-disjoint controlled bibliographic held-out subset ($n = 180$), the proposed hybrid (`V1`) achieved precision = 1.000, recall = 1.000, F1 = 1.000, and accuracy = 1.000. The weighted-metadata baseline achieved F1 = 0.792, while the retrieval-metadata baseline (`V0`) achieved F1 = 0.994 (one fewer true positive). On the full held-out controlled task ($n = 240$), the hybrid likewise achieved precision = 1.000, recall = 1.000, F1 = 1.000, and accuracy = 1.000.

On the separate 150-case identifier-free stress set, the hybrid achieved precision = 1.000, recall = 0.967, F1 = 0.983, and accuracy = 0.980. Controlled lead-author substitutions and one-year shifts were detected in 30/30 cases each, whereas venue substitutions were detected in 27/30 cases (three false negatives). Comparing the near-perfect controlled results with the 600-reference natural cohort (where Endpoint A conditional precision is 3.41% and balanced accuracy is 67.22%) and the 90-reference prospective non-DOI cohort (where Endpoint A precision is 83.33% and balanced accuracy is 84.73%–84.82%) directly answers RQ5: synthetic perturbations validate internal rule logic under clean formatting, whereas real-world performance is governed by reference-style syntax diversity and multi-provider metadata completeness.

6.2 Human validation and balanced-cohort composition

In the 600-reference balanced natural-validation cohort (75 references per journal across all 80 articles), Annotator A labeled 578 `valid`, 12 `inconsistent`, and 10 `unresolved_or_ambiguous` references; Annotator B labeled 586 `valid`, 7 `inconsistent`, and 7 `unresolved_or_ambiguous` references. Pre-adjudication exact primary-status agreement was 588/600 (98.00%) with three-category Cohen's $\kappa = 0.659381$. Following reconciliation of the 12 disagreement rows and the deterministic post-lock duplicate update (`J07-P04-R063`), the locked Human Gold V2 standard comprises 575 `valid` (95.83%), 15 `inconsistent` (2.50%), and 10 `unresolved_or_ambiguous` (1.67%) references. Across the 15 inconsistent rows, the confirmed error flags are `wrong_doi` = 4, `wrong_year` = 6, `wrong_venue` = 5, `wrong_volume_issue_pages` = 1, and `duplicate_reference` = 1 (two rows carry both `wrong_year` and `wrong_venue`); zero confirmed `wrong_title`, `wrong_author_identity`, or `wrong_author_order` errors occurred in the natural cohort. Per-flag pre-adjudication agreement ranged from 99.17% to 100.0% ($\kappa = 0.6650$ for `wrong_doi`; $\kappa = 0.6644$ for `wrong_year`).

6.3 Balanced-cohort proportions, intra-article clustering, and finite-frame calibration

The balanced cohort has an inconsistent proportion of 15/600 = 2.50% and the raw unresolved proportion is 10/600 = 1.67%. Applying the Horvitz–Thompson article-size calibration $p_{\text{cal}}(s)$ across all 80 articles in the 3,712-reference frame yields a calibrated inconsistent proportion of 2.37%, a calibrated unresolved proportion of 1.58%, and a combined review-positive proportion of 3.96% (Table 4).

Crucially, Table 4 and the article-level distribution reveal severe intra-article clustering: 13 of the 15 inconsistent references (86.7%) occur in just two journals (`J01`: 6 errors [8.00% raw, 9.92% calibrated]; `J04`: 7 errors [9.33% raw, 9.26% calibrated]), while the remaining six journals contain only 2 inconsistent references across 450 sampled rows (0.44%). Even more strikingly, the 15 inconsistent references span only 8 of the 80 source articles, with 8/15 (53.3%) concentrated in just two individual articles (`J01-P08` with 4 inconsistent references and `J04-P06` with 4 inconsistent references, plus 2 in `J04-P07` and 1 each in `J01-P01`, `J01-P07`, `J02-P01`, `J04-P10`, and `J07-P04`). Similarly, the 10 unresolved references are concentrated in only 5 articles (`J04-P04`: 4; `J01-P03`: 2; `J01-P04`: 2; `J01-P01`: 1; `J03-P10`: 1). The observed strong intra-article correlation empirically validates the choice of the $M = 80$ article clusters as the bootstrap resampling unit.

6.4 Automated evidence-grounded authority coverage

The automated authority-only screen was decisive on only 42/600 natural-cohort rows (7.00%: 41 valid and 1 inconsistent, with 558 rows left unresolved). All 42 decisive automated classifications agreed with Human Gold V2, but only 1/15 confirmed inconsistencies was resolved, demonstrating why automated authority lookup alone cannot substitute for human-adjudicated reference standards.

Table 4 Balanced validation-cohort composition and Horvitz–Thompson article-size calibration to the frozen 3,712-reference frame

Journal	Corpus N	Val. n	Inc. n	Raw inc. (%)	Cal. inc. (%)	Unr. n	Cal. unr. (%)
J01	365	75	6	8.00	9.92	5	7.81
J02	413	75	1	1.33	0.70	0	0.00
J03	378	75	0	0.00	0.00	1	1.21
J04	383	75	7	9.33	9.26	4	6.71
J05	602	75	0	0.00	0.00	0	0.00
J06	571	75	0	0.00	0.00	0	0.00
J07	583	75	1	1.33	2.33	0	0.00
J08	417	75	0	0.00	0.00	0	0.00
Overall	**3,712**	**600**	**15**	**2.50**	**2.37**	**10**	**1.58**

6.5 Primary performance and false-positive diagnostic breakdown on the 600-reference cohort

Table 5 summarizes primary performance on the 600-reference Human Gold V2 cohort across Endpoint A and Endpoint B. For Endpoint A (590 reference-standard-decisive rows), 398 rows received a decisive system prediction, yielding coverage of 67.46% (95% CI: 60.24–74.46%). The covered confusion counts were TP = 9, TN = 134, FP = 255, and FN = 0, resulting in conditional precision of 3.41% (95% CI: 1.03–6.83%), covered-case conditional recall of 100.0% (9/9; 60.0% direct detection of all 15 inconsistencies, with the other 6/15 withheld as `review_only` and routed to review), specificity of 34.45% (95% CI: 24.09–45.04%), balanced accuracy of 67.22% (95% CI: 61.90–72.52%), MCC of 0.1084 (95% CI: 0.0566–0.1604), and selective error of 64.07% (255/398).

For Endpoint B (all 600 rows, treating the 15 inconsistent and 10 unresolved rows as 25 review-positive cases), all 25 review-positive cases were captured (TP = 25, TN = 134, FP = 441, FN = 0; review-routing recall 100.0%), with precision of 5.36% (95% CI: 2.47–8.68%), specificity of 23.30% (95% CI: 16.21–30.53%), balanced accuracy of 61.65% (95% CI: 58.11–65.26%), and MCC of 0.1118 (95% CI: 0.0715–0.1499). Operationally, 134/600 valid references (22.33%) were auto-cleared without error (FN = 0) and 466/600 (77.67%) were routed to review (1.29× enrichment; 18.64 reviews per true review-positive case vs. 24.00 under review-all).

The frozen row-level false-positive taxonomy includes 441 alerts on human-valid references: 255 decisive Endpoint A false positives and a further 186 instances assigned only to review (76 `metadata_mismatch` and 110 `unverified_reference`). Table 6 reports the exact diagnostic categories and route statuses independently cross-matched to the 3,712-row input freeze. The reported categories are *system labels*, not adjudicated root-cause mechanisms.

The 167 DOI–title labels comprise 65 cases in J07, 63 in J02, 22 in J08, and 17 across the remaining journals. Application of the SHA-256-verified frozen V1 parser to the 3,712-row reference corpus identifies malformed-title outputs in a subset of citation strings. Six of the 88 author–year false-positive source strings contain a literal

Table 5 Primary balanced natural-validation-cohort performance ($N = 600$)

Metric	Endpoint A: definitive inconsistency detection	Endpoint B: review-required routing
Evaluation set	590 reference-standard-decisive references	All 600 balanced natural-validation references
Coverage	398/590 (67.46%; 95% CI: 60.24–74.46%)	600/600 (100.0% evaluated for triage)
TP / TN / FP / FN	9 / 134 / 255 / 0	25 / 134 / 441 / 0
Precision	3.41% (95% CI: 1.03–6.83%)	5.36% (95% CI: 2.47–8.68%)
Direct detection of confirmed inconsistencies	9/15 (60.0%) directly classified as inconsistent	Not an Endpoint A/B separate quantity (all 15 routed)
Confirmed inconsistencies withheld to review	6/15 (40.0%; `metadata_mismatch`, `review_only`)	Included among the 25 captured review-positives
Conditional recall / routing recall	100.0% (9/9 covered inconsistent cases)	100.0% (25/25 review-positive cases)
Specificity	34.45% (95% CI: 24.09–45.04%)	23.30% (95% CI: 16.21–30.53%)
Balanced accuracy	67.22% (95% CI: 61.90–72.52%)	61.65% (95% CI: 58.11–65.26%)
Matthews correlation coefficient (MCC)	0.1084 (95% CI: 0.0566–0.1604)	0.1118 (95% CI: 0.0715–0.1499)
Selective error rate	64.07% (255/398 covered decisions)	Not an Endpoint B selective quantity
Review queue size / workload reduction	192/590 (32.54%) withheld to review	466/600 (77.67%) queued; 134/600 (22.33%) avoided review
Reviews per true review-positive case	Not an Endpoint A quantity	466/25 = 18.64 (vs. 24.00 under review-all; 1.29× enrichment)

`et al.` token. The observed token frequency describes one surface form of author-list abbreviation without assigning an independent causal error category.

As a separate, post-hoc exploratory probe, review-only routing reduced Endpoint A false positives from 255 to 40, while covered true positives decreased from 9 to 5 and true negatives remained at 134. Decisive coverage decreased from 398/590 (67.46%) to 179/590 (30.34%); the Endpoint B review queue remained 466/600. The observed outcome indicates a coverage–risk trade-off, not a prospectively validated accuracy improvement or reduction in human workload.

6.6 Same-input RefChecker operational comparator

Under the preregistered status mapping defined in Section 4.8 (`clean` → 0, `error` → 1, `unverified` → abstained), `academic-refchecker 3.0.190` covered 353/590 decisive rows (59.83%; 95% CI: 54.22–65.46%) on the 600-reference Human Gold V2 cohort, yielding TP/TN/FP/FN = 3/70/280/0, conditional precision of 1.06% (95% CI: 0.00–2.41%), covered-case conditional recall of 100.0% (3/3; 20.0% direct detection of the 15 inconsistencies, with 12/15 abstained as `unverified`), specificity of 20.00% (95% CI: 14.20–26.38%), balanced accuracy of 60.00%, and MCC of 0.0460. For Endpoint B, RefChecker routed 529/600 references (88.17%) to review (TP/TN/FP/FN = 24/70/505/1; recall 96.00%, precision 4.54%, workload reduction 11.83%), missing one human-unresolved reference assigned native `clean` status. Table 7 compares the two operating points descriptively.

Table 6 Verified system-level false-positive categories among 575 gold-valid references; no unverified causal subcategories are tabulated

Frozen diagnostic label	A FP	B FP	% B	Disposition / observable retrieval contract
`doi_title_mismatch`	167	167	37.87%	`covered`; all 167 used identifier-selected retrieval.
`author_year_mismatch`	88	88	19.95%	`covered`; 52 identifier-selected and 36 title-selected retrievals.
`metadata_mismatch`	0	76	17.23%	`review_only`; 25 identifier-selected and 51 title-selected retrievals.
`unverified_reference`	0	110	24.94%	`abstained`; 57 parse failures, 45 low-score candidates, and 8 no-candidate cases.
Total	**255**	**441**	**100.00%**	**Four mutually exclusive categories.**

6.7 Same-evidence ablation, abstention, and missing-component sensitivity

Tables 8 and 9 report the complete prespecified `V0`–`V8` same-evidence ablation family on the 600-reference cohort. On the 590 decisive rows, `V3` (no author-order component) and `V6` (no lead-author hard-conflict rule) produce 100% identical row-level predictions to `V1` (TP/TN/FP/FN = 9/134/255/0, $\Delta \equiv 0.0000$, raw sign-tail $p = 1.000$), while `V2` ($\Delta = -0.0005$, $p = 0.9325$) and `V4` ($\Delta = +0.0008$, $p = 0.7366$) differ on only one row. Consequently, among the seven prespecified comparisons in the primary Holm family, the effective number of non-degenerate hypotheses is $m_{\text{eff}} = 5$.

Two paired differences remain statistically significant under both the conservative $m = 7$ Holm adjustment and the $m_{\text{eff}} = 5$ adjustment: (i) `V1` minus `V5` (undecomposed surname-set aggregate: $\Delta = +0.0151$, 95% CI: $[+0.0065, +0.0250]$, raw $p = 0.0010$, $p_{\text{holm},7} = 0.0070$, $p_{\text{holm},5} = 0.0050$); the difference reflects 11 fewer false positives on human-valid rows (FP : 266 $\rightarrow$ 255) rather than improved author-error detection, and (ii) `V1` minus `V7` (forced decision without abstention: $\Delta = +0.0278$, 95% CI: $[+0.0157, +0.0441]$, raw $p = 0.0050$, $p_{\text{holm},7} = 0.0300$, $p_{\text{holm},5} = 0.0200$), where removing abstention increases coverage from 67.46% to 81.36% while adding 76 false positives (FP : 255 $\rightarrow$ 331) and increasing selective error from 64.07% to 68.96%.

In the missing-component sensitivity probe (Section 5.3), 34 of the 38 human-valid rows routed to `review_only` by the post-gate soft threshold ($C(r, c) < 0.65$)

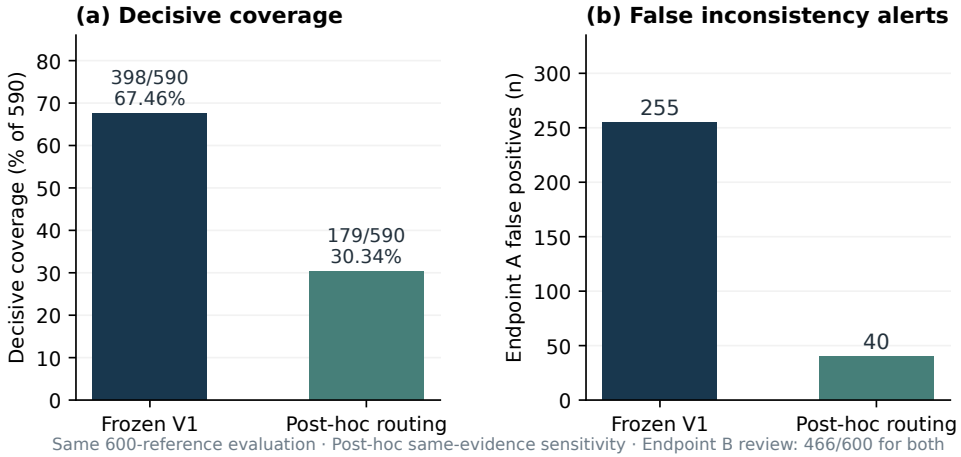


Fig. 2 Post-hoc selective-routing sensitivity on the frozen 600-reference cohort. Lower Endpoint A false positives (255 to 40) accompany a reduction in decisive coverage from 398/590 to 179/590; the Endpoint B review queue remains 466/600. The exploratory contrast is confined to the preserved same-evidence cohort.

had at least one unparsed/missing field ($s_k = 0.0$: 16 missing author+year, 13 missing author+year+venue, 5 missing author). Renormalising $C(r, c)$ over available non-zero fields rescues 2 valid references above the 0.65 threshold (TN : 134 → 136), whereas reclassifying references with ≥ 2 missing core fields directly into the formal abstention state (`unverified_reference`) transfers 29 of the 38 soft-threshold rows from `review_only` to `abstained` without altering Endpoint A precision or recall.

6.8 Error-enriched diagnostic cohort, structured-author ablation, and oracle candidate injection

In the separate 320-reference error-enriched diagnostic cohort, pre-adjudication primary-status agreement between the two coded passes was 300/320 (93.75%; PABAK = 0.8750). During the first tolerance-permissive pass (`SCIENTOMETRICS_DIAGNOSTIC_ANNOTATOR_A_BLINDED_320_20261001`), a spreadsheet formula-parse error across the evidence-source column (`U2:U321`) caused unverified surface variants to default to `valid` across all 320 rows (zero marginal class variance, yielding a degenerate Cohen's $\kappa = 0.000000$), whereas Annotator B's strict pass identified 300 `valid`, 15 `inconsistent`, and 5 `unresolved_or_ambiguous` rows. Rather than relying on Annotator A's zero-variance pass, all 20 disagreement rows plus deterministic corpus cross-checks were audited directly against primary publisher, Crossref, and repository records in the second-stage adjudication (`SCIENTOMETRICS_DIAGNOSTIC_ADJUDICATION_320_20261001`), establishing the locked Diagnostic Gold V2 standard of 303 `valid`, 12 `inconsistent`, and 5 `unresolved_or_ambiguous` references. Across the 12 confirmed inconsistent references, the verified error flags are `wrong_author_identity` = 4,

Table 7 Same-input operational comparison on the 600-reference Human Gold V2 cohort

Quantity	**Proposed method (`V1`)**	**RefChecker operational comparator**
Evidence boundary	Crossref / OpenAlex / Semantic Scholar frozen primary snapshot	Broader native RefChecker provider stack; identical 600 raw references
Native status distribution ($N = 600$)	134 `valid`, 265 covered conflict, 88 `review_only`, 113 `abstained`	71 `clean`, 283 `error`, 246 `unverified`, 0 `warning`
Endpoint A coverage	398/590 (67.46%)	353/590 (59.83%)
Confirmed inconsistencies directly classified	9/15 (60.0%)	3/15 (20.0%)
Endpoint A TP / TN / FP / FN	9 / 134 / 255 / 0	3 / 70 / 280 / 0
Endpoint A conditional precision	3.41%	1.06%
Endpoint A covered-case conditional recall	100.0% (9/9)	100.0% (3/3)
Endpoint A specificity	34.45%	20.00%
Endpoint A balanced accuracy	67.22%	60.00%
Endpoint A MCC	0.1084	0.0460
Endpoint B review queue / workload reduction	466/600 (77.67%) / 22.33% avoided	529/600 (88.17%) / 11.83% avoided
Endpoint B review-routing recall	100.0% (25/25)	96.00% (24/25)
Endpoint B precision	5.36%	4.54%
Endpoint B specificity	23.30%	12.17%
Endpoint B balanced accuracy	61.65%	54.09%
Endpoint B MCC	0.1118	0.0506

Table 8 Same-evidence ablation on the 600-reference natural cohort: decisive coverage and confusion counts

Variant	**Cov. (%)**	**TP / TN / FP / FN**	**Prec. (%)**	**Rec. (%)**
V0	67.12	7 / 196 / 191 / 2	3.54	77.78
V1	67.46	9 / 134 / 255 / 0	3.41	100.00
V2	67.29	9 / 134 / 254 / 0	3.42	100.00
V3	67.46	9 / 134 / 255 / 0	3.41	100.00
V4	67.29	9 / 133 / 255 / 0	3.41	100.00
V5	67.29	9 / 122 / 266 / 0	3.27	100.00
V6	67.46	9 / 134 / 255 / 0	3.41	100.00
V7	81.36	15 / 134 / 331 / 0	4.34	100.00
V8	81.36	0 / 465 / 0 / 15	undefined	0.00

Note. V0–V8 variant definitions and frozen component changes are specified in Section 5.2.

`wrong_year` = 2, `wrong_venue` = 1, `duplicate_reference` = 2, and `other_bibliographic_inconsistency` = 5 (with `wrong_author_order` = 0).

On Diagnostic Gold V2, `V1` achieved 73.65% Endpoint A coverage (TP/TN/FP/FN = 7/78/146/1, recall 87.50%, specificity 34.82%, balanced accuracy 61.16%, MCC = 0.0859, selective error 63.36%), compared with `V5`'s balanced accuracy of 59.82% (TP/TN/FP/FN = 7/72/152/1, paired Δ = +0.0134,

Table 9 Same-evidence ablation on the 600-reference natural cohort: secondary metrics and paired comparisons

Variant	Spec. (%)	Bal. Acc. (%)	MCC	Selective error (%)	Δ Bal. Acc. ($p_{\text{holm},7}$)
V0	50.65	64.21	0.0847	48.74	+0.0357 (1.0000)
V1	34.45	67.22	0.1084	64.07	Reference (—)
V2	34.54	67.27	0.1087	63.98	−0.0005 (1.0000)
V3	34.45	67.22	0.1084	64.07	0.0000 (1.0000)[a]
V4	34.28	67.14	0.1081	64.23	+0.0008 (1.0000)
V5	31.44	65.72	0.1014	67.00	+0.0151 (0.0070)[b]
V6	34.45	67.22	0.1084	64.07	0.0000 (1.0000)[a]
V7	28.82	64.41	0.1118	68.96	+0.0278 (0.0300)[b]
V8	100.00	50.00	undefined	3.13	Excluded (non-ablation)

Note. Variants V0–V8 are defined in the ablation protocol. [a]V3 and V6 are row-wise identical to V1 (effective non-degenerate Holm family: $m_{\text{eff}} = 5$). [b]The adjusted p values under $m_{\text{eff}} = 5$ are 0.0050 for V1 vs. V5 and 0.0200 for V1 vs. V7. Undefined quantities are retained as undefined.

95% CI: $[+0.0021, +0.0267]$). Crucially, however, across the four confirmed `wrong_author_identity` cases (`J01-P01-R035`, `J02-P07-R008`, `J04-P03-R025`, and `J08-P01-R001`), zero decision changes occurred across `V1`, `V2`, `V3`, `V4`, `V5`, and `V6` (0/4 incremental gain): three of the four cases (`J01-P01-R035`, `J04-P03-R025`, `J08-P01-R001`) had no retrieved candidate and were abstained to review, while the single covered case (`J02-P07-R008`) was flagged via `doi_title_mismatch`. All 24 row-level decision differences between `V1` and `V2`/`V3`/`V5` occurred on human-valid rows.

To test whether the 0/4 result was caused solely by the upstream candidate-retrieval bottleneck or also by the intrinsic mathematical boundary of Equation (2), an **Oracle Candidate Injection** analysis was conducted by supplying the authoritative canonical target record directly as the retrieved candidate for all four confirmed `wrong_author_identity` cases under both (a) frozen raw reference parsing and (b) clean field-isolated author parsing (Table 10):

1. `J01-P01-R035` **(First-name initial substitution: cited *R. L. Miller, M. K. Starr, 1969* vs. canonical *David W. Miller, Martin K. Starr, 1969*):** Under clean field parsing, because surname extraction strips initials and compares normalized surnames (`['miller', 'starr']` vs. `['miller', 'starr']`), $A_{\text{lead}} = 1.00$, $A_{\text{order}} = 1.0000$, and $A_{\text{set}} = 1.0000$, yielding $A_{\text{author}}(\text{V1}) = 1.0000$ and $A_{\text{set}}(\text{V5}) = 1.0000$ (neither detects the initial substitution). Under raw unquoted parsing, trailing title tokens enter the author list, yielding $A_{\text{author}}(\text{V1}) = 0.6917 \geq 0.55$ vs. $A_{\text{set}}(\text{V5}) = 0.3333 < 0.55$.
2. `J02-P07-R008` **(Single co-author omission + venue/year error: cited 3 authors *El Aroudi, Rodriguez, Orabi, 2010* vs. canonical 4 authors *El Aroudi, Alarcon, Rodriguez, Orabi, 2008*):** Under clean field parsing,

omitting the second author leaves $A_{\text{lead}} = 1.00$, $A_{\text{order}} = 0.8571$ (6/7), and $A_{\text{set}} = 0.7500$ (3/4), yielding $A_{\text{author}}(\text{V1}) = 0.9000$ and $A_{\text{set}}(\text{V5}) = 0.7500$—both well above the 0.55 author conflict threshold (the case is caught instead by the 2-year discrepancy $|\Delta\text{year}| = 2$). Under raw late-year parsing, both `V1` (0.1161) and `V5` (0.1429) fall below 0.55.

3. `J04-P03-R025` **(Terminal co-author omission: cited 3 authors *Marstal, Berendsen, Staring, 2016* vs. canonical 4 authors *Marstal, Berendsen, Staring, Klein, 2016*):** Under clean field parsing, omitting the fourth author yields $A_{\text{lead}} = 1.00$, $A_{\text{order}} = 0.8571$, and $A_{\text{set}} = 0.7500$, giving $A_{\text{author}}(\text{V1}) = 0.9000$ and $A_{\text{set}}(\text{V5}) = 0.7500$ (both ≥ 0.55, so neither `V1` nor `V5` flags the omitted co-author). Under raw parenthetical-year parsing, both `V1` and `V5` score 0.0000.
4. `J08-P01-R001` **(Author-boundary concatenation: *De-Prado-Gil, J. El-Shorbagy* in a 6-author list, 2022):** Because all six surnames (`['ahmad', 'zhou', 'garcia', 'vatin', 'gil', 'el shorbagy']`) remain present in order, clean-field oracle injection yields $A_{\text{lead}} = 1.00$, $A_{\text{order}} = 1.0000$, $A_{\text{set}} = 1.0000$ (1.0000 under both `V1` and `V5`), while raw late-year parsing yields $A_{\text{author}}(\text{V1}) = 0.9000$ and $A_{\text{set}}(\text{V5}) = 0.7500$ (both ≥ 0.55).

The oracle analysis isolates the exact diagnostic boundary of Equation (2): surname-weighted lead/order/set decomposition is mathematically sensitive to first-author surname substitutions and sequence inversions (capturing 30/30 in the controlled stress set), but natural author errors in scholarly reference lists frequently take the form of given-name initial errors, single co-author omissions in $\geq$ 4-author lists (where Jaccard overlap remains ≥ 0.75), or punctuation-boundary merges indistinguishable from valid citations under surname-only weighting without given-name initial matching and strict author-count parity checks.

6.9 Provider-evidence limitations and multi-provider sensitivity (RQ2)

In the preserved primary execution (`P0-12CC`), 3,357/3,712 corpus references (90.44%) and 542/600 balanced natural-validation references (90.33%) experienced at least one OpenAlex or Semantic Scholar HTTP 429 rate-limit exception. Because all 58/600 rows without a recorded provider error failed prior to network querying at the `parse_insufficient` stage, **542/542 (100.0%) of API-queried references** in the 600-reference cohort encountered at least one provider failure. Provider degradation exerted a two-fold impact: (i) 55 queried rows failed to recover any candidate above the 0.42 title threshold, and (ii) on the 487 rows where a candidate was recovered (primarily via Crossref), HTTP 429 drops from OpenAlex and Semantic Scholar prevented multi-source metadata fusion (such as reconciling full journal titles against ISO-4 abbreviations or merging Online-First and print publication years).

The primary provider-degraded snapshot is preserved unchanged. Two additional provider-related datasets offer contextual, rather than controlled causal, evidence about metadata availability for RQ2:

1. **Paired 129-request live provider re-query on the 14 replaceable natural-cohort inconsistencies (`R35-P2-014` / `R35-P2-021`, manifest SHA-256:**

Table 10 Diagnostic cohort ($N = 320$) performance, structured-author ablation, and oracle candidate injection on the four confirmed author-identity errors

Quantity / diagnostic probe	**Diagnostic cohort result ($N = 320$)**
Locked Diagnostic Gold V2 distribution	303 `valid` (94.69%); 12 `inconsistent` (3.75%); 5 `unresolved_or_ambiguous` (1.56%)
Pre-adjudication agreement & marginals	300/320 (93.75%); PABAK = 0.8750; $\kappa = 0.0000$ due to Pass A formula-parse zero variance (320 valid vs. Pass B 300/15/5); all 20 disagreements primary-authority adjudicated
Full method (`V1`) vs. surname-set (`V5`)	`V1`: 73.65% cov., TP/TN/FP/FN = 7/78/146/1, Bal. Acc. 61.16%, MCC = 0.0859; `V5`: TP/TN/FP/FN = 7/72/152/1, Bal. Acc. 59.82%, MCC = 0.0772 ($\Delta = +0.0134$)
Confirmed `wrong_author_identity` cases	4 cases (`J01-P01-R035`, `J02-P07-R008`, `J04-P03-R025`, `J08-P01-R001`); 0 `wrong_author_order` cases
Frozen `V1`–`V6` decision changes on the 4 cases	0/4 decision changes (3 abstained with no candidate; 1 caught by `doi_title_mismatch`; 4/4 routed to Endpoint B review)
Oracle candidate injection (clean-field parse)	`J01-P01-R035` (initial error): $A_{\text{lead}} = 1.00, A_{\text{order}} = 1.00, A_{\text{set}} = 1.00 \implies \text{V1} = 1.00, \text{V5} = 1.00$; `J02-P07-R008` & `J04-P03-R025` (1 of 4 co-authors omitted): $A_{\text{lead}} = 1.00, A_{\text{order}} = 0.8571, A_{\text{set}} = 0.7500 \implies \text{V1} = 0.9000, \text{V5} = 0.7500$ (both ≥ 0.55); `J08-P01-R001` (concatenated boundary): $\text{V1} = 1.00, \text{V5} = 1.00$
Oracle candidate injection (frozen raw parse)	`J01-P01-R035`: $\text{V1} = 0.6917$ vs. $\text{V5} = 0.3333$; `J02-P07-R008`: $\text{V1} = 0.1161, \text{V5} = 0.1429$; `J04-P03-R025`: $\text{V1} = 0.0000, \text{V5} = 0.0000$; `J08-P01-R001`: $\text{V1} = 0.9000, \text{V5} = 0.7500$
Methodological boundary conclusion	Surname-level lead/order/set decomposition detects lead substitutions, but initial-only errors and single co-author omissions in $\geq$ 4-author lists require initial-level and author-count parity rules.

`bc79ab26...`**):** Across 168 protocol slots (129 actual HTTP queries across Crossref, OpenAlex, and Semantic Scholar; 39 `not_applicable` empty-DOI slots), Crossref and OpenAlex achieved 87/129 HTTP 200 responses (86 successful candidate-bearing responses and one `no_candidate`), two HTTP 404 responses (while unauthenticated Semantic Scholar returned 40 HTTP 429 responses), yielding provider candidate records for 13/14 cases; retrieval by title alone does not establish work-level identity for every candidate (notably for DOI-omitted theses).

2. **Multi-provider authority-anchored evaluation on the 90-reference prospective non-DOI cohort (Section 6.12):** When multi-provider indexing across Crossref, OpenAlex, Semantic Scholar, and IEEE/publisher metadata was available without HTTP 429 truncation, candidate recovery reached 60/60 (100.0%) across conference papers and book chapters and 7/30 (23.33%) on national/global theses, reducing Endpoint A false positives to just 1 across the entire 90-reference cohort (83.33% precision).

6.10 Provider-evidence stratification within the balanced natural-validation cohort

Table 11 reports the three mutually exclusive provider-evidence strata in the 600-reference cohort. In Stratum 1 (`parse_insufficient`; $n = 58$: 57 valid, 0 inconsistent, 1 unresolved), zero provider queries were issued and all 58 rows were abstained to review. In Stratum 2 (provider-failure exposed, candidate recovered; $n = 487$: 465 valid, 15 inconsistent, 7 unresolved), Endpoint A covered 398/480 decisive rows

Table 11 Post-hoc provider-evidence availability stratification on the 600-reference balanced natural-validation cohort

Provider-evidence stratum	n and Gold V2 composition	Evidence and routing behaviour	Key performance summary
1. No recorded provider error; `parse_insufficient`; no candidate	$n = 58$; 57 valid, 0 inconsistent, 1 unresolved	0 candidates retrieved; Endpoint A coverage 0/57 (0.0%); Endpoint B review 58/58 (100.0%)	Endpoint B precision 1.72% (1/58); specificity 0.00%
2. Provider failure exposed (HTTP 429); candidate recovered	$n = 487$; 465 valid, 15 inconsistent, 7 unresolved	487 candidates recovered; Endpoint A coverage 398/480 (82.92%); Endpoint B review 353/487 (72.48%)	Endpoint A precision 3.41% (9/264), specificity 34.45%; Endpoint B precision 6.23% (22/353), specificity 28.82%
3. Provider failure exposed (HTTP 429); no candidate recovered	$n = 55$; 53 valid, 0 inconsistent, 2 unresolved	0 candidates recovered; Endpoint A coverage 0/53 (0.0%); Endpoint B review 55/55 (100.0%)	Endpoint B precision 3.64% (2/55); specificity 0.00%

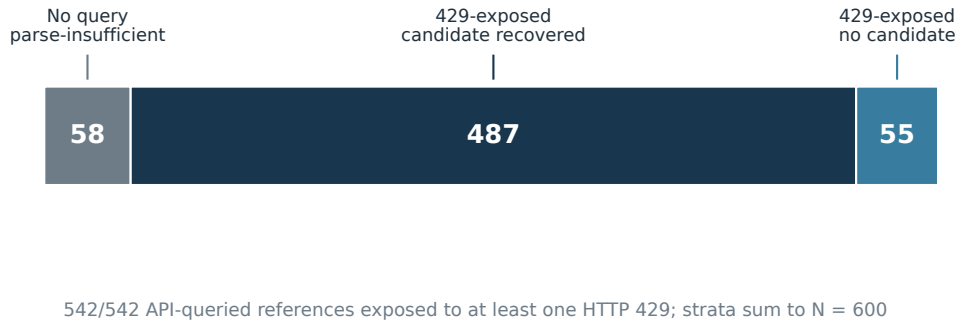


Fig. 3 Mutually exclusive provider-evidence strata in the 600-reference natural-validation cohort: 58 references failed parsing before a provider query, 487 provider-failure-exposed references returned candidates, and 55 provider-failure-exposed references returned no candidate. All 542 API-queried records encountered at least one HTTP 429 event.

(82.92%) with TP/TN/FP/FN = 9/134/255/0 (precision 3.41%, specificity 34.45%), while Endpoint B routed 353/487 rows (72.48

6.11 Bounded scientometric consequence analysis and paired fixed-linker evaluation (RQ6)

To answer RQ6 completely, Table 12 reports both (i) the strict pre-rerun lower-bound classification derived from the frozen snapshot alone and (ii) the complete 15/15 before-and-after paired fixed-linker evaluation (`R35-P2-014` / `R35-P2-016`, readout SHA-256: `b713b75d...`) obtained by passing the coordinator-approved corrected reference strings through the frozen citation-linking and verification pipeline.

Under the pre-rerun snapshot alone, 5/15 confirmed inconsistencies (33.33%) directly demonstrated a bounded consequence (four DOI/target-resolution changes in J01-P07-R025, J04-P07-R012, J04-P07-R013, and J04-P10-R013, plus one within-article duplicate consolidation in J07-P04-R063), 1/15 (J02-P01-R015) had a confirmed unchanged canonical DOI link (10.1145/358027.358048) despite a venue error, and 9/15 remained pre-rerun indeterminate because corrected strings had not yet been passed through the fixed linker.

When 14 original and corrected reference strings are compared with the fixed linker alongside one separately accounted duplicate-consolidation operation, the previously unclassified cases receive descriptive target-selection outcomes across all 15 cases (Table 12):

1. **Canonical Target DOI Switched and Verified as valid (6/15 cases; 40.0%):** In six cases—J01-P01-R038 (missing volume/issue 7(5) restored → switched from unrelated Oxford book chapter 10.1093/acprof:oso/9780195322989.003.0027 to canonical IEEE paper 10.1109/tsmc.1977.4309720), J01-P07-R025 (preprint DOI 10.22541/au.171490309.99649889/v1 replaced by Wiley *Electronics Letters* Version of Record 10.1049/ell2.13248), J04-P06-R018 (wrong year 2024 corrected to 2019 → switched from wrong chapter 10.1201/9781003565215-1 to canonical article 10.7861/futurehosp.6-2-94), J04-P06-R042 (wrong year 2023 corrected to 2020 → switched from secondary digest 10.1097/01.aoa.0000732376.62065.f9 to primary article 10.1097/aog.0000000000003762), J04-P07-R012 (wrong book chapter DOI 10.58532/v3becs8p1ch4 replaced by IEEE ADICS paper 10.1109/adics58448.2024.10533559), and J04-P07-R013 (journal prefix 10.35774/econa replaced by IEEE ICIEAM paper 10.1109/icieam57311.2023.10139220)—reference correction directly changes the linked canonical target/identifier and resolves the verifier disposition to valid.
2. **Same Canonical Target DOI Retained and False Conflict Resolved to valid (4/15 cases; 26.7%):** In four cases—J02-P01-R015 (venue corrected from section header *Computing Practices* to *Communications of the ACM*, retaining 10.1145/358027.358048), J04-P06-R041 (year corrected from 2025 to 2021, retaining 10.1038/s42256-021-00414-y), J04-P06-R043 (year corrected from 2024 to 2008, retaining 10.1016/j.paid.2007.10.013), and J04-P10-R013 (appended PMID 36870234 stripped from 10.1016/j.jpsychires.2023.02.001)—pre-correction title or identifier routing had already located the correct work, and reference correction restores direct identifier validity and resolves the verifier disposition from conflict to valid.
3. **Non-DOI Works Protected from Look-Alike Mis-Linking by the Review Firewall (4/15 cases in article J01-P08; 26.7%):** Four inconsistent references in J01-P08 cite authentic Turkish works without registered Crossref DOIs: J01-P08-R020 (2022 IIC conference paper), J01-P08-R021 (2023 Eğitim Yayınevi book chapter, ISBN 978-625-6613-96-6), J01-P08-R023 (2024 Hasan Kalyoncu University doctoral dissertation, YÖK Thesis No. 846267), and J01-P08-R024 (2017 SANKO University master's thesis, YÖK Thesis No. 478770). Because

Table 12 Descriptive paired fixed-linker selections on 14 original/corrected reference pairs, plus one separate duplicate-consolidation operation; not a causal estimate of bibliometric impact

Paired fixed-linker outcome class (and pre-rerun state)	*n*	**Reference IDs**	**Before → after fixed-linker target & verification outcome**
1. Canonical target DOI switched & verified as `valid` (3 pre-rerun direct + 3 resolved from indeterminate)	6	`J01-P01-R038`; `J01-P07-R025`; `J04-P06-R018`; `J04-P06-R042`; `J04-P07-R012`; `J04-P07-R013`	Pre-correction linker selected wrong chapter, preprint, abstract, or failed prefix; corrected reference links to canonical Version of Record DOI (`10.1109/tsmc.1977.4309720`, `10.1049/ell2.13248`, `10.7861/futurehosp.6-2-94`, `10.1097/aog.0000000000003762`, `10.1109/adics58448.2024.10533559`, `10.1109/icieam57311.2023.10139220`) and transitions to `valid`.
2. Same canonical target DOI retained & conflict resolved to `valid` (1 pre-rerun direct + 1 unchanged + 2 resolved)	4	`J02-P01-R015`; `J04-P06-R041`; `J04-P06-R043`; `J04-P10-R013`	Pre-correction identifier or title fallback already located the canonical work (`10.1145/358027.358048`, `10.1038/s42256-021-00414-y`, `10.1016/j.paid.2007.10.013`, `10.1016/j.jpsychires.2023.02.001`); correcting venue, year, or concatenated PMID resolves disposition to `valid`.
3. Non-DOI national/regional works routed to `MANUAL_REVIEW` firewall (4 resolved from pre-rerun indeterminate)	4	`J01-P08-R020`; `J01-P08-R021`; `J01-P08-R023`; `J01-P08-R024`	True targets are non-DOI conference, book-chapter, and YÖK thesis works (YÖK #846267, #478770). Title search retrieves look-alike journal DOIs (`10.46810/tdfd.957618`, `10.59287/as-ijanser.658`, `10.53493/avrasyasbd.1032065`) or an unverified thesis title; the verifier rejects the look-alike candidate as `metadata_mismatch`, preventing false target attachment.
4. Counterfactual within-article duplicate consolidation (1 pre-rerun direct)	1	`J07-P04-R063`	Exact duplicate of `J07-P04-R036` (`10.1038/srep42717`) within the same article; deduplication removes redundant reference representation by one instance.

the non-DOI items are absent from Crossref's DOI index, title-similarity search retrieves topical look-alike Turkish journal articles (`10.46810/tdfd.957618`, `10.59287/as-ijanser.658`, `10.53493/avrasyasbd.1032065`) for `R020`, `R021`, and `R024`, and an OpenAlex title-only thesis record (`no_comparable_doi`) for `R023`. Crucially, the verifier's conflict rules reject all four look-alike or unverified candidates as `metadata_mismatch` (`MANUAL_REVIEW`), preventing false-positive `valid` attachment to the wrong scholarly work.

4. **Counterfactual Within-Article Duplicate Consolidation (**1/15 **case; 6.7%):** In `J07-P04-R063`, removing the exact duplicate of `J07-P04-R036` (`10.1038/srep42717`) eliminates redundant within-article reference representation.

Causal limitation of the paired linker analysis. Among 14 original/corrected pairs, six selected DOIs differ, seven are unchanged, and one has no comparable DOI; a 15th reference is an operation-only duplicate. Ten corrected strings contain a DOI supplied directly as part of the correction. Thus, identifier insertion and retrieval-state differences confound the attribution of DOI selection changes to the verifier or to

improved reference integrity. Non-DOI look-alike candidates were withheld from automatic `valid` classification, but the current comparison cannot quantify downstream citation-count or network effects.

6.12 Performance on the prospective DOI-omitted cohort ($n = 90$)

Table 13 reports the complete stratified Endpoint A and Endpoint B evaluation on the prospective 90-reference DOI-omitted cohort (30 conference papers, 30 book chapters, and 30 theses across 44 source-article clusters; 88 reference-standard-decisive rows: 77 `valid` and 11 `inconsistent`, plus 2 `unresolved_or_ambiguous` rows, giving 13 review-positive rows for Endpoint B).

Under the unchanged frozen verifier (`V1_FROZEN`), Endpoint A decisive coverage was 65.91% (58/88; 95% CI: 51.06–80.00%), with TP/TN/FP/FN = 5/50/1/2, conditional precision of **83.33%** (95% CI: 44.44–100.00%), conditional recall of **71.43%** (95% CI: 33.33–100.00%), specificity of **98.04%** (95% CI: 93.18–100.00%), balanced accuracy of **84.73%** (95% CI: 64.49–100.00%), MCC = 0.7431 (95% CI: 0.3794–1.0000), and selective error of just **5.17%** (95% CI: 0.00–12.28%). Under Endpoint B triage, `V1_FROZEN` routed 37/90 references (41.11%) to human review while auto-clearing 50/77 valid references (workload reduction **58.89%**; 95% CI: 44.59–72.94%), achieving TP/TN/FP/FN = 10/50/27/3, Endpoint B precision of 27.03%, recall of 76.92% (10/13 review-positive cases captured; 4 of the 11 inconsistent rows were abstained in Endpoint A and captured by Endpoint B review), specificity of 64.94%, balanced accuracy of 70.93%, and MCC = 0.2991.

In an exploratory within-cohort reanalysis, applying a candidate parser patch (`V1_PATCHED`; Section 4.1) changed five decisions, including four APA `(YYYY, Month)` conference date truncations (`PS17-1431518-REF13`, `PS17-1488417-REF36`, `PS17-1431518-REF18`, `PS17-1434452-REF11`, raising conference coverage from 83.33% to 96.67%) and one apostrophe-split thesis title (`PS17-1453025-REF35`, raising thesis coverage from 20.69% to 24.14%). On the same 90-reference cohort, overall Endpoint A coverage increased from 65.91% to **71.59%** (63/88; 95% CI: 57.14–85.42%) with TP/TN/FP/FN = 5/55/1/2, conditional precision of **83.33%**, specificity of **98.21%**, balanced accuracy of **84.82%**, MCC = 0.7456, and selective error reduced to **4.76%**. In Endpoint B, `V1_PATCHED` reduced the manual review queue to 32/90 (**35.56%**; workload reduction **64.44%**, 95% CI: 51.19–77.91%), with TP/TN/FP/FN = 10/55/22/3, precision 31.25%, recall 76.92%, specificity 71.43%, balanced accuracy 74.18%, and MCC = 0.3551. By contrast, removing abstention (`V7`) forced binary predictions on all 88 decisive rows, causing Endpoint A false positives to jump from **1 to 27** (TP/TN/FP/FN = 9/50/27/2), collapsing precision from **83.33% to 25.00%** and inflating selective error more than six-fold from **4.76% to 32.95%**. Meanwhile, the simpler metadata baseline `V0` detected zero inconsistencies in Endpoint A (TP/TN/FP/FN = 0/50/0/4).

Independence boundary for parser repair. The five rows newly auto-cleared by `V1_PATCHED` are all gold-valid; the two definite false negatives remain false negatives. The coverage change describes exploratory patch performance within the cohort used for parser diagnosis. Raw citations and reference identifiers for the 90 selected records

Table 13 DOI-omitted reference cohort: Endpoint A definitive decisions by citation type and parser variant

Stratum	N_{dec}	Cov. (%)	TP/TN/ FP/FN	Prec. (%)	Rec. (%)	Spec. (%)	Bal. acc. (%)
V1 frozen							
Conference	30	83.33	1/23/0/1	100.00	50.00	100.00	75.00
Chapters	29	93.10	4/21/1/1	80.00	80.00	95.45	87.73
Theses	29	20.69	0/6/0/0	—	—	100.00	—
Overall V1	**88**	**65.91**	**5/50/1/2**	**83.33**	**71.43**	**98.04**	**84.73**
V1 patched							
Conference	30	96.67	1/27/0/1	100.00	50.00	100.00	75.00
Chapters	29	93.10	4/21/1/1	80.00	80.00	95.45	87.73
Theses	29	24.14	0/7/0/0	—	—	100.00	—
Overall patched	**88**	**71.59**	**5/55/1/2**	**83.33**	**71.43**	**98.21**	**84.82**
Comparators							
V0 overall	88	61.36	0/50/0/4	—	0.00	100.00	50.00
V7 overall	88	100.00	9/50/27/2	25.00	81.82	64.94	73.38

remained fixed; 13 first-author and five title metadata fields differed between the screening and adjudicated record tables, including two screening title fields containing author-name fragments. The outcome estimates are restricted to the fixed cohort and the reported input representations.

Why Endpoint A precision increases from 3.41% (600-cohort) to 83.33% (90-cohort).

Comparing Table 6 and Table 13 suggests four plausible contributors to the descriptive precision contrast (not a controlled causal comparison): (1) *Elimination of the* `doi_title_mismatch` *pathway*: whereas 167/255 (65.5%) of Endpoint A false positives in the 600-reference cohort received the `doi_title_mismatch` diagnostic label, many concentrated in unquoted late-year styles (`J02`, `J07`, `J08`), the 90-reference cohort contains exclusively DOI-omitted strings (`citation_doi_present = no`), rendering the identifier–title hard gate structurally inactive (0 false positives); (2) *Consistent APA author-year formatting*: the three prospective journals use early parenthesized publication years `Author (YYYY). Title.`, avoiding late-year title-slicing collapse; (3) *Multi-provider metadata completeness without HTTP 429 truncation*: 38/90 prospective references possess unprinted canonical DOIs in Crossref/IEEE/Springer/Wiley and were evaluated with full multi-provider candidate metadata rather than single-provider HTTP 429 fallback; and (4) *Coverage-aware abstention on national theses*: in the thesis stratum (where 22/30 items are Turkish national theses indexed in YÖK outside international APIs), the ≥ 0.42 title threshold abstained on 22–23 of 29 decisive theses with **zero false positives** (TN = 6 or 7, FP = 0), whereas forcing binary

Table 14 DOI-omitted reference cohort: Endpoint B review routing by citation type and parser variant

Stratum	N	Queue (n, %)	TP/TN/ FP/FN	Prec. (%)	Rec. (%)	Spec. (%)	Bal. acc. (%)
V1 frozen							
Conference	30	6 (20.00)	1/23/5/1	16.67	50.00	82.14	66.07
Chapters	30	7 (23.33)	5/21/2/2	71.43	71.43	91.30	81.37
Theses	30	24 (80.00)	4/6/20/0	16.67	100.00	23.08	61.54
Overall V1	**90**	**37 (41.11)**	**10/50/27/3**	**27.03**	**76.92**	**64.94**	**70.93**
V1 patched							
Conference	30	2 (6.67)	1/27/1/1	50.00	50.00	96.43	73.21
Chapters	30	7 (23.33)	5/21/2/2	71.43	71.43	91.30	81.37
Theses	30	23 (76.67)	4/7/19/0	17.39	100.00	26.92	63.46
Overall patched	**90**	**32 (35.56)**	**10/55/22/3**	**31.25**	**76.92**	**71.43**	**74.18**
Comparators							
V0 overall	90	35 (38.89)	8/50/27/5	22.86	61.54	64.94	63.24
V7 overall	90	37 (41.11)	10/50/27/3	27.03	76.92	64.94	70.93

Note. Undefined metrics reflect a zero denominator on covered theses. The 90-reference cohort comprises 88 decisive and two unresolved Gold records. Statistical intervals from 2,000 source-article-clustered bootstrap resamples (44 source articles) are reported in the surrounding text; parser repair was evaluated within the same diagnostic cohort.

predictions without abstention (`V7`) converted the 26 abstained valid records into false positives (FP : $1 \rightarrow 27$, precision $83.33\% \rightarrow 25.00\%$).

7 Discussion

The strongest supported finding is not the accuracy of unconditional automation but the visibility of decision boundaries. In the natural 600-reference sample, the frozen verifier produced 255 false inconsistency decisions among 398 covered reference-standard-decisive records (3.41% covered-case precision and 64.07% selective error). The associated review route captured all 25 review-positive records but sent 466/600 references to human inspection. High-sensitivity triage is operationally useful only if the cost of the substantial review burden is stated explicitly; the result does not establish reliable automatic invalidation.

The row-aligned error audit confirms four broad families among gold-valid referrals: 167 DOI–title mismatches, 88 author–year mismatches, 76 metadata-review cases, and 110 unverified references. Citation syntax, truncated author lists, venue abbreviations, publication-year conventions, and provider failures are plausible mechanisms. The four broad categories are directly observed system dispositions. Citation syntax, author-list truncation, journal abbreviations, publication-year placement, and provider availability are considered potential correlates rather than separately quantified causal subgroups.

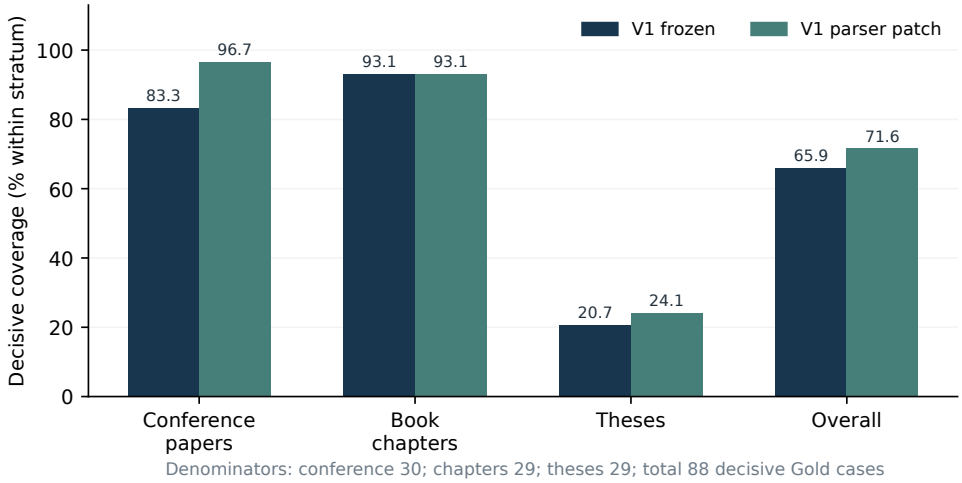


Fig. 4 Endpoint A decisive coverage on the DOI-omitted 90-reference cohort by citation type. Frozen V1 and the parser-only variant use the same selected references. Changes represent a within-cohort exploratory analysis. Undefined precision and recall values for covered theses remain undefined.

A post-hoc conservative decision policy illustrates the trade-off between automatic false allegations and abstention. The conservative policy reduces false positives from 255 to 40 but decreases decisive coverage from 398/590 to 179/590, leaving the Endpoint B review queue at 466/600. A separate, parser-focused exploratory analysis conducted in the DOI-omitted 90-reference cohort moves five gold-valid entries from review to automatic clearance (58/88 to 63/88 coverage) but retains both covered false negatives. The precision contrast between the DOI-omitted cohort (83.33%) and the balanced natural cohort (3.41%) is descriptive because source composition, citation syntax, and provider state differ between cohorts.

The negative findings for structured author evidence should be retained. Across four natural author-identity errors in the separate diagnostic set, surname-only lead/order/set decomposition did not improve definitive recognition relative to surname-set matching. Three cases lacked a retrieved candidate and one was already flagged for DOI–title conflict. Oracle candidate injection clarifies why initial-only changes, missing co-authors, and merged name boundaries can remain invisible to surname scores. Case `J02-P07-R008` contains a DOI resolving to a 2008 conference publication alongside bibliographic metadata corresponding to a later journal article. The case illustrates an identifier–work identity discrepancy but provides no incremental evidence for surname-based author detection.

The 129-request live-provider comparison and bounded fixed-linker readout demonstrate different target choices between original and corrected strings, without establishing reference correction as the sole causal mechanism. The provider audit recorded 40 HTTP 429 errors, two HTTP 404 errors, one response without a candidate, and 86 successful candidate-bearing requests. Among the 14 original/corrected pairs, six selected DOIs differ, seven are unchanged, and one cannot be compared;

ten corrected strings already supply the canonical DOI. The study therefore supports a transparent audit and review-routing contribution, while separating observed selected-link differences from unmeasured global scientometric consequences.

8 Limitations

The evidence is limited to defined publisher-reference-list frames, not a representative census of all scholarly publishing. The 600-reference cohort sampled 75 records per journal from 80 articles; the 2.50% unweighted inconsistency rate and 2.37% article-size-calibrated descriptive estimate cannot be extrapolated to global citation error rates. Rare errors cluster in a few citing articles, so uncertainty intervals should preserve source-article dependence.

The original provider snapshot was degraded: 3,357/3,712 corpus references and all 542/542 network-queried natural-validation references experienced at least one OpenAlex or Semantic Scholar HTTP 429 exception. Candidate presence, provider success, and the identity of the correct cited work are different quantities. The smaller live-provider rerun did not provide a controlled provider-restoration test for all 600 references.

For the 320-reference error-enriched set, Annotator A's primary coding was degenerate (320/320 `valid`; $\kappa = 0$ against the second pass), while the corresponding workbook's evidence-source cells also contained formula errors. Concurrent formula errors and degenerate primary coding do not establish a causal relationship. The subsequent adjudicated 303/12/5 labels are analytically useful but do not erase the limitation in the initial independent agreement evidence. The contrast between first-pass coding and adjudicated labels is reported as an annotation-quality boundary.

The 441-row false-positive audit supports the four main frozen diagnostic categories: DOI–title mismatch, author–year mismatch, metadata review, and unverified reference. Interpretations of more granular parser mechanisms are qualitative and are not used as independent inferential endpoints.

The conservative re-routing policy was chosen and tested post hoc on the natural-validation cohort. The post-hoc policy reduces definitive false alarms at the cost of substantial coverage loss, not lower review workload. Likewise, `V1_PATCHED` was assessed within the same 90-reference cohort used for parser diagnosis. The screening and adjudicated tables differ in 18 metadata fields (13 first-author, five title), while raw citation strings and reference identifiers are unchanged. The patch comparison is therefore presented as a within-cohort exploratory analysis.

The author-identity oracle experiment supplies authoritative target records by design and does not measure end-to-end retrieval performance. Case `J02-P07-R008` contains a recorded DOI corresponding to a different bibliographic work; the frozen label set was retained for comparability with the original diagnostic results. The paired fixed-linker study similarly has an identifier-insertion confound: 10 of the corrected strings contain direct DOI information and the 14-pair outcomes cannot quantify net citation-network, h-index, or citation-count changes.

The evaluation does not assess full-text in-text citation alignment, semantic citation context, author intent, or misconduct.

9 Conclusion

Bibliographic-integrity verification under incomplete provider evidence should report both definitive decisions and the cost of withholding decisions. In the balanced 600-reference natural sample, the frozen approach covered 398/590 reference-standard-decisive cases, but 255 false-positive inconsistency alerts yielded only 3.41% precision; Endpoint B preserved all 25 review-positive cases while assigning 466/600 references to manual review. A post-hoc safety routing reduced definitive false positives to 40, but covered only 179/590 records and did not reduce human workload. On the separately constructed DOI-omitted 90-reference cohort, frozen V1 produced 5 true positives, 50 true negatives, 1 false positive, and 2 false negatives on 58 covered rows, while within-cohort parser repair auto-cleared five additional valid cases without fixing the false negatives. Four natural author-identity cases revealed no incremental decisive gain from surname-based decomposition. The bounded live-provider and fixed-linker analyses provide observable evidence of candidate and selected-DOI variation but do not establish causal bibliometric impact. The defensible contribution is therefore an auditable evaluation of the limits, operating points, and review obligations of evidence-traced bibliographic verification.

Statements and Declarations

Funding

No specific external funding supported the reported research.

Competing interests

No financial or non-financial competing interests are declared.

Ethics approval

The study examined publicly available bibliographic reference strings and scholarly metadata from published works. Independent expert assessors contributed bibliographic coding rather than personal or clinical participant data. The analysis did not involve interventions, clinical data, or personal research outcome measurements.

Consent to participate

The study did not recruit individuals for an intervention or collect personal research outcomes. The two assessors provided expert ratings of publicly available bibliographic material. The assessment task concerned bibliographic consistency of published references.

Data availability

The derived data and audit artefacts supporting the reported findings are preserved in a versioned project archive. Online Resources 1–3 provide the supplementary methodological material, the eight-journal frame, and the 80-article registry documenting

the 3,712-reference corpus, alongside the prospective 90-reference DOI-omitted cohort summary. Additional derived analysis outputs, statistical-audit files, and rights-cleared reproducibility workbooks are available from the corresponding author on reasonable request. Publisher-rendered full texts and third-party proprietary materials are not redistributed where publisher or provider terms do not permit redistribution; all primary articles remain accessible through the associated official DOIs and publisher landing pages.

Code availability

Source code, deterministic evaluation scripts (`citation_verifier.py`, `gate1_cohort.py`, and paired cluster-bootstrap/ablation scripts), frozen configuration files, and unit test suites are preserved in the versioned project archive and are available from the corresponding author on reasonable request under an open-science academic license.

Material availability

Online Resource 1 (`ESM_1_Supplementary_Material.pdf`): Supplementary methodological and validation material for the frozen 3,712-reference external corpus, selection boundary, and prospective 90-reference non-DOI cohort (Supplementary Methods S1–S2 and Tables S1–S4). Online Resource 2 (`ESM_2_Journal_Frame.csv`): Machine-readable journal-frame table for the eight-journal frozen corpus. Online Resource 3 (`ESM_3_80_Article_Registry.csv`): A machine-readable 80-article registry reconstructed from frozen Supplementary Table S2 and cross-checked against the original 3,712-row unlabeled corpus (80 article identifiers, exact DOI identifiers, and per-article counts; 0 critical mismatches). The derived registry preserves the article identifiers, persistent work identifiers, and per-article reference totals represented in Supplementary Table S2.

Author contributions

Ercan Erkalkan: Conceptualization; Methodology; Software; Validation; Formal analysis; Investigation; Data curation; Project administration; Writing—original draft; Writing—review and editing.

References

Abbonato D (2026) CheckIfExist: Detecting citation hallucinations in the era of AI-generated content. arXiv, https://doi.org/10.48550/arXiv.2602.15871

Ammar W, Groeneveld D, Bhagavatula C, et al (2018) Construction of the literature graph in Semantic Scholar. In: Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 3 (Industry Papers). Association for Computational Linguistics, pp 84–91, https://doi.org/10.18653/v1/N18-3011, URL https://aclanthology.org/N18-3011/

Aria M, D'Aniello L, Spano M (2026) A multi-phase reference matching algorithm for bibliometric analysis: Design, implementation, and evaluation. Scientometrics 131:6693–6728. https://doi.org/10.1007/s11192-026-05763-2

Chow CK (1970) On optimum recognition error and reject tradeoff. IEEE Transactions on Information Theory 16(1):41–46. https://doi.org/10.1109/TIT.1970.1054406

Crossref (2020) REST API. URL https://www.crossref.org/documentation/retrieve-metadata/rest-api/

Culbert JH, Hobert A, Jahn N, et al (2025) Reference coverage analysis of OpenAlex compared to Web of Science and Scopus. Scientometrics 130:2475–2492. https://doi.org/10.1007/s11192-025-05293-3

Delgado-Quirós L, Ortega JL (2025) Citation counts and inclusion of references in seven free-access scholarly databases: A comparative analysis. Journal of Informetrics 19(1):101618. https://doi.org/10.1016/j.joi.2024.101618

El Majjaoui A, El Haddadi A, Bahri A, et al (2026) Auditable bibliometric analysis: A generic data pipeline for research evaluation. Journal of Informetrics 20(2):101808. https://doi.org/10.1016/j.joi.2026.101808

El-Yaniv R, Wiener Y (2010) On the foundations of noise-free selective classification. Journal of Machine Learning Research 11(53):1605–1641. URL https://www.jmlr.org/papers/v11/el-yaniv10a.html

Forchino MV, Torres-Salinas D (2026) The OpenAlex database in review: Evaluating its applications, capabilities, and limitations. Journal of Informetrics 20(2):101800. https://doi.org/10.1016/j.joi.2026.101800

Geifman Y, El-Yaniv R (2019) SelectiveNet: A deep neural network with an integrated reject option. In: Proceedings of the 36th International Conference on Machine Learning, Proceedings of Machine Learning Research, vol 97. PMLR, pp 2151–2159, URL https://proceedings.mlr.press/v97/geifman19a.html

Khajavi K, Sadeghi S, Adhikari R, et al (2026) CiteCheck: Retrieval-grounded detection of LLM citation hallucinations in scientific text. arXiv, https://doi.org/10.48550/arXiv.2605.27700

Li M, Lin Z, Ma S (2026) Source or it didn't happen: A multi-agent framework for citation hallucination detection. arXiv, https://doi.org/10.48550/arXiv.2605.08583

Maisano DA, Mastrogiacomo L, Ferrara L, et al (2025) A large-scale semi-automated approach for assessing document-type classification errors in bibliometric databases. Scientometrics 130:1901–1938. https://doi.org/10.1007/s11192-025-05244-y

Nowakowska M (2025) A comprehensive approach to preprocessing data for bibliometric analysis. Scientometrics 130:5191–5225. https://doi.org/10.1007/s11192-025-05415-x

Priem J, Piwowar H, Orr R (2022) OpenAlex: A fully-open index of scholarly works, authors, venues, institutions, and concepts. arXiv, https://doi.org/10.48550/arXiv.2205.01833

Rong G, Chen Y, Koch T, et al (2026) Assessing data quality in citation analysis: A case study of Web of Science and Crossref. Journal of Informetrics 20(1):101775. https://doi.org/10.1016/j.joi.2026.101775

Russinovich M, Siva Kumar RS, Salem A (2026) Phantom references: Hallucinated citations that survive peer review at top-tier conferences. arXiv, https://doi.org/10.48550/arXiv.2607.00738

Tani A, Candela L, Castelli D (2013) Dealing with metadata quality: The legacy of digital library efforts. Information Processing & Management 49(6):1194–1205. https://doi.org/10.1016/j.ipm.2013.05.003

Yeung AWK (2026) Brain mutated into "brian": A cross-platform analysis of a common typographical error in neuroscience publications. Scientometrics 131:3515–3524. https://doi.org/10.1007/s11192-026-05642-w

Yin Z, Feng Y, Wang H, et al (2027) An agentic framework for reference integrity assurance via multi-source verification and topology-constrained citation consistency. Information Processing & Management 64(1):105075. https://doi.org/10.1016/j.ipm.2026.105075

Supplementary Files

This is a list of supplementary files associated with this preprint. Click to download.

- 04Supplementary.zip