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.

Citation traceability and inter-coder agreement in one scoping review: an exploratory audit

Yu Liu

ciciqq1105@hotmail.com

General Hospital of Central Theater Command of PLA

Xiang Wang

General Hospital of Central Theater Command of PLA

Jun Hou

General Hospital of Central Theater Command of PLA

Jianjun Tang

General Hospital of Central Theater Command of PLA

Meihua Guo

General Hospital of Central Theater Command of PLA

Research Article

Keywords: Meta-research audit, Multimodal artificial intelligence in dentistry, Dental AI evaluation, Data extraction, Inter-rater reliability, Citation traceability

Posted Date: November 1st, 2026

DOI: https://doi.org/10.21203/rs.3.rs-11208912/v1

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

Additional Declarations: Competing interest reported. Yu Liu, Xiang Wang, Jun Hou, Jianjun Tang and Meihua Guo declare that they have no financial competing interests. They declare one non-financial interest: the audited object is the team's own unpublished scoping review, in which the team has no commercial interest. No author has a relationship with BMC Medical Research Methodology or with BMC (for example as an Editorial Board Member or Guest Editor).

Abstract

Background

Agreement between extractors does not show whether their citations lead back to the supporting text, nor whether screening eligibility rules reach the fields later extracted.

Methods

We audited a scoping review by our team: 277 included reports (coded corpus 390) and 29 study-level fields. Co-primary measures were field-level proportion of agreement (Po) across the 29 fields and the whole-passage verbatim-anchoring rate of 9,415 citation judgments (coder A, pre-adjudication), with Krippendorff's alpha (α) secondary. Citation checks establish relocation, not support. The registered screening protocol required the visual channel to be used in the pipeline of the evaluated model; that rule had not reached the extracted image-channel field, so we re-applied it. Two reviewers independently re-assessed 115 records (Cohen's kappa = 0.83) with third-party arbitration; 275 coded records were not re-verified.

Results

Re-applying that pre-specified rule excluded 107 records, confirmed 2, and left 6 unresolved (excluded from synthesis; sensitivity 283), reducing the 390-record coded corpus to 277 included reports. Reliability and anchoring denominators remain the 390-record coded corpus, not 277 included reports. Eight of 29 fields fell below Po 0.80; of 26 fields with estimable α, 11 fell below 0.67, seven of those with Po ≥ 0.80. Three free-text fields (exact-match only) ranged from 0.04 to 1.00. Of 9,415 citation judgments, 6,215 (66.0%) were whole-passage verbatim, 2,229 (23.7%) first-25-character only, and 971 (10.3%) unmatched; a tolerant search relocated 946 without verifying support. Gate 1 returned 15/23 (65.2%) verbatim re-hits, below the ≥ 90% threshold, and stopped Gate 2; Gate 3 returned Po = 0.6435 across the 115-row attribution ledger (115 attribution-judgment cells; 74/115).

Conclusions

This audit shows the instrument built to classify extraction failures failed its own checks in the way it was designed to detect; agreement and citation anchoring need separate checks. Five operationalisations (field-type classification, free-text exclusion, floors, 26-field grouping, prevalence-adjusted analyses) and the co-primary measures were set after the data were visible. These exploratory results do not establish the correctness of extracted values or support a clinical conclusion.

1. Background

Reliable extraction is essential to systematic reviews, including work that aims to automate evidence synthesis. [1, 2] Studies of automated extraction report continuing challenges and a need for human oversight. [1, 3–5] Agreement alone cannot establish whether an extracted value can be traced to its source text. This audit therefore examines agreement and citation traceability separately.

The first check is reliability: whether independent coders assign the same value from the same source. Po, the proportion of coding cells in which both coders chose the same category, counts that concordance at face value. When one category dominates the marginal distribution, Po can stay high even when coders disagree on the rare categories that carry the field's meaning. Krippendorff's α [6] is a chance-corrected nominal coefficient that removes agreement expected from the margins; when the margins concentrate, chance agreement rises, so α falls while Po holds.

The second check is traceability: whether the cited text can be located in the paper. Claim–source misquotation is documented in biomedical literature. [7, 8] String comparison tests whether the quoted text matches the source word for word; a string mismatch alone does not establish that the claim is inaccurate. A whole-passage verbatim match establishes relocation only, not support; a match on the first 25 characters only fixes where the passage starts and leaves the rest unconfirmed; an unmatched citation leaves the link open.

We audited the extraction records of our team's own unpublished scoping review (277 included reports; coded corpus 390) across 29 study-level fields, re-anchoring all 9,415 citation judgments against source text. The two co-primary measures were field-level Po across the 29 fields and the whole-passage verbatim-anchoring rate of the 9,415 citation judgments. The audit's own diagnostic instrument — the failure-attribution classification — failed the same checks it was designed to apply, so consistency was not equivalent to citation traceability in this record. The co-primary measures were specified after the extracted data were visible and the registration was amended during the study, so the analysis is exploratory and contains no clinical conclusion. It examines the review's extraction record, not the clinical question the review was built to answer.

2. Methods

2.1 Design and registration

This meta-research study audited the authors' own scoping review: one formal extraction round of 390 coded records across 29 study-level fields. The audited object is described minimally in five parts: (1) 842 full texts were assessed; 401 raw rows minus 3 superseded rows left 398 records, and 398 minus 8 duplicate records of 7 studies left 390 coded records (398 is a record-level count, not a screening-stage count); a post-hoc re-adjudication of eligibility (§2.9) reduced these to 277 included reports; (2) the 29 study-level fields comprise 25 Tier-1 extraction fields plus 4 analysis-derived fields (Supplementary Table S1d); (3) the 9,415 citation judgments are coder A's non-empty evidence quotes, graded before adjudication by string comparison; (4) the configuration-level count is not closed in the frozen ledger and

is not reported; (5) the frozen rule set, de-identified data and reproduction code are deposited on Zenodo (DOI 10.5281/zenodo.23005564). The audited review itself is not cited as a reference: it is an unpublished observation with OSF registration osf.io/mhjtz and the Zenodo data deposit. This is not external validation. The review was registered on OSF (osf.io/mhjtz; DOI 10.17605/osf.io/mhjtz) on 12 September 2026; later amendments are logged in Supplementary Table S5. The analyses follow this registration, not an earlier one (osf.io/rx8sm, 30 August 2026). The 12 September registration fixed the extraction protocol; the reliability thresholds were registered later (Supplementary Table S5 V3). The co-primary measures (field-level Po across 29 fields and the verbatim-anchoring rate of 9,415 citation judgments, with α secondary) were specified after the extracted data were visible; the analytic axis changed twice and the gate order was revised. These changes are reported as post hoc (Table 5). We therefore present the audit as exploratory and hypothesis-generating: it demonstrates the procedure rather than testing a prespecified hypothesis. Each question has its own unit of analysis: study-level fields for reliability, citation judgments for anchoring, and attribution-judgment cells for Gate 3 (Figure 1; Table 1; Table 4; Supplementary Table S8 #2, #7, #8, #10).

2.2 Frozen rules

The field definitions (the 29 analysis fields include four analysis-derived fields defined in Supplementary Table S1d), adjudication criteria, three anchoring tiers, Gate 1 threshold and gate definitions were fixed before the relevant audit steps. This did not freeze every analytic choice: the Po floor, the 0.67 α floor, field grouping and the co-primary measures were set after the data were visible (§2.8). The gate order also changed after the data were visible (§2.6; Table 5). The criteria table was version-bound, so every adjudication traces to the rule text in force at the time. Gate 2 was not initiated (Figure 4; Table 4).

2.3 Extraction and adjudication

Two coders independently coded the 29 study-level fields for each report; a third, independent arbiter resolved disagreements. The arbiter's decisions were recorded separately against the criteria-table version in force, so disagreement cells remain visible after adjudication. One batch was voided after review showed that the criteria table covered only part of the observed coding paths. It was sealed with a hash manifest, so any reuse of the voided batch is detectable. The table was rebuilt so that every path mapped to a defined category; the rebuilt table governs only the post hoc re-run (Supplementary Tables S4, S7; Table 4).

2.4 Reliability

For each field we computed Po and, where the denominator permits, Krippendorff's α, keeping category counts so that marginal distributions stay visible. α was computed for 26 fields: 20 single-valued nominal fields and six binary endpoint-level indicators (L1–L6, the per-label decomposition of one multi-label variable). Three open-string transcription fields (intended_use_author, model_version,

primary_instance_id) carry Po only. They have no closed category set, so chance-expected agreement over literal strings would not measure the semantic agreement of interest. Pc is used only descriptively in Figure 5 as the chance-expected agreement computed from the product of the two coder-arm marginal proportions; it is not the chance model underlying Krippendorff's α. Table 2 summarises fields by raw and chance-corrected status; Supplementary Table S1 lists every field with its Po and, where computed, its α. The Figure 2 caption records eight fields below Po 0.80 and seven fields that cleared Po while α fell below 0.67.

The Po floor of ≥0.80 was adopted after the data were visible and is not registered. The α ≥ 0.80 floor was registered (Supplementary Table S5 V3) but the analysis used 0.67, Krippendorff's tentative-conclusions convention — a post hoc relaxation. A field that clears the Po floor but misses the α floor is the case this audit targets. Alternative chance-corrected coefficients include prevalence- and bias-adjusted forms [9] and analyses of the high-agreement, low-kappa case [10]; conventional strength-of-agreement bands [11] are descriptive only (Figure 2; Tables 1–2; Supplementary Table S1).

2.5 Citation anchoring

All 9,415 citation judgments were re-anchored against source text and assigned to one of three mutually exclusive tiers. Whole-passage verbatim: the full cited span matches character for character. First-25-character only: the opening 25 characters match and the remainder does not. Unmatched (failed the strict string rule): comparison stopped short of the 25-character threshold. Re-check showed these 2,229 quotes matched well beyond 25 characters (only 34 stopped at exactly 25), so this tier records trailing non-verbatim text, not truncated excerpts. Text was NFKC-normalised with whitespace removed. The whole-passage verbatim and first-25-character-only tiers were graded under first-round comparison conditions only; only the unmatched tier was later re-reviewed under complete full-text conditions. Each non-empty quote in coder A's evidence columns counted as one judgment and took the first tier it satisfied; coder B's quotes were not graded in this round, so the 9,415 are single-coder (coder A), pre-adjudication citations. Coder B's per-cell citation evidence is retained in the frozen extraction file (23 evidence-class columns overall; Supplementary Table S3b lists 22 columns with remaining single-citation failures; 387 of 390 records with quoted text); a two-coder anchoring assessment is listed as a pending manual step.

Unmatched citations fell into three groups: citations in multi-citation cells (721); single-citation zero-prefix citations with no 8-character prefix match in the source (50); and other single-citation failures (200). A null model gave a descriptive random-match baseline of 0.0016% (30 hits in 1,883,000 draws); it gauges chance coincidence, not support (Figure 3; Table 3; Supplementary Tables S2–S3). The pre-declared stop rule triggered a full re-run of the 721 multi-citation-cell citations (594 unique cells), and 610 disagreement rows across three first-round batches went to a third arbiter. These parallel records do not replace the frozen main-text values for coder A's pre-adjudication judgments (6,215 / 2,229 / 971; 66.0% / 23.7% / 10.3%; Supplementary Table S7).

2.6 Quality gates

Three gates were defined, each with its own denominator; their evaluation order was revised after the data were visible. Gate 1 (interim futility) required a verbatim re-hit rate of ≥90% plus three co-conditions — original-text evidence, an applicable rule, and an independent attribution record — on a simple random subset of 23 cells (seed 20260925) drawn from the disagreement-enriched 115-row attribution ledger (76 disagreement, 39 control); only verbatim retrieval was scored in the frozen main run; the three co-conditions were assessed only in the post hoc exploratory record (Supplementary Table S4c), so the 65.2% is not a general extraction-cell verbatim rate. Gate 2 depended on Gate 1 and was not initiated. Gate 3 required Po≥0.80 on the 115-row attribution ledger (76 disagreement, 39 control) and was evaluated out of the frozen order; the departure was registered on OSF (osf.io/xm6kq, 25 September 2026). Gate 1's subset was drawn from the same 115-row attribution ledger that Gate 3 reads, so the two samples are not independent (Figure 4; Table 4; Supplementary Tables S4c, S8).

2.7 Supplementary material and reporting

Supplementary Tables S1–S10 are online only, including the numbered subsections of S1 and S4. TRIPOD+AI [12], PRISMA-trAIce [13] and TRIPOD-LLM [14] are cited as background only; none was applied. Reporting followed PRISMA 2020 [15] for the items that apply to a meta-research audit of a completed review; no recognised reporting checklist specific to meta-research audits of completed reviews exists, so applicable PRISMA 2020 items are used and non-applicable items are given item-by-item reasons in the adapted checklist. The audited scoping review's protocol was prepared in accordance with PRISMA-ScR and JBI guidance for scoping reviews; the present audit is reported against PRISMA 2020 items that apply to a meta-research audit, so the two reporting frameworks apply to different objects. This study performed no new literature screening; the inclusion flow of the audited corpus is summarised from the frozen screening records (results §3.1; Supplementary Table S-map). The underlying review's flow counts come from the internal frozen log and are supplied in the package but not publicly archived, which limits independent verification of corpus formation. The adapted checklist is supplied as Supplementary Material. Supplementary tables map each field to its agreement metrics and each citation tier to its counts, so denominators and category margins can be inspected. Only the unmatched citation tier was re-reviewed under complete full-text conditions; the whole-passage verbatim and first-25-character-only tiers were not subject to the same full-text check. For coder A's pre-adjudication citation judgments, depending on which tiers are counted as anchored, the string-level relocatability estimate ranges from 66.0% (whole-passage verbatim only) to 89.7% ((6,215 + 2,229)/9,415) to 99.7% ((6,215 + 2,229 + 946)/9,415; Supplementary Table S9.2). Twelve cells remained unverified at submission and are listed, outside the frozen denominators, in Supplementary Table S6. The frozen rule set, de-identified analysis files and reproduction code are deposited on Zenodo (version DOI 10.5281/zenodo.23005564; concept DOI 10.5281/zenodo.22953344). On request, the de-identified extraction records, the frozen rule-set versions and the hash manifest can be made available to the editor or reviewers under confidential access.

2.8 Deviations from the registered agreement-reporting plan

The registration specified percentage agreement plus, where the denominator permits, a chance-corrected statistic. It did not pre-specify the field count, the thresholds, or the handling of binary and free-text fields. Five operationalisations were therefore made during analysis, after the data were visible: (1) the field-type classification; (2) excluding three free-text transcription fields from α; (3) the Po 0.80 floor and the α 0.67 floor (the latter relaxing the registered α ≥ 0.80 floor); (4) replacing the earlier 20-nominal-field grouping with the 26-field α set; and (5) the post hoc prevalence-adjusted agreement analyses for binary indicators, distinct from the unresolved-record sensitivity analysis in §2.9. Supplementary Table S1a compares registration and practice in full.

2.9 Post-hoc re-adjudication of eligibility

Eligibility was re-checked against the screening rule (§6.1 TA-3; §6.2 FT-2) requiring the visual channel to be exercised in the evaluated pipeline. The re-adjudication used Screening SOP v1.0 (12 September 2026) §6.1 TA-3 and §6.2 FT-2; V4 (2026-09-21) revised the categorisation and reporting of “full text not obtainable” only and did not change TA-3 or FT-2.

Screening rule text (Screening SOP v1.0, 12 September 2026):

§6.1 Title/abstract stage exclusion code: TA-3 No visual input (text-only LLM study) §6.2 Full-text stage exclusion code: FT-2 Visual input not exercised §6 fixed priority (highest first): TA-2 (not an original study) -> TA-1 (not dental/oral) -> TA-4 (non-LLM/VLM route) -> TA-3 (no visual input) -> TA-5 (duplicate). That rule had not been applied to the extracted image-channel field (E1_image_input) at screening: 115 of the 390 coded records (the 115-record eligibility re-adjudication subset) carried E1 = no or unable to determine. The remaining 275 coded records were not re-verified; this re-adjudication therefore addresses protocol conformance for the flagged subset only. Two reviewers re-adjudicated this 115-record eligibility re-adjudication subset independently (Cohen κ = 0.8311; 95% CI [0.58, 1.00], bootstrap 5,000 resamples; 3 categories: eligible / ineligible / indeterminate; chance model: Cohen κ with observed marginal distributions; n = 115; based on the two reviewers’ pre-arbitration judgments; dichotomised sensitivity ineligible vs other: Cohen κ = 0.8355); three disagreements went to a third adjudicator. 107 did not exercise visual input, two did, and six remained unresolved; unresolved records were not counted as included. The six unresolved records were unresolved because the supplied full texts or records did not contain enough information to determine whether the visual channel was exercised (evidence insufficiency), not because the evidence showed ineligibility. The re-adjudicated included set is 277 reports; the remaining 275 coded records were not re-verified and are not eligibility-confirmed. Two records judged eligible carried E1 values of no or unable to determine in the frozen extraction, indicating a field-value error that is documented here (2/115 ≈ 1.7%). Sensitivity analysis (unresolved records): if the six unresolved records were counted as included, the included set would be 283; this is a sensitivity figure only and does not change the main result of 277; it does not alter any reported estimate, because all reliability and citation-anchoring values are computed on the 390-record coded corpus. If the three

arbitration-disagreement records were instead counted as eligible, the included set would be 280; this directionality concerns only 3 records and likewise does not change the main result of 277 or any reported estimate. The κ = 0.83 was computed on the 115-record eligibility re-adjudication subset; three disagreements went to a third adjudicator; in all three, reviewer A judged ineligible, reviewer B judged indeterminate or eligible, and the arbiter took reviewer A's ineligible ruling. This directionality warrants caution in treating arbitration as neutral, particularly given the same arbiter's arm-B preference in the separate first-round batches (Supplementary Table S7); it does not change the reported 107 excluded, 6 unresolved, 2 confirmed or 277 included.

Accordingly, the reliability and anchoring denominators remain the 390-record coded corpus, and the post-adjudication included set is 277. The two denominators are shown side by side and are not interchangeable. For field-level reliability estimates that depend on the marginal distributions of E1_image_input, claimed_modality, and actual_input_modality, restricting to the 277 included records would condition on the measured variable and would collapse those margins, degrading their α (e.g., E1_image_input 0.849 to 0.097; claimed_modality 0.717 to 0.303; actual_input_modality 0.755 to 0.372), so those field-level estimates would not be interpretable for those fields. Citation anchoring units are the citation records from the coding round and do not depend on the E1 margin; they therefore likewise remain on the 390-record coded corpus and do not share that restriction. The diagnostic re-estimates were computed on the coded corpus for diagnosis only, and not substituted for the reported estimates.

3. Results

3.1 Audited corpus (limited map)

The audited review comprised 277 included reports (coded corpus: 390 records), reached from 842 full texts assessed (401 raw rows − 3 superseded rows = 398 records; 398 − 8 duplicate records of 7 studies = 390 coded records; 398 is a record-level count, not a screening-stage count); the reduction from 390 to 277 follows the eligibility re-adjudication described in §2.9. In the coded corpus of 390 records, by claimed modality 293 (75.1%) were image-plus-text, 82 (21.0%) text-only, seven image-only and eight unclear; the descriptive task groupings were led by education/examination (118) and imaging diagnosis (85); codebook field 11 separately specifies nine permitted task-domain values; the most common sample units were questions (128), images (112) and cases (92); and 312 reports were peer-reviewed journal publications, 54 preprints, 23 conference papers and 1 unclear or not reported (frozen extraction-A.csv; Supplementary Table S-map). This distribution describes the 390-record coded corpus, not the 277-report included set. This map is a descriptive summary of the corpus and must be read together with the reliability and traceability audits that follow; it is not a verified clinical-evidence summary. The configuration-level count was not closed in the frozen ledger and is not reported.

3.2 Reliability

Eight of 29 study-level fields fell below the Po floor of 0.80. Of the 26 fields with estimable α, 11 fell below 0.67 (α 0.155–0.606): six single-valued nominal fields and five of six binary endpoint-level indicators. Under the registered α ≥ 0.80 floor, only 4 of the 26 fields would have cleared (population_stratum 0.8597, E1_image_input 0.8494, evidence_category 0.8661, source_type 0.9042; Supplementary Table S1c), fewer than the 15 that cleared the 0.67 floor used in analysis. Seven of the 26 cleared Po≥0.80 while α fell below 0.67; observed agreement alone would have passed them. Binary indicators tracked prevalence: endpoint_L4 (1.5% yes) paired Po 0.974 with positive-specific agreement 0.167 (Supplementary Table S1c.2). The open-string field intended_use_author, assessed by exact match only, had the lowest Po of all 29 fields (0.041). Figure 2 shows raw and chance-corrected agreement side by side; Supplementary Tables S1 and S9 list every field's metrics and 95% intervals (Figure 2; Figure 5; Tables 1–2; Supplementary Table S1).

3.3 Citation traceability

Of coder A's 9,415 citation judgments, 6,215 (66.0%; Supplementary Table S9.1) were whole-passage verbatim matches, 2,229 (23.7%) matched on their first 25 characters only, and 971 (10.3%) were unmatched. Unmatched means failure of the strict string rule, not absence of source text: a separate, typesetting-tolerant search located source text for 946 of 971 (Supplementary Table S7). Finding text does not establish that it supports the extracted value (Supplementary Table S9.2). Only the unmatched tier was re-reviewed under complete full-text conditions (Supplementary Table S7); the whole-passage verbatim and first-25-character-only tiers were not re-checked under the same conditions. The stopping rule used the strict tier. All proportions are descriptive counts on the judgment-level denominator of 9,415; no inferential interval is reported because the judgments cluster within reports and cells. The 971 unmatched judgments comprised 721 in multi-citation cells and 250 in single-citation cells; of the 250, 50 (0.53% of 9,415) matched no 8-character prefix (zero-prefix failures) and 200 were other single-citation failures. The whole-passage tier lies well above the descriptive null-model baseline (30 hits in 1,883,000 draws, 0.0016%); first-25-character-only matches still need a human reader to confirm the rest of the passage (Figure 3; Table 3; Supplementary Tables S2–S3).

3.4 Gates

In the 23-cell Gate 1 diagnostic subset, 15 cells (65.2%) were verbatim re-hits, below the ≥90% threshold, so Gate 2 was not advanced; the frame is enriched for disagreements, so 65.2% is not a general extraction-cell verbatim rate. Gate 3 returned Po=0.6435 across the 115-row attribution ledger (76 disagreement, 39 control; 74/115), below its 0.80 floor. Across the two arms' 230 primary attributions (115 cells × two coders), the largest single category accounted for 74 of 230 (32.2%), so no failure category was separable. All 23 Gate 1 keys occur in the Gate 3 ledger; the frozen denominator remains 115 rows (Figure 4; Table 4).

3.5 Reading the two checks together

Reliability and traceability answer different questions. Weak agreement marks fields that coders read differently from the same source; an unanchored citation marks a value whose evidential link is still open. A field can pass one check and fail the other; their combinations suggest different next steps, though not counted here. High agreement with intact anchoring shows consistent, locatable coding, not verified correctness. High agreement with weak anchoring means the supporting text still has to be found. Lower agreement with intact anchoring points to the field definition or coder training. Lower agreement with weak anchoring combines both; re-reading the field definition is a reasonable first step.

3.6 Voided batch

The voided batch is excluded from these results. Its removal is recorded, not reconstructed: analyses used the surviving rule set, and the batch appears only as a disposition with its hash manifest (Supplementary Table S4; Table 4).

4. Discussion

The audit separates two properties that review reports often merge: reliability (whether judgments converge) and traceability (whether the supporting text can be recovered). In this sample the two properties were measured separately; no joint estimate is offered. Eight fields fell short on raw agreement and 11 on chance-corrected agreement; seven cleared the raw floor but missed the chance-corrected one while their marginal distributions concentrated in a single category. Reporting Po beside α, with marginal tables attached, makes that gap visible. That raw-versus-chance-corrected divergence is well documented [9, 10]; what the audit adds is locating it field by field in a real extraction record and pairing it with graded citation anchoring on the same round.

Citation traceability (coder A, pre-adjudication, string-level relocatability) was incomplete. Two-thirds of these judgments (6,215 of 9,415) were recovered as whole passages, 23.7% only as a leading fragment, and 10.3% failed the strict string rule. Multi-citation cells accounted for 721 unmatched judgments; the other 250 were single-citation failures, of which 50 (0.53% of 9,415) matched no 8-character prefix (zero-prefix failures). The denominator was left at 9,415 and not recalculated for full-text availability. A separate, tolerant search located source text for 946 of the 971 strict-rule failures (Supplementary Table S7). This did not reclassify the strict matches or verify support for the extracted values.

In the revised gate order, Gate 1 returned 15 of 23 (65.2%) verbatim retrievals, below the ≥ 90% threshold, so Gate 2 was not initiated; Gate 3, evaluated out of sequence, returned Po = 0.6435 on the 115-row attribution-ledger denominator (76 disagreement, 39 control).

The central observation is that the diagnostic instrument built for this audit failed its own checks in the same way it was designed to detect. Three records support this: (1) the K100 judgment batch was

voided and sealed: its rubric categories were neither mutually exclusive nor exhaustive, and the rule table was rebuilt (Supplementary Table S4); (2) the first-round anchoring comparison counted 104 “source not found” rows, whereas a complete full-text re-run left 20, so the first-round measurement reflected truncated comparison excerpts — the same evidence-boundary artifact this audit attributes to the reviewed studies (Supplementary Table S7); (3) arbitration across the three first-round batches matched arm B in 82.7%, 56.6% and 66.5% of rows, a systematic favouring of one arm rather than neutral resolution (Supplementary Table S7).

A practical audit should test the codebook before freezing it, retain independent coding and adjudication as separate records, and link each citation to its source. It can then report Po and α with category counts, grade citation matches, and re-run affected work under a dated rule version. This sequence is a candidate procedure that requires prospective validation before use in other reviews; a validation protocol is registered separately (Supplementary Table S10).

Several constraints bound these estimates. The co-primary measures were post hoc and the registration was amended during the study, so not every analytic decision was prespecified. The sample comprises the extraction record of one scoping review of multimodal AI in dentistry; transfer to other reviews or clinical domains remains untested. The authors audited their own review, limiting independence from prior extraction decisions. Gate 1 and Gate 3 share records (all 23 sampled keys occur in the 115-row ledger; Supplementary Table S4b), so the samples are not independent. In Supplementary Table S4c the three co-conditions are completely collinear with the verbatim-hit outcome (the eight non-hit cells are exactly the cells where condition (1) is unmet), and “rule applies” was assigned in the opposite direction from hit status; this circularity limits the independent evidentiary value of that co-condition record. Gate 2 stayed contingent, three open-string fields carry Po only, and the voided batch was set aside rather than reused. The design yields reliability and traceability estimates, not a pooled effect estimate or a clinical claim. It measures traceability and consistency, not the correctness of extracted values; whether each value rests on the correct source, location and supporting text remains to be verified. Three procedural records are held in the study log and the OSF record: the registration update log, the 115-row attribution-ledger composition, and the authorisation of the out-of-sequence Gate 3 evaluation (25 September 2026). The eligibility re-adjudication reported in § 2.9 was likewise conducted within the author team: the agreement between the two reviewers (κ = 0.83) indicates internal reproducibility rather than external validation, it does not establish that the remaining records are eligible, and it does not extend the pre-adjudication reliability estimates to the 277-report included set. The 390-record denominator and the 277-report included set are reported side by side by design; recomputing field-level reliability estimates on the 277-report set would condition on an extracted field and would not be interpretable for those fields. Citation anchoring is not an E1 marginal problem; the diagnostic values were computed on the coded corpus for diagnosis only, and not substituted for the reported estimates. Independent verification of the included set remains outstanding.

5. Conclusion

In this audit of the authors' own review, raw agreement alone would have passed seven fields that fell below the chance-corrected floor, and 6,215 of coder A's 9,415 (66.0%) citation judgments were recoverable as whole passages. The proposed audit pairs agreement measures with category counts and graded citation matching. Its gate order changed after the data were visible, so the procedure and its stopping decisions need independent testing before use in other reviews.

Declarations

Declaration of generative AI and AI-assisted technologies in the manuscript preparation process

During preparation of this work the authors used large language model assistants (Anthropic Claude, OpenAI GPT, Google Gemini and DeepSeek) for a deliberately restricted set of tasks: language editing, critical reading of drafts, assistance in writing the analysis code, and checking of formatting and reference formatting. Their use was bounded as follows. No AI tool screened, extracted, coded or adjudicated any data, and no reported value was produced by an AI tool: every coded value was produced by two independent human extractors with a third human arbiter. No AI tool generated the research question, the analytical reasoning or the conclusions, and no AI tool selected which analyses to report. The statistical analyses were run on the frozen coded data, and all AI-assisted output was reviewed, verified and edited by the authors, who take full responsibility for the content of the publication. This use is disclosed in accordance with the journal's policy on artificial intelligence. Model names such as BLIP, CLIP, GPT or Gemini appearing in the audited extraction or supplementary tables identify the models studied in the audited literature, not the AI assistants used to prepare this manuscript.

Role of automated processing in the reported work. The audited extraction round was performed entirely by humans: two coders worked independently under the codebook described in §2.3 to code every field, with disagreements resolved by a third human arbiter, and every coded value in the extraction record was produced by human coders, not by any automated process. Automated processing was limited to non-extraction tasks: structural validity checks (headers, closed vocabularies, value ranges), computation of the agreement statistics reported here, and the citation-anchoring scan, which compared extracted quotations with source text by string matching.

Availability of data and materials. The frozen rule set, de-identified analysis files and reproduction code are available in the Zenodo repository — "Human-frozen extraction and verification data, and pre-frozen protocol artifacts, for an audit of extraction reliability and citation traceability in a scoping review of multimodal artificial intelligence in dentistry" — https://doi.org/10.5281/zenodo.23005564 (version DOI; concept DOI 10.5281/zenodo.22953344) [16]. In addition, the extraction codebook (field definitions, permitted values and decision rules) and the record-level inclusion list of the audited corpus are supplied

with this submission as Supplementary-Codebook and Supplementary-Corpus-List. The process counts derived from the internal screening log are not publicly archived; they are available to the editor or reviewers on request. Reuse terms are stated in the license attached to the Zenodo record. Quotations from audited reports are reproduced only as short evidence spans for audit verification.

Pre-registration. The review was registered on OSF at osf.io/mhjtz (DOI 10.17605/osf.io/mhjtz) on 2026-09-12; later amendments are logged in Supplementary Table S5. An earlier registration of the audited review was made on OSF (osf.io/rx8sm) on 30 August 2026; the audited review was conducted under the protocol registered on 12 September 2026 (osf.io/mhjtz) and its update log, including separately registered amendments (Table 5). The present meta-research audit does not rely on the earlier registration and includes the post hoc analyses marked as such (§2.8). The co-primary measures were specified after the extracted data were visible and are reported as post hoc (§2.1). The OSF registration lists the corresponding author as its sole registered contributor; the author block lists the full analysis team, and individual roles are given in the CRediT statement. The OSF registration title is *A Scoping Review Protocol* and refers to the audited review, not to the present meta-research audit; this study is a meta-research audit of that review and was not separately registered as a review protocol.

Reporting guideline. PRISMA 2020 (Reference 15), applied to the items that bear on an audit of a completed review; non-applicable items carry a recorded rationale, and the adapted checklist is supplied as Supplementary Material.

Ethics approval and consent to participate. Not applicable. The study audited publicly available reports, preprints, conference proceedings and bibliographic records and involved no human participants and no identifiable participant data. Consent to participate is therefore not required.

Clinical trial number: not applicable.

Consent for publication. Not applicable. The manuscript contains no individual person's data in any form.

Trial registration. Not applicable; this study is a meta-research audit of an unpublished scoping review and of publicly available reports, which did not prospectively enrol participants.

Funding. This research did not receive any specific grant from funding agencies in the public, commercial, or not-for-profit sectors.

Competing interests. The authors declare that they have no financial competing interests. The authors declare a non-financial interest in that the audited object is their team's unpublished scoping review; the team has no commercial interest in its results.

Authors' contributions. Yu Liu: conceptualization, screening, data extraction, formal analysis, supervision, project administration, writing - original draft; corresponding author. Xiang Wang: screening, data extraction, formal analysis, writing - original draft. Jun Hou: arbitration of disagreements, validation,

writing - review and editing. Jianjun Tang: software, formal analysis, data curation, visualization. Meihua Guo: methodology, validation, investigation, resources, writing - review and editing. Eligibility re-adjudication was performed independently by Yu Liu and Xiang Wang, with Jun Hou as third adjudicator. All authors read and approved the final manuscript and agree to be accountable for all aspects of the work.

Acknowledgements. None.

References

1. Shankar R, Lim A, Qian X. Performance of large language models in data extraction for evidence synthesis: a systematic review. J Biomed Inf. 2026;181:105086. 10.1016/j.jbi.2026.105086.
2. Li L, Mathrani A, Susnjak T. Transforming evidence synthesis: a systematic review of the evolution of automated meta-analysis in the age of AI. Res Synth Methods. 2026;17(3):403–50. 10.1017/rsm.2025.10065.
3. Li L, Mathrani A, Susnjak T. What level of automation is good enough? A benchmark of large language models for meta-analysis data extraction. Res Synth Methods. 2026;17(4):671–92. 10.1017/rsm.2025.10066.
4. Mahmoudi H, Chang D, Lee H, Ghaffarzadegan N, Jalali MS. Critical assessment of large language models' (ChatGPT) performance in data extraction for systematic reviews: exploratory study. JMIR AI. 2025;4:e68097. 10.2196/68097.
5. O'Connor AM, Clark J, Thomas J, Spijker R, Kusa W, Walker VR, et al. Large language models, updates, and evaluation of automation tools for systematic reviews: a summary of significant discussions at the eighth meeting of the International Collaboration for the Automation of Systematic Reviews (ICASR). Syst Rev. 2024;13(1):290. 10.1186/s13643-024-02666-2.
6. Krippendorff K. Content Analysis: An Introduction to Its Methodology. 4th ed. Thousand Oaks, CA: SAGE; 2019. 10.4135/9781071878781.
7. Peoples N, Østbye T, Yan LL. Burden of proof: combating inaccurate citation in biomedical literature. BMJ. 2023;383:e076441. 10.1136/bmj-2023-076441.
8. Baethge C, Jergas H. Systematic review and meta-analysis of quotation inaccuracy in medicine. Res Integr Peer Rev. 2025;10(1):13. 10.1186/s41073-025-00173-z.
9. Byrt T, Bishop J, Carlin JB. Bias, prevalence and kappa. J Clin Epidemiol. 1993;46(5):423–9. 10.1016/0895-4356(93)90018-V.
10. Feinstein AR, Cicchetti DV. High agreement but low kappa: I. The problems of two paradoxes. J Clin Epidemiol. 1990;43(6):543–9. 10.1016/0895-4356(90)90158-L.
11. Landis JR, Koch GG. The measurement of observer agreement for categorical data. Biometrics. 1977;33(1):159–74. 10.2307/2529310.
12. Collins GS, Moons KGM, Dhiman P, Riley RD, Beam AL, Calster BV, et al. TRIPOD + AI statement: updated guidance for reporting clinical prediction models that use regression or machine learning

methods. BMJ. 2024;385:e078378. 10.1136/bmj-2023-078378.
13. Holst D, Moenck K, Koch J, Schmedemann O, Schüppstuhl T. Transparent reporting of AI in systematic literature reviews: development of the PRISMA-trAIce checklist. JMIR AI. 2025;4:e80247. 10.2196/80247.
14. Gallifant J, Afshar M, Ameen S, Aphinyanaphongs Y, Chen S, Cacciamani G, et al. The TRIPOD-LLM reporting guideline for studies using large language models. Nat Med. 2025;31(1):60–9. 10.1038/s41591-024-03425-5.
15. Page MJ, McKenzie JE, Bossuyt PM, Boutron I, Hoffmann TC, Mulrow CD, et al. The PRISMA 2020 statement: an updated guideline for reporting systematic reviews. BMJ. 2021;372:n71. 10.1136/bmj.n71.
16. [dataset] Liu Y, Wang X, Hou J, Tang J, Guo M. Citation traceability and inter-coder agreement in one scoping review: an exploratory audit (frozen rule set, de-identified extraction and verification data, and reproduction code). Zenodo. 2026. 10.5281/zenodo.23005564

Tables

Table 1. Analytic units and denominators

analytic_unit	n	role_in_audit
Reports (formal extraction round)	390	sampling frame
Study-level fields	29	reliability units
Fields with computed alpha	26	20 nominal + 6 binary endpoint indicators
Free-text transcription fields (Po only)	3	open-string, no closed category set
Fields with Po below 0.80	8	raw agreement failures
Fields with alpha below 0.67 (of the 26)	11	chance-corrected failures (nominal 6/20, binary 5/6)
Fields passing raw but failing chance-corrected floor	7	nominal 3 + binary 4
Coder-A pre-adjudication citation judgments	9415	anchoring denominator
Coder-A judgments: verbatim	6215	tier 1
Coder-A judgments: prefix-only	2229	tier 2
Coder-A judgments: unmatched	971	tier 3
unmatched in multi-citation cells	721	subset
unmatched zero-prefix (single-citation cells)	50	subset (disjoint from 721)

Table 2. Reliability summary

summary	value
Fields audited	29
Po >= 0.80	21
Po < 0.80	8
Alpha computed	26
Alpha >= 0.67	15
Alpha < 0.67	11
Alpha < 0.67, by type (post hoc)	nominal 6/20; binary endpoint labels 5/6
Free-text transcription fields (Po only)	3
Po >= 0.80 and alpha < 0.67	7
Per-field detail	Supplementary Table S1c

Table 3. Anchoring tiers

tier	n	denominator	pct
Whole-passage verbatim	6215	9415	66.0
First-25-character only	2229	9415	23.7
Unmatched	971	9415	10.3
Multi-citation subset	721	9415	7.7
Zero-prefix subset (single-citation cells)	50	9415	0.53
Other single-citation failures	200	9415	2.1
Null-model anchorability (descriptive base rate; hits / valid draws)	30	1883000	0.0016
Detail	Supplementary Tables S2 to S3		

Table 4. Gate definitions, thresholds and outcomes

gate	threshold	unit_and_denominator	observed	outcome	source
Gate 1 (process)	>=90% and all three conditions	diagnostic subset, n = 23 (of the 115-cell sampled frame)	15/23 (65.2%)	not passed	frozen gate specification
Gate 2 (attribution)	frozen before evaluation: >=60% holds; <40% voids; 40-60% referred to PI review; conditional on Gate 1, not applied	not applicable	n/a	not initiated	frozen gate order
Gate 3 (agreement)	Po >= 0.80	attribution judgement cells (disagreement 76 + control 39); 74/115	0.6435 (113 unique-key sensitivity: 73/113 = 0.646)	not met	frozen gate specification
Voided quality-control batch	n/a	one judgement batch	n/a	voided, sealed with hash manifest	rubric classified non-exhaustively; audited and withdrawn

Table 5. Registration-layer revisions and analytic deviations

date_2026	event	type
09-12	Amendment 1 (recovery searches)	registered revision
09-13 to 09-20	Update 2 (seed definition and internal QC statement)	registered disclosure
09-20	Update 3 (registered the α ≥ 0.80 reliability floor)	registered disclosure
09-21	Screening and eligibility revision; Update 4 (pre-screening)	registered revision
09-22	Amendment 2	registered revision
09-23	Amendment 3 and Amendment 4	registered revision
09-25	Latest update effective 17:28:11 (+08), osf.io/xm6kq (update to osf.io/mhjtz)	post hoc disclosure
09-25 to 09-26	Two changes of analytic axis; principal measure set after data were visible; α floor relaxed from the registered 0.80 to 0.67	post hoc
09-26	Tool defect identified in the judgement rubric; batch voided and sealed	post hoc
09-27	Update registered before the full-text re-analysis of the 971 unmatched citation judgments (Supplementary Table S7)	registered update

Figures

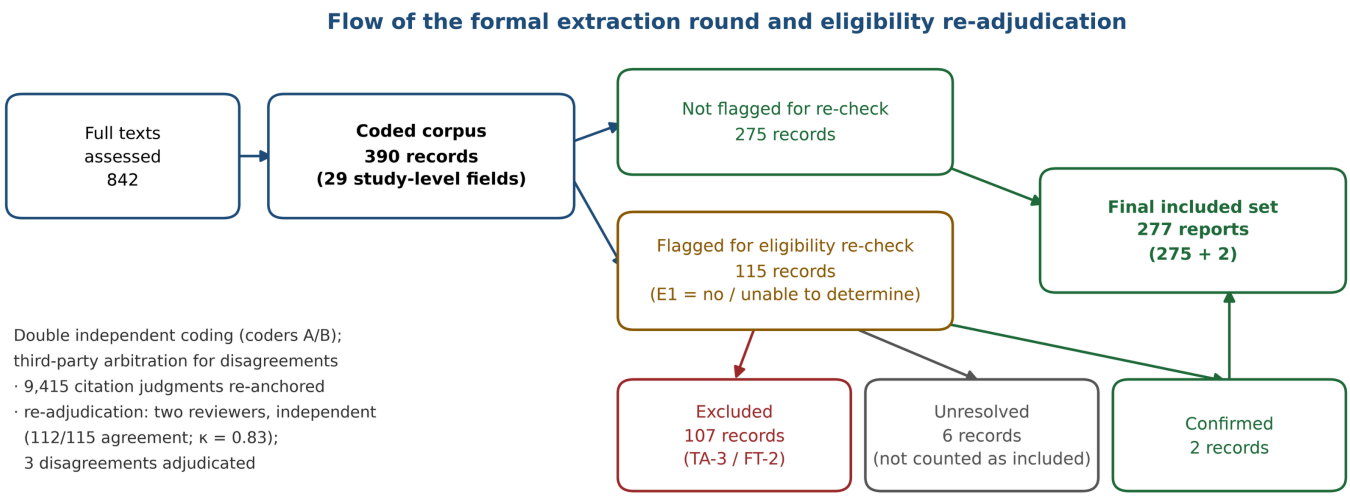


Figure 1

Flow of the formal extraction round and eligibility re-adjudication. 842 full texts assessed; 390 coded-corpus records, of which 275 were not flagged and 115 were re-adjudicated under the existing visual-channel rule (107 excluded, 6 unresolved, 2 confirmed); 275 coded records were not re-verified; 277 included reports (275 + 2). 29 study-level fields were coded by two independent coders with third-party arbitration, and 9,415 citation judgments were audited.

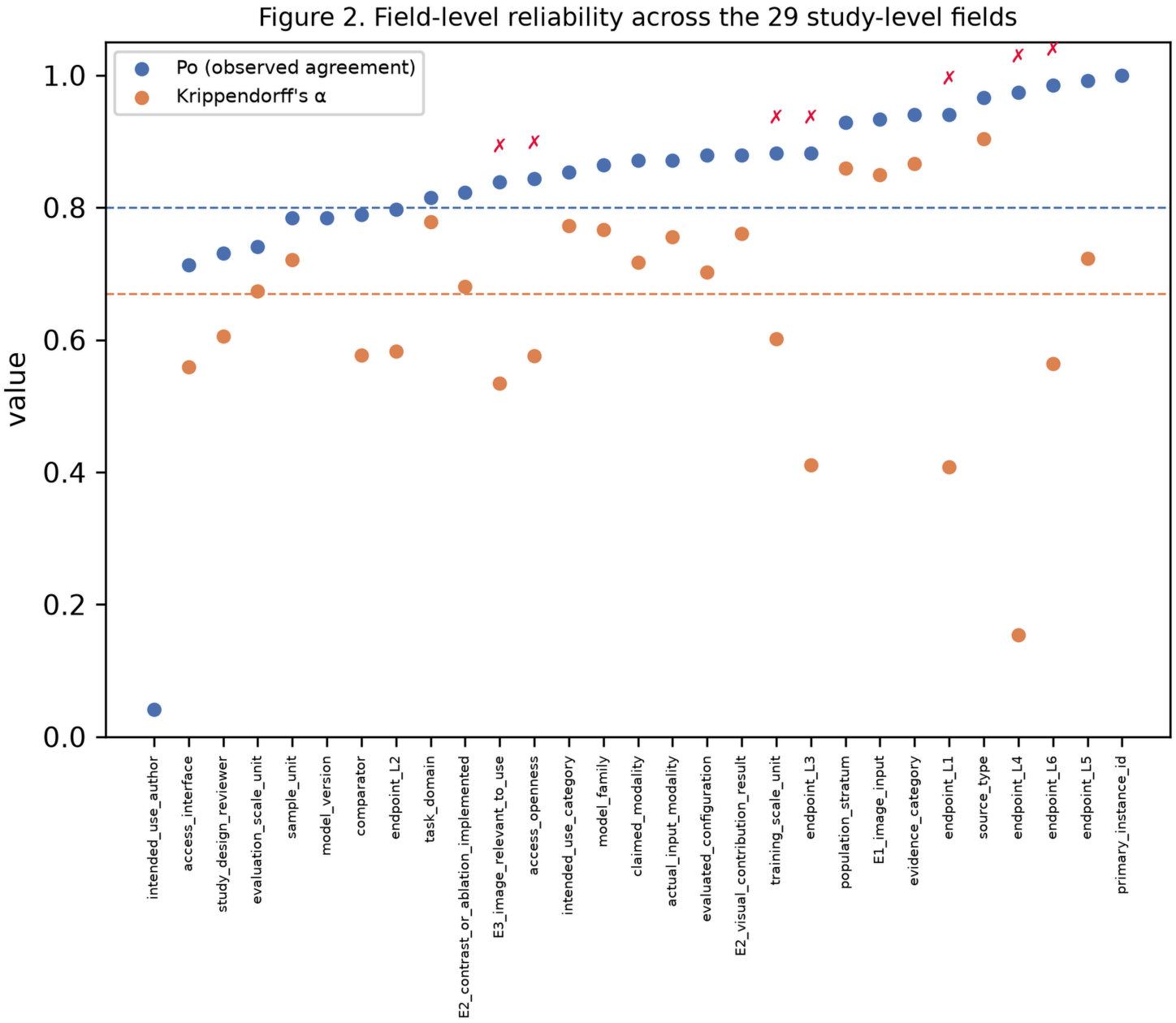


Figure 2

Field-level reliability across the 29 study-level fields. Based on the pre-adjudication 390-record coded corpus. The denominator is the 390-record coded corpus, not the 277-report post-adjudication included set. Observed agreement (Po) and, where computed, Krippendorff's alpha per field. Eight fields fell below the Po floor of 0.80. Seven fields cleared the Po floor while failing the alpha floor (three nominal fields and four binary endpoint indicators). Three free-text transcription fields (intended_use_author, model_version, primary_instance_id) carry exact-match Po only, without a chance-corrected coefficient.

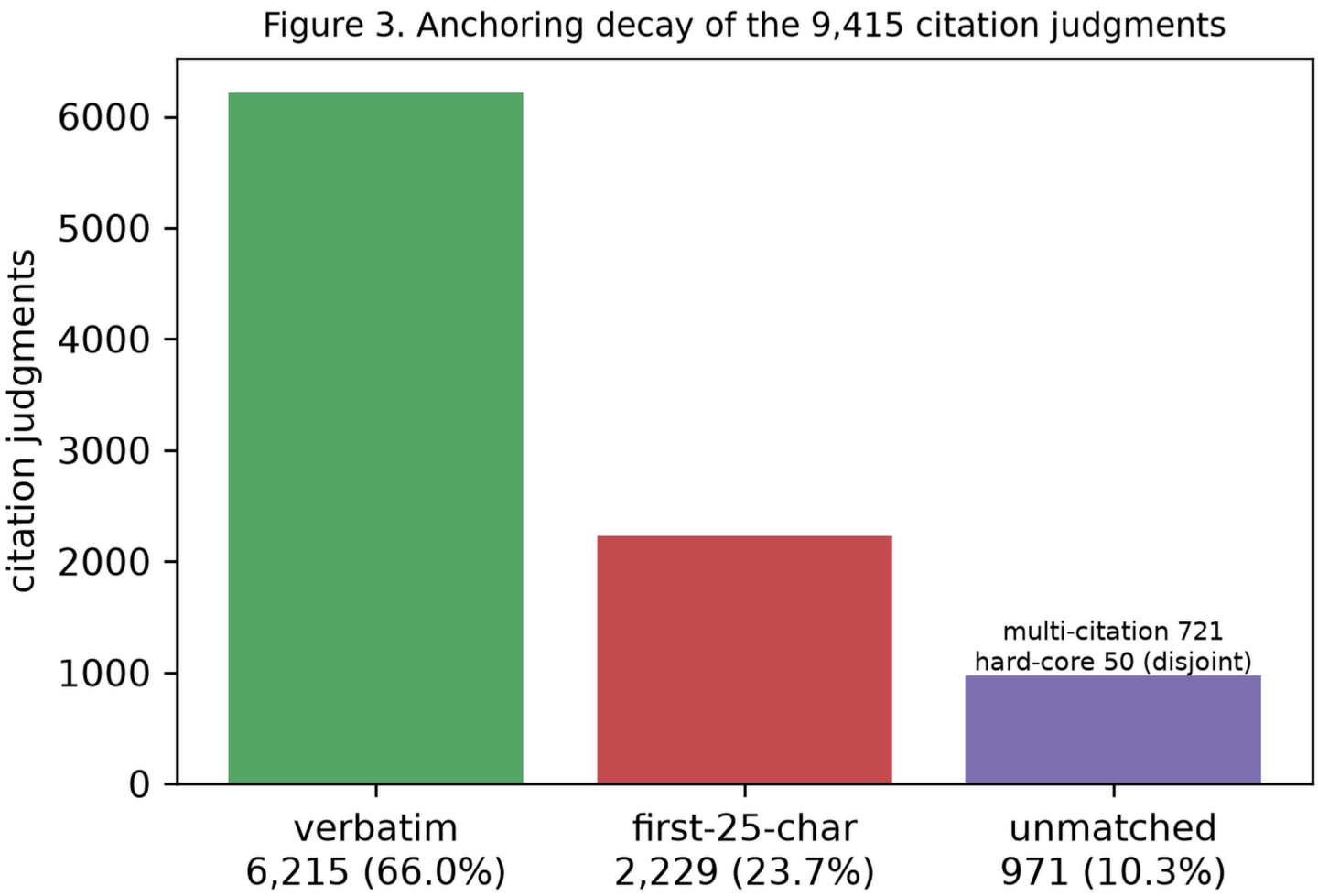
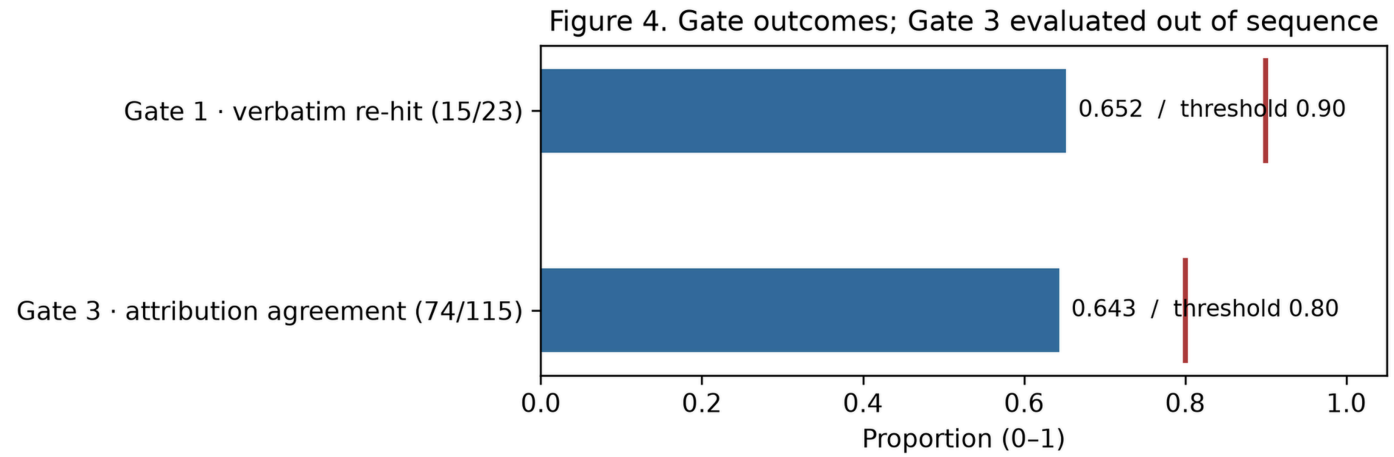


Figure 3

Anchoring decay of the 9,415 citation judgments. Three mutually exclusive tiers: whole-passage verbatim 6,215 (66.0%), first-25-character only 2,229 (23.7%) and unmatched 971 (10.3%). Within the unmatched tier, 721 arose in multi-citation cells, 50 in single-citation cells were zero-prefix (no match even for an eight-character prefix), and 200 were other single-citation failures. Unmatched means failure of the strict string rule; locating the cited text does not establish that it supports the extracted value.



Gate 2 not initiated. Gate 1 co-conditions were not scored. The gates have different denominators.

Figure 4

Frozen gate topology and outcomes. Gate 1's diagnostic subset (23 cells drawn from the 115-row attribution ledger) required a verbatim re-hit rate of at least 90% (observed 15/23, 65.2%); only verbatim retrieval was scored, not the three co-conditions. Gate 2 was contingent on Gate 1 and was not initiated because Gate 1 did not pass; Gate 3 was evaluated out of the frozen sequence, using the 115-row attribution ledger (76 disagreement plus 39 control), and returned Po 0.6435. The two gate-frame 115-row attribution-ledger labels name one frame read by different gates: the frame from which Gate 1 drew its 23-cell diagnostic subset, and the 76-disagreement-plus-39-control set on which Gate 3 tested agreement.

Figure 5. Marginal concentration of the three nominal fields that cleared Po while missing alpha

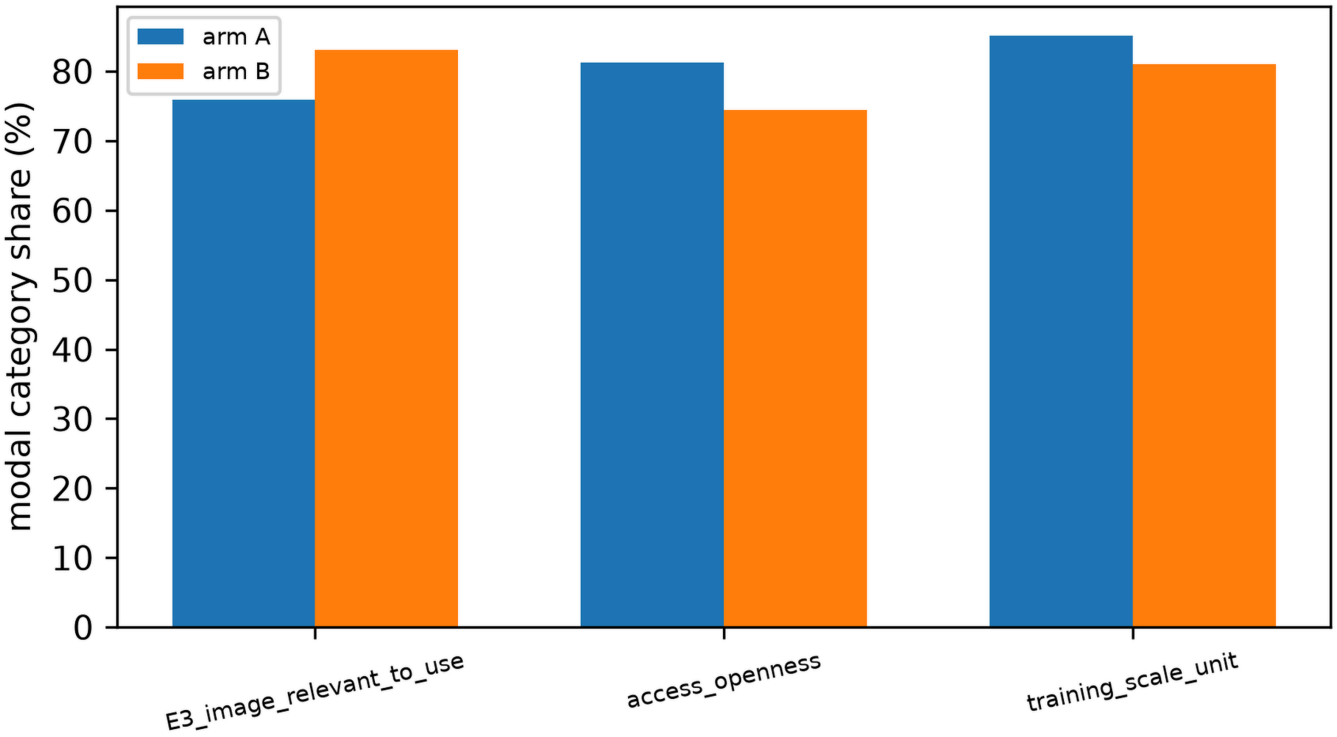


Figure 5

Marginal concentration of the three nominal fields that cleared Po while missing alpha. Based on the pre-adjudication 390-record coded corpus. The denominator is the 390-record coded corpus, not the 277-report post-adjudication included set. Share of the modal category per field and per coder arm, computed mechanically from the frozen coded data; the reported Pc and Po minus Pc are post hoc descriptive statistics, not the chance model inside α.

Supplementary Files

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

- SupplementaryMaterial.docx
- SupplementaryCodebook.docx
- SupplementaryCorpusList.csv
- PRISMA2020Checklist.docx