

Measuring coverage of cited fields by author publication histories

hana kim

hanakim@kaist.ac.kr

Korea Advanced Institute of Science and Technology

Research Article

Keywords: reference coverage, expertise diversity, research teams, publication histories, bibliometric measurement, artificial intelligence

Posted Date: October 19th, 2026

DOI: <https://doi.org/10.21203/rs.3.rs-11069371/v1>

License:   This work is licensed under a Creative Commons Attribution 4.0 International License.
[Read Full License](#)

Additional Declarations: No competing interests reported.

Measuring coverage of cited fields by author publication histories

Hana Kim

School of Business and Technology Management (BTM), KAIST
Daejeon, Republic of Korea
Corresponding author: Hana Kim; email: hanakim@kaist.ac.kr

Abstract

Reference diversity describes what a paper cites but not how that literature relates to its authors' records. We develop a reference-weighted decomposition into maximum individual coverage, additional coauthor coverage, and an uncovered remainder. Exploratory diagnostics use 355 Korean grant-linked papers. In discovery, the partial correlation between diversity and preproject uncovered share is .525 after author- and reference-count adjustment. Adding diversity to models that also control history volume raises R^2 from .328 to .502 in discovery and from .196 to .397 in first confirmation. Within-paper comparisons hold histories, cited subfields, diversity, and broader-field citation totals fixed while reallocating citation counts. Actual uncovered shares are 8.66 and 7.60 percentage points below the respective benchmarks. After excluding references in each paper's own primary subfield, differences remain at 4.27 points (95% CI 2.45–6.21; $n=114$) and 3.26 points (2.28–4.31; $n=238$). Other teams' citation frequencies explain much of the full-portfolio contrast. Aggregation also matters: additional coauthor coverage of artificial intelligence (AI) references averages 10.89% per paper but 3.45% per reference occurrence in first-confirmation AI-use papers. A prospective second confirmation of 194 papers does not confirm an AI-use association: adjusted differences are +0.455 points with parent-field controls (95% CI –6.831 to 7.742) and +0.517 with detailed-field controls (–7.395 to 8.428). Coverage identifies how citation emphasis relates to documented histories; it does not establish unfamiliarity or learning.

Keywords: reference coverage; expertise diversity; research teams; publication histories; bibliometric measurement; artificial intelligence

1 Introduction

A paper's references connect it to an observable body of prior research. Bibliometric studies use those references to characterize disciplinary diversity, combinations of knowledge, and the age of the literature on which new work draws. A different question concerns the relationship between the cited literature and the authors who use it. The same reference distribution can be well represented in one team's publication histories and largely absent from another's. Measuring that relationship requires a reference point specific to the authors, as well as a description of the references themselves.

This distinction matters for interpreting team research. Increasing specialization can make collaboration important for assembling the expertise needed to produce a paper (Jones, 2009; Wuchty et al., 2007). Yet a diverse team need not provide distinct coverage of the literature cited in a

particular output. Conversely, similar author profiles can contain different areas that are relevant to that output. Research on interdisciplinary teams and divisions of labor therefore motivates a more specific measurement question: how much of the cited reference distribution is represented in individual publication histories, and how much additional coverage appears when those histories are combined (Luncheon et al., 2014; Haeussler & Sauermann, 2020)?

For a focal paper, our reference-weighted decomposition defines M as the share covered by the most covering individual history, D as the additional share covered by combining the coauthors' histories, and E as the share outside every author's observed history. These quantities partition the same classified references. Holding those references fixed and comparing five-year preproject and prepublication windows then separates baseline uncovered share from subsequent net coverage change. The framework answers an output-specific question: which cited fields are represented in an individual record, only in the wider team, or in neither?

Publication histories have been used to characterize expertise (Luncheon et al., 2023; Salatino et al., 2026), and collaboration measures have been related to interdisciplinary output (Nakhoda et al., 2025). We connect these histories to the references of a particular paper, measuring individual and additional team coverage on the same denominator. Weighting the cited categories also distinguishes a paper that draws on uncovered fields at the margins from one that places those fields at the center of its bibliography.

AI-enabled research offers a first application because debates about AI concern both changes in scientific production and the knowledge that supports it (Wang et al., 2023; Gao & Wang, 2024; Hao et al., 2026). Greater AI use could accompany different reference choices, different team compositions, or both. Neither a higher AI-reference share nor higher research impact establishes that cited fields lie outside the authors' prior records. The application links Korean research grants to focal publications and compares projects classified as substantively using AI with sampled non-AI projects. It retains a discovery stage, an initial confirmation stage, and a second confirmation with its own prospectively fixed primary outcome.

The AI application follows its original inferential sequence. The first confirmation failed its primary test of positive net coverage change. Preproject uncovered share became the focus after those results were examined and was designated primary before the second-confirmation outcomes were observed. In that new sample, AI-use coefficients are near zero under both parent-field and detailed-field adjustment. The intervals allow negative and positive differences; the results neither confirm the proposed association nor establish equivalence.

Two comparisons establish what the measure records. Reference diversity remains associated with uncovered share after adjustment for history volume and author and reference counts. Within-paper reallocations then hold histories and diversity fixed to identify how citation weight is distributed between covered and uncovered fields. Actual uncovered shares fall below the structural benchmarks in both earlier stages, including after references in the focal paper's own primary subfield are removed. Other teams' citation frequencies account for much of the full-portfolio difference. These exploratory results show what reference weighting adds to category overlap, while leaving claims about actual expertise open.

2 Cited knowledge and documented expertise

2.1 Individual histories and collective coverage

Salatino et al. (2026) provide the closest publication-history comparison, relating expertise diversity to citation impact across 944,508 computer-science papers. Their between-author distances characterize how collaborators differ. Our framework evaluates each history against the fields cited by a particular output. A large distance can arise in fields the focal paper never cites; output-specific coverage changes only when the observed histories intersect its cited fields.

This distinction also follows the separation of author and reference-list diversity in Abramo et al. (2018). Their comparison establishes that these are related but noninterchangeable empirical objects. Nakhoda et al. (2025), in an institutional study, further distinguish collaboration characteristics from interdisciplinary output. Their reported results motivate attention to prior knowledge profiles rather than author counts alone. The present contribution links the objects at the reference-category level and partitions their intersection into individual coverage, additional team coverage, and an uncovered remainder.

The growth of team science and specialization motivates this question (Wuchty et al., 2007; Jones, 2009). Research on interdisciplinary team assembly and text-based expertise profiles supplies relevant antecedents (Lungeanu et al., 2014, 2023). Team composition nevertheless differs from division of labor and from novelty or impact (Haeussler & Sauermann, 2020; Lee et al., 2015). The maximum-covering author is a reproducible benchmark, not an identified project leader or principal intellectual contributor. Coauthors may perform essential work while adding little indexed field coverage beyond this benchmark. D therefore describes a marginal reference share rather than collaborators' labor or value.

Team assembly models distinguish team size, newcomers, and repeated collaboration (Guimerà et al., 2005; Milojević, 2014). Zeng et al. (2022) compare collaborators' participation across topics with controlled surrogate data and find participation concentrated in fewer topics. This motivates an explicit benchmark for observed overlap; their collaborator-topic statistic differs from the focal-reference marginal coverage D used here. Zeng et al. (2021) also distinguish the freshness of collaboration within a team. That concept differs from excluding authors already present in an earlier analysis sample.

2.2 Diversity, coherence, and output-specific coverage

The measurement literature cautions against treating all indicators of interdisciplinarity as interchangeable. Leydesdorff and Rafols (2011) use correlations and factor analysis to examine relationships among diversity, centrality, and citation indicators. Rafols and Meyer (2010) distinguish diversity from network coherence, and Du et al. (2025) study integration and segregation through concept networks in infectious-disease research. Coverage is neither a network-coherence statistic nor a direct measure of knowledge integration.

Stirling (2007) distinguishes variety, balance, and disparity in diversity measurement. Shannon and Gini–Simpson indices describe category frequencies without an independently estimated distance between categories. Similarity-aware reference diversity makes this additional choice explicit (Zhang

et al., 2016), and debate over diversity measures addresses their scaling and mathematical properties (Leydesdorff et al., 2019). Matching Shannon therefore does not match every dimension of a reference portfolio.

Reference diversity describes the distribution of a paper's citations across categories, while atypical-combination measures assess how unusual combinations are relative to a wider literature (Wagner et al., 2011; Uzzi et al., 2013). These approaches answer questions about the cited portfolio and its position in science. A portfolio can be diverse or atypical without extending beyond the authors' documented areas. A concentrated portfolio can also be poorly represented in those areas. Coverage is therefore defined relative to an author-specific record rather than a population distribution of combinations.

Consistency between a construct and its operationalization is a separate requirement from obtaining a new numerical index (Zwanenburg et al., 2022). Cantone (2024) makes the measurement design and unit of analysis explicit. Here the intended object is the intersection of cited categories with time-bounded publication records; it is not latent familiarity. Comparisons with scientists' perceptions, such as Ávila-Robinson et al. (2021), illustrate the kind of external criterion that bibliographic accounting alone does not supply.

Reference age adds a temporal description of cited knowledge (Yang, 2024). Older references can lie inside or outside recent publication areas. Age contrasts describe that portfolio but require an external criterion to support claims about familiarity, since citations record only part of the cognitive influence of earlier work (Tahamtan & Bornmann, 2018).

2.3 AI research as an application

AI methods span research activities and fields (Bianchini et al., 2022; Wang et al., 2023). Studies of their diffusion, scientific benefits, and association with impact or focus motivate attention to disciplinary composition (Duede et al., 2024; Gao & Wang, 2024; Bianchini et al., 2026; Hao et al., 2026). The coverage framework asks how the cited fields in such outputs relate to the eventual authors' preceding publication records.

The present application separates three observable features: whether a project is classified as using AI, whether its focal paper cites AI-classified literature, and whether that literature is represented in the authors' publication histories. These are related but different variables. In particular, a conditional decomposition of AI references can identify which observed histories cover those references. It cannot establish who introduced AI to a collaboration or who performed an AI-related task. The distinction between producing research outputs and understanding the underlying work is especially relevant here (Messerli & Crockett, 2024).

Sourati and Evans (2023) incorporate researchers' activities into AI models of scientific discovery. Their work motivates attention to human research records alongside algorithmic capabilities. Our accounting uses those records as a descriptive reference point; testing a discovery model or observing expertise used in a project requires different evidence.

2.4 Coverage across observation windows

Prior related knowledge is central to accounts of absorptive capacity (Cohen & Levinthal, 1990), but an indexed publication history is only a partial trace of that knowledge. The preproject measure asks how much of a later paper's cited portfolio is outside the eventual authors' records before the project starts. The prepublication measure asks the corresponding question using a more recent window. Holding the references fixed makes the difference interpretable as a change in recorded coverage.

Because both windows contain five complete calendar years, the comparison includes entry and expiry. A field can enter the recent publication record without being new to an author over their lifetime; it can leave that record without being forgotten. Net change is thus decomposed into reference-weighted gains and losses in window-specific coverage. This accounting supports a temporal description of the record while leaving learning and knowledge acquisition as separate empirical questions.

Research strategies can be defined relative to established knowledge and new relationships (Foster et al., 2015), while individual publication records also exhibit topic switching over time (Zeng et al., 2019). These studies motivate separating a changing record from a lifetime stock of expertise. They do not establish five years as a uniquely appropriate observation window or equate a newly covered category with learning.

3 Data and methods

3.1 Research frame and inferential sequence

The data combine Korean National Science and Technology Information Service (NTIS) records with National Research Foundation (NRF) award acknowledgments and OpenAlex publications (Priem et al., 2022). The conservatively linked frame contains 17,959 grants: 1,266 classified as AI use, 16,376 as non-AI, and 317 as AI development. Project-content classifications are inherited from the earlier frame construction. They do not constitute a new validation of actual AI use or its intensity. The conservative frame uses a first retained project year equal to the project-start year and stable observed organization identifiers. These restrictions define a selected linked population.

The paired-history analyses in the earlier discovery and confirmation stages contain 115 and 240 observations. Their pooled sample of 355 observations in 151 organizations is a post-result synthesis. The original confirmation protocol designated net coverage change A as primary. Preproject uncovered share was a component of that accounting, but its later prominence followed inspection of the earlier results. The second confirmation protocol fixed preproject uncovered share as primary on 14 September 2026 before its new outcomes were examined. This is an internally dated protocol, not a public preregistration.

The second confirmation selected all 416 remaining AI-use grants in the linked frame and 416 non-AI grants. Controls were allocated across baseline-year-by-stored-field cells using largest-remainder allocation and a fixed deterministic order. Rounding left 26 cells containing 235 grants with zero selection probability. Completing selection of the linked AI-use frame is distinct from observing outcomes for every selected grant. No grant or focal publication was replaced in response to the

estimates or to a failed quality screen. Appendix Tables A1–A4 preserve the stage-specific counts and their units.

3.2 Focal articles and publication histories

For each grant, the focal publication is the earliest nonretracted, DOI-bearing journal article one to three calendar years after the baseline year and no later than 2024. Publication date determines priority, with OpenAlex work identifier resolving ties. A first article that subsequently fails an attribution or quality requirement is not replaced by a later one. Attribution requires a single complete normalized NRF code matching the grant and no unparseable NRF code; acknowledgments of other funders are permitted. A focal work already selected in an earlier stage is excluded by work identifier or canonical DOI.

Funding-acknowledgment coverage and interpretation vary, and extracted records can disagree with manual coding (Paul-Hus et al., 2016; Grassano et al., 2017). The complete-code rule conservatively assigns one focal output to one grant. Ambiguous cases can still be genuinely connected to the grant, so attribution remains a source of selection.

Histories are constructed for the eventual focal authors. The baseline window contains the five complete years before the project-start year; the publication window contains the five complete years before the focal publication year. Eligible histories include nonretracted articles, conference papers, reviews, preprints, and book chapters. Works are deduplicated by canonical DOI within a window, retaining the smallest numeric OpenAlex identifier; works without a DOI remain distinct by identifier. References with missing publication year, a year later than the focal article, or a retraction flag are omitted from the eligible reference set. OpenAlex primary topics supply the field assignments observed at collection, rather than historical taxonomic snapshots.

The fixed quality rules require 1–15 identified authors without duplicate author identifiers, at least ten classified distinct references, and classified-reference coverage of at least 70% of listed reference identifiers. All required history queries must be complete. At least 80% of authors must have two classified publications in the relevant window, without a detected identity conflict within that window. The primary second-confirmation analysis uses the baseline-history requirement alone. A and the paired outcome matrix require both windows. Invalid baseline budget or headcount records are excluded. Prior author overlap is retained and flagged in the new primary sample; a separate sensitivity excludes it.

Author-level histories depend on name disambiguation, whose evaluation requires labeled identity information (Tekles & Bornmann, 2020; Kim & Owen-Smith, 2021). The conflict and duplicate-ID exclusions detect specific inconsistencies; they do not constitute a validated identity-error rate for this sample. Culbert et al. (2025) document reference-coverage and metadata differences between OpenAlex, Web of Science, and Scopus. This supports explicit coverage assessment, but the 70% threshold is our operational criterion. Its denominator is the reference identifiers listed by OpenAlex, so it does not measure completeness relative to the article’s full bibliography.

3.3 Reference coverage and its decomposition

For paper i , let q_{ik} denote the fraction of its eligible classified references assigned to category k . Let H_{iat} contain the categories meeting the publication threshold for author a in window t , and let U_{it} be the union across its authors. At the primary threshold, one eligible publication by an individual author is sufficient to cover a category. The uncovered share E_{it} is the sum of q_{ik} over categories outside U_{it} . E_{pre} and E_{pub} apply the preproject and prepublication windows to the same focal reference distribution.

$$E_{it} = \sum_k q_{ik} 1\{k \notin U_{it}\}$$

Net coverage change is $A = E_{pre} - E_{pub}$. Gross gain G is the reference share in U_{pub} but outside U_{pre} ; gross loss L is the share in U_{pre} but outside U_{pub} . Accordingly, $A = G - L$. For the prepublication decomposition, M is the largest individual reference coverage. Ties are resolved by author identifier. D is team-union coverage minus M , yielding $M + D + E_{pub} = 1$. Total coauthor coverage includes both D and the share also covered by the anchor. The reported shared ratio divides mean shared coverage by mean total coauthor coverage; it is a ratio of project means.

$$A_i = E_{i,pre} - E_{i,pub} = G_i - L_i$$

$$M_i + D_i + E_{i,pub} = 1$$

AI references are assigned to subfields 1702 or 1707. Their coverage decomposition is conditional on at least one such reference in the focal paper. Baseline AI history indicates an eligible AI-field publication by a focal author in the preproject window; prepublication AI history uses the later window. Under the two-work threshold, a category is covered only if an individual author has two eligible publications in it. One publication from each of two authors is insufficient. Appendix Tables A5 and A6 define all variables.

Holding the reference distribution and taxonomy fixed, adding an author to the roster can only enlarge the union of covered fields. E therefore cannot increase. Duplicate coverage leaves the union unchanged. The measure is not normalized for the opportunity that a larger team has to cover a field, and NTIS headcount in the application is distinct from the observed paper-author count. For a fixed roster, nested thresholds that require more within-author publications cannot enlarge coverage. Taxonomy changes need not behave like such nested thresholds.

Observed history size is a separate source of coverage opportunity. We count the DOI-deduplicated union of eligible works by the focal authors in the five preproject years, including eligible works without a subfield label. The accompanying field count is the union of classified subfields in those authors' deduplicated histories, following the coverage engine. Holding the roster, references, and taxonomy fixed, adding history records cannot increase E_{pre} . A later descriptive diagnostic therefore reports history size as well as author count; its timing and caliper are specified in Section 3.6.

A constructed example isolates the distinction while holding size fixed. Two authors each have one prior publication in one field, and the focal paper cites X and Y equally. Histories $\{X\}$ and $\{Y\}$ yield $M=.5$, $D=.5$, $E=0$; histories $\{X\}$ and $\{Z\}$ yield $M=.5$, $D=0$, $E=.5$. Author count, per-author and team history counts, history-field counts, and maximum binary profile distance are identical, as is reference Shannon diversity $\ln(2)$. Only the alignment of recorded fields with the cited portfolio

changes. This establishes nonidentity of the measures under the construction; empirical validity requires further evidence (Appendix Table A7).

3.4 Models and fixed decision rules

The primary model regresses E_{pre} on AI-use status, parent-field and baseline-year indicators, $\log(1 + \text{baseline budget})$, and $\log(1 + \text{baseline NTIS team headcount})$. Recorded headcount differs from the focal article's author count. Estimates use unweighted ordinary least squares (OLS), organization-cluster standard errors with a finite-sample correction (CR1), and t inference with one fewer degree of freedom than the number of organizations. Estimation requires at least 30 observations per group, 30 organizations, and a full-rank design matrix.

A positive primary coefficient with a two-sided p-value below .05 is required to activate a subsequent test of positive A at the same level. If that criterion fails, A is reported as a descriptive estimate. Other outcomes cannot replace the primary endpoint. A dated amendment, fixed before the new outcomes were computed, substitutes detailed-field indicators for parent-field indicators while preserving the remaining covariates, history gates, and inference. Detailed fields use the first two characters of the stored NTIS code, with missing fields assigned to an explicit unknown category. This amendment is a sensitivity analysis and does not modify the sequence.

Specified sensitivities address the paired-history gate, prior author and organization overlap, publication thresholds, classification resolution, discordant DOI groups, common field-year support, and inverse allocation weights. Weights are restricted to cells with positive selection probability and do not correct outcome-observation probabilities. Organization influence is assessed by deleting each organization in turn; the primary model also has a restricted-null wild cluster bootstrap with 9,999 draws. Coverage-change checks address reference-matching histories, conference papers, dates, initial coverage, publication lag, and intervening output. Nonengineering and baseline AI-history interactions remain descriptive. Both field-adjustment estimates for E_{pre} are reported in the abstract irrespective of significance.

After the second-confirmation outcomes were known, an additional retention sensitivity modeled arrival in the primary sample among the 832 selected grants. Logistic regression used AI-use status, parent field, baseline year, log budget, and log NTIS headcount. Seven grants lacked required covariates, leaving 825. Retained observations received the inverse fitted retention probability, without trimming, in the original outcome specification. Organization-cluster CR1 intervals are conditional on these fitted weights and do not propagate propensity-estimation uncertainty. This is distinct from the frozen inverse-allocation sensitivity and cannot correct unobserved selection or supply well-defined outcomes for grants without an attributable focal paper.

After observing the second-confirmation results, we assessed precision using the primary standard error and 96 degrees of freedom. A two-sided noncentral-t calculation gives the difference detectable with 80% power at the .05 level and power against the earlier confirmation estimate. These calculations use the observed precision and an uncertain earlier benchmark; they are not prospective power estimates. Appendix Table A8 reports both field specifications.

3.5 Scope of interpretation

E measures cited fields outside indexed publication histories. It does not measure semantic distance, familiarity, latent ability, understanding, or learning. Missing indexed evidence is not verified absence of expertise, and citations are incomplete records of intellectual influence (Tahamtan & Bornmann, 2018). The ex-post maximum gives the anchor a definitional advantage. D consequently describes additional coverage beyond that benchmark, rather than the total knowledge or contribution of the remaining authors.

Language coverage warrants particular attention in this Korean sample. Céspedes et al. (2025) find broader linguistic coverage in OpenAlex than in Web of Science, but also language-label errors that overstate English. A mislabeled language is not itself a missing publication, and this study does not filter histories by language. Nevertheless, Korean-language or domestic-journal works missing from indexed histories could reduce individual and union coverage. With the reference distribution, roster, and taxonomy held fixed, such omissions can depress M and inflate E; D has no fixed directional bias. If references are also missing, the net direction is undetermined. We have not measured Korean-specific history undercoverage in this sample.

Recorded history volume constrains the interpretation of coverage. In discovery, mean E_{pre} falls from 35.60% to 8.13% across history-work quartiles (Appendix Table A21). This gradient can reflect both the volume of the indexed record and its alignment with the cited fields. Section 4.1 examines these relationships through size-adjusted regressions and a separate matched-case search.

The AI comparisons are observational. Project labels do not verify use intensity, budgets are recorded amounts, and eventual authorship does not identify the project-start roster. Authorship and acknowledgments capture different aspects of collaboration (Paul-Hus et al., 2017). Linkage, attribution, and history requirements select the analyzed observations. Complete retrieval and inverse-allocation weights address specific observation processes, without making this sample representative of all NTIS projects.

The earlier pooled estimates remain a post-result synthesis. The second-confirmation primary model and amendment are prospective with respect to its new outcomes only. Nonsignificant estimates neither establish equivalent groups nor rule out coverage gains for individual teams.

3.6 Exploratory measurement diagnostics

All measurement diagnostics are exploratory. The initial comparisons followed examination of discovery and first-confirmation outcomes but preceded the second-confirmation results. The matching extension to 355 papers, history-size adjustment, organization-exclusion diagnostics, and within-paper citation-emphasis analysis were added on 16 September 2026, after all three stages were known. The citation-emphasis specification was recorded before its diagnostic values were computed; shared-author resampling was added subsequently. The primary-subfield exclusion was added on 17 September 2026. These dated records are not preregistrations, and none changes the prospective second-confirmation test.

The supplied replacement exercise is first reproduced unchanged. It holds the focal references, observed anchor, and team size fixed while drawing replacement author appearances with similar

classified publication counts. A revised benchmark then uses unique donor identities, reconstructs histories in the focal paper’s window, excludes every focal-team member, prohibits duplicates within a simulated team, and retains the $\pm 30\%$ publication-count caliper. Each paper must admit 200 valid draws. On each draw, D is evaluated both relative to the original anchor and after reselecting the maximum-covering member. The latter applies the observed statistic symmetrically. Mean observed-minus-benchmark differences use organization-cluster CR1 inference. This is calibration against a specified donor mechanism, not an exact randomization test.

Neighboring-measure comparisons include reference-field Shannon entropy and Gini–Simpson diversity, reference-age descriptors, and mean and maximum author-profile cosine distances. An additional binary top-ten-topic analogue uses OpenAlex profiles; it does not reproduce Salatino et al.’s (2026) background-adjusted Computer Science Ontology (CSO) construction. Age descriptors address Yang’s (2024) temporal-diversity question without claiming replication of an unverified formula. Rao–Stirling diversity requires a choice of category dissimilarities (Stirling, 2007); it is not estimated because an independent field-dissimilarity matrix is unavailable. Pearson and Spearman correlations are accompanied by 1,999-draw organization-bootstrap intervals.

Reference-level diagnostics first compare covered and uncovered references within papers containing both types. The resulting age and review-share differences are averaged across papers with organization-resampling intervals. An internal follow-up measures how much baseline-uncovered reference weight enters the intervening author histories, with and without the cited works themselves. Because this follow-up uses the same authors, categories, and histories as the coverage measures, it cannot independently validate familiarity or learning.

The conditional AI-reference decomposition is reported with two aggregation rules. For paper i with n_i AI references and a component share c_i , the paper average is the unweighted mean of c_i ; the reference-occurrence estimate is the sum of $n_i c_i$ divided by the sum of n_i . Each reference is deduplicated within its focal paper, but a work cited by two focal papers contributes two occurrences. The reference-occurrence estimate minus the paper average equals the population-form covariance of n_i and c_i divided by mean n_i . Thus a gap between the summaries is not itself a calculation error: it identifies a change of estimand. Both are retained, together with denominator strata.

The original discovery comparison residualizes E_{pre} and Shannon on author count and log classified-reference count. We extend it by adding $\log(1+\text{history works})$, using the same four nested OLS specifications in both earlier stages and adding a stage indicator when pooling. Organization-bootstrap percentile intervals use 1,999 draws; coefficient inference uses CR1 and $t(G-1)$. Leave-one-organization-out prediction refits each specification after removing every paper from the held-out organization, with no tuning or clipping. Its root mean squared error (RMSE) evaluates prediction of recorded E_{pre} . This resampling does not guarantee separation of authors affiliated with different organizations and is not an independent validation sample. Appendix Table A18 reports all specifications and prediction-range diagnostics.

The case search separately enumerates every unordered pair among the 115 discovery and 240 first-confirmation papers. It requires the same stored NTIS field, publication year, and author count, a classified-reference-count ratio ≤ 1.20 , and an absolute Shannon difference $\leq .10$. All eligible pairs are retained, regardless of E_{pre} gap or AI status. The later history-work-count ratio ≤ 1.5 diagnoses these

same pairs without replacement or relaxed rules. This matched-case search and the regressions answer different descriptive questions.

3.7 Citation emphasis conditional on histories and diversity

To assess what reference weighting adds, we compare each of the 355 papers with a constrained reallocation of its citation counts. Authors, history records, covered categories, and cited subfields remain fixed. Within each OpenAlex field, we permute the positive reference counts across the subfields that the paper cites. This preserves total references, Shannon and Gini–Simpson diversity, broader-field citation totals, and every author-profile comparison. Only the allocation of reference weight to the occupied subfields changes. OpenAlex fields differ from the NTIS project-field controls used in the AI application.

For paper i and occupied field g , let Q_{ig} be the total reference count, K_{ig} the number of cited subfields, U_{ig} the number of those subfields outside all preproject author histories, and N_i the paper’s total classified references. The mean uncovered share over these reallocations and its difference from the observed share are:

$$\mu_i = \sum_g \frac{Q_{ig} U_{ig}}{N_i K_{ig}}; \quad \Delta_i = \mu_i - E_i$$

A positive Δ means that actual citations place less weight on uncovered subfields than the field-conditioned mean. We calculate this mean, the attainable range, and the permutation variance exactly. Papers whose uncovered share cannot vary contribute $\Delta=0$. The comparison describes the observed category counts; it does not assume uniform citation choice or generate substitute bibliographies. Because the result depends on the benchmark (Cantone, 2024), the permutations are not used as a randomization test.

A separate frequency benchmark distributes each field’s reference mass according to citations in other papers from the same stage, excluding the focal organization and every paper sharing a focal author. We restrict the calculation to the focal paper’s occupied subfields and add 0.5 to each category count before normalization. Histories, occupied categories, and field totals remain fixed, but this benchmark does not preserve Shannon. It tests sensitivity to ordinary differences in subfield citation frequency. Further checks use domain strata, topics within subfields, prepublication histories, a two-work per-author coverage threshold, and removal of references present in the preproject histories. All specifications and both stages are reported. Intervals use 4,999 whole-organization resamples; connected organization–author components provide an additional dependence sensitivity (Appendix Table A19).

We also remove every reference assigned to the focal paper’s own primary subfield. This tests whether the structural contrast is confined to overlap in the paper’s main subject. Histories and coverage flags remain unchanged, while reference weights and field-conditioned benchmarks are recomputed on the remaining bibliography. Papers with no remaining classified references are excluded from this diagnostic. This exploratory check was added after all stage outcomes and the original structural diagnostics were known.

4 Results

4.1 Coverage and neighboring measures

The measurement analyses use 355 papers with reconstructed reference lists and publication histories. Author-distance comparisons use 325 papers in 141 organizations; 30 single-author papers have no author pair. Appendix Table A22 reports the neighboring descriptors, and Table 1 reports the size-adjusted relationships.

E_{pre} correlates with reference-subfield Shannon diversity at $r=.459$ (95% CI .379 to .542), mean author-profile distance at .235 (.127 to .333), and reference-age standard deviation at .142 (.016 to .257). The measures share information without describing the same object. The top-ten binary maximum distance equals one in 197 of 325 papers (60.6%); its ceiling limits the interpretation of weak correlations as evidence of distinctness.

Size adjustment preserves the association with reference diversity. In discovery, the partial correlation controlling author count and log reference count is .525 (new organization-bootstrap 95% CI .403 to .657), compared with raw .515. R^2 rises from .140 to .378 when Shannon is added to those two controls. Adding observed history volume to the controls leaves a partial correlation of .509 (.378 to .637); Shannon then raises R^2 from .328 to .502, an increment of .174 (.085 to .289). The corresponding first-confirmation partial correlation is .501 (.415 to .592), with R^2 increasing from .196 to .397, an increment of .202 (.133 to .285). The two stages therefore show a similar conditional association despite different baseline model fit.

Table 1 Coverage and reference diversity after size adjustment

Statistic / specification	Discovery	First confirmation
Papers / organizations	115 / 73	240 / 123
Partial correlation with Shannon diversity		
Author and reference counts adjusted	.525	.479
History volume also adjusted (95% CI)	.509 [.378, .637]	.501 [.415, .592]
Model fit with uncovered share E_{pre} as the outcome		
R ² : author and reference counts	.140	.103
R ² : counts + history volume	.328	.196
R ² : counts + history volume + Shannon	.502	.397
Shannon increment in R ² (95% CI)	.174 [.085, .289]	.202 [.133, .285]
Leave-one-organization-out prediction		
RMSE without Shannon (pp)	19.33	20.00
RMSE with Shannon (pp)	16.79	17.39

Notes. Exploratory models use E_{pre} as the outcome. Counts denote author count and log classified-reference count; history volume is $\log(1 + \text{history works})$. Brackets give 95% confidence intervals from 1,999 organization resamples. R² increments describe additional model fit. RMSE is in percentage points (pp); held-out predictions are neither tuned nor clipped. Appendix Table A18 reports all specifications, pooled models, and out-of-range predictions.

The matched-case search finds no eligible pair among the 115 discovery papers and four among all 355 papers under the same field, year, author-count, reference-count, and Shannon restrictions. Their E_{pre} gaps are 57.05, 40.63, 34.46, and 4.63 percentage points. None meets the additional history-work-count ratio of at most 1.5. The four cases remain in Appendix Table A14, but they cannot establish a coverage contrast between papers with balanced history size.

History volume also has a strong association with coverage. In discovery, $\log(1 + \text{history works})$ correlates with E_{pre} at $-.556$ and author count correlates with history works at $.568$. After history volume is controlled, the author-count coefficient is close to zero ($p=.898$; Appendix Table A17). Team size is relevant to reference characteristics (Wu et al., 2019), but these conditional associations do not identify a causal mechanism. In particular, the R² increase from $.328$ to $.502$ measures what Shannon adds when recorded E_{pre} is the outcome.

Adding Shannon also reduces leave-one-organization-out RMSE, from 19.33 to 16.79 percentage points in discovery and from 20.00 to 17.39 in first confirmation. The models are refitted without tuning. Their predictive improvement supports the conditional relationship between the bibliographic measures, although held-out organizations can share authors with the training sample. Some OLS predictions fall outside $[0,1]$; Appendix Table A18 reports them without clipping.

4.2 Citation emphasis at fixed histories and diversity

Actual references place more weight in subfields represented in the authors' records than the field-conditioned benchmark does. In discovery, mean preproject uncovered share is 21.25%, compared with 29.91% under reallocation, a difference of 8.66 percentage points (95% CI 6.22 to 11.36). First confirmation yields 18.59% versus 26.19%, a difference of 7.60 points (5.94 to 9.33; Table 2). The histories, occupied subfields, diversity, and broader-field citation totals are identical within each comparison.

Table 2 Citation emphasis and preproject coverage

Stage	N	Observed E_{pre} (%)	Benchmark E_{pre} (%)	Difference (pp)	95% CI
Field-conditioned reallocation · All classified references					
Discovery	115	21.25	29.91	+8.66	[6.22, 11.36]
First confirmation	240	18.59	26.19	+7.60	[5.94, 9.33]
Other-team frequency benchmark · All classified references					
Discovery	115	21.25	23.04	+1.79	[−0.66, 4.76]
First confirmation	240	18.59	21.44	+2.84	[1.11, 4.90]
Field-conditioned reallocation · Own primary subfield excluded					
Discovery	114	36.07	40.34	+4.27	[2.45, 6.21]
First confirmation	238	33.59	36.86	+3.26	[2.28, 4.31]

Notes. Difference = benchmark minus observed uncovered share; positive values indicate lower actual uncovered weight. Full-sample organization counts are 73 and 123. Under full-portfolio reallocation, uncovered share can vary in 77/115 and 141/240 papers, respectively. The structural benchmark preserves histories, occupied subfields, broader-field citation totals, Shannon, and Gini–Simpson. Exclusion removes the focal paper's primary-subfield references and recomputes weights and the benchmark; one discovery and two first-confirmation papers have no references left. The frequency benchmark uses other teams' citation frequencies and does not preserve entropy. Intervals use 4,999 organization resamples, rebuilding peer frequencies where applicable. All diagnostics are exploratory. Appendix Table A19 gives specifications and dependence checks.

The uncovered share can vary under the permitted reallocations in 77 of 115 discovery papers and 141 of 240 first-confirmation papers. The remaining papers contribute zero differences. Citation-count allocations can change in a larger set—111 and 222 papers—because a different allocation need not change the total weight assigned to uncovered categories. Table 2 reports the count for variable uncovered share; Appendix Table A19 separates the two concepts.

The focal paper's own primary subfield accounts for a mean 50.42% of references in discovery and 49.40% in first confirmation. That subfield appears in the preproject histories of 103 of 115 and 214 of 240 papers, respectively. After removing all references in it, actual uncovered share is 36.07% versus a 40.34% benchmark in discovery (114 papers), a difference of 4.27 points (95% CI 2.45 to 6.21). First confirmation gives 33.59% versus 36.86% (238 papers), a difference of 3.26 points (2.28 to 4.31). The contrast therefore persists outside the focal paper's main subfield. Because exclusion changes the denominator and category support, these values are not an additive decomposition of the full-portfolio difference.

The frequency benchmark reduces the difference to 1.79 points in discovery (95% CI -0.66 to 4.76) and 2.84 in first confirmation (1.11 to 4.90). Common subfield citation patterns therefore explain much of the structural contrast; an excess beyond those patterns is not established in both stages. The structural differences remain positive at topic resolution (7.69 and 8.85 points), with a two-work coverage threshold (11.50 and 11.25), and after removing references present in the preproject histories (8.82 and 7.47). Appendix Table A19 reports these exploratory sensitivities.

The examples illustrate why reference weights matter. In the water-network design paper W3156089142, 50 of 64 references belong to Civil and Structural Engineering, a subfield represented in the authors' preproject histories. Its actual uncovered share is 20.31%, even though eight of its ten cited subfields are uncovered; the field-conditioned benchmark is 76.17%. W2910755680 instead assigns 33 references to Epidemiology, absent from its recorded histories. Its actual uncovered share is 79.55%, compared with 34.85% under the benchmark. Category counts and reference diversity alone do not distinguish these peripheral and dominant uses of uncovered fields. Appendix Table A20 gives the four systematically selected examples and their category records.

These within-paper comparisons hold each paper's records fixed; they do not repair the history imbalance in the original matched pairs. Results are similar when resampling accounts for both organizations and shared authors. The resulting 73 and 121 dependence components give structural-difference intervals of 6.23 to 11.45 and 6.03 to 9.34 points, respectively (Appendix Table A19).

4.3 Individual and additional team coverage

In the earlier paired samples, maximum individual coverage M averages 77.76% in discovery and 80.24% in first confirmation. Additional coauthor coverage D averages 3.78 and 2.61 percentage points (Table 3, Panel A). Coauthors cover a much larger total reference share, 63.46% and 64.27%, because most of it overlaps the anchor's coverage. The ratios of mean shared coverage to mean total coauthor coverage are 94.0% and 95.9%; they concern the cited references rather than the team's total knowledge.

Among AI-use projects with AI-classified references, occurrence-weighted anchor, additional-coauthor, and uncovered shares are 84.58%, 3.33%, and 12.08% in discovery (41 papers; 480 occurrences) and 79.50%, 3.45%, and 17.05% in first confirmation (62 papers; 522 occurrences). Table 3, Panel B compares these estimates with equal-paper averages. Appendix Table A12 also includes the smaller non-AI samples. These quantities attribute indexed coverage and do not identify who introduced AI to the team.

Equal-paper averages give a different account: 69.81%, 6.16%, and 24.03% in discovery and 63.93%, 10.89%, and 25.17% in first confirmation. The latter sample has a median of four AI references. Mean additional-coauthor coverage is 23.68%, 12.82%, 1.68%, and 2.31% for papers with 1–2, 3–4, 5–9, and at least 10 AI references, respectively. A single covered reference has greater influence in a short AI bibliography. Thus 10.89% describes the average paper, whereas 3.45% describes the pooled reference occurrences (Appendix Tables A12–A13).

The exploratory within-paper AI-versus-other-reference contrast in additional coverage is +7.52 points in first-confirmation AI-use papers ($p=.0445$), +1.60 when requiring at least three AI references ($p=.6111$), and -2.29 when requiring five ($p=.1987$). These unadjusted comparisons

change the sample. Their instability reinforces the need to report denominators before attributing an AI-specific coauthor advantage.

Table 3 Individual and additional team coverage

Panel A. All classified references

Stage	N	Anchor M (%)	Additional D (pp)	Coauthor coverage (%)	Shared / coauthor (%)
Discovery	115	77.76	3.78	63.46	94.04
First confirmation	240	80.24	2.61	64.27	95.93

Panel B. AI references in AI-use papers

Stage	N	Weighting	Anchor (%)	Additional (%)	Uncovered (%)
Discovery	41	Equal-paper	69.81	6.16	24.03
		Occurrence-weighted	84.58	3.33	12.08
First confirmation	62	Equal-paper	63.93	10.89	25.17
		Occurrence-weighted	79.50	3.45	17.05

Notes. Panel A reports equal-paper averages at the primary subfield threshold. M is maximum individual coverage; D is coverage added by coauthors. Shared / coauthor is the ratio of mean shared coverage to mean total coauthor coverage. Mean paper-specific ratios are 89.80% and 92.47%, with 12 and 29 undefined ratios. Panel B includes AI-use papers with at least one AI reference and uses the same prepublication histories and focal-paper anchor. Equal-paper and occurrence-weighted shares use the same records, with cited works deduplicated within each focal paper. Appendix Tables A12–A13 report all groups and denominator strata.

4.4 Replacement team calibration

The replacement benchmark uses unique donors, excludes focal-team members, reconstructs donor histories in the focal window, and reselects the maximum-covering author in each simulated team. Complete sets of 200 draws are available for 105 discovery papers in 70 organizations and 218 first-confirmation papers in 111 organizations. Single-author papers have $D=0$; one additional multiauthor paper per stage lacks a complete assignment. Table 4 reports this symmetric comparison. Appendix Table A23 preserves the initial and fixed-anchor versions.

With anchor reselection, the subfield observed-minus-benchmark contrast is +1.143 points in discovery (95% CI -0.095 to 2.381 ; $p=.0699$) and +0.417 in confirmation (-0.175 to 1.009 ; $p=.1656$). Both intervals include zero. Topic contrasts are +2.914 points (1.383 to 4.444 ; $p=.000311$) and +2.167 (1.510 to 2.824 ; $p=2.032 \times 10^{-9}$). Thus small D alone does not establish unusually redundant teams at the primary subfield scale. The topic results indicate greater marginal coverage than under the specified replacement mechanism, whose donor-pool bias and residual dependence remain unresolved.

The benchmark adapts the controlled-surrogate logic of Zeng et al. (2022) to reference-weighted marginal coverage. It makes the necessary comparison explicit: small observed D must be assessed against simulated teams evaluated by the same maximum-anchor rule.

Table 4 Replacement-team coverage with anchor reselection

Classification	Observed D (%)	Benchmark D (%)	Difference (pp)	95% CI	p
Discovery · 105 papers in 70 organizations					
Subfield	4.14	3.00	+1.14	[-0.10, 2.38]	.0699
Topic	3.63	.72	+2.91	[1.38, 4.44]	.0003
First confirmation · 218 papers in 111 organizations					
Subfield	2.85	2.44	+0.42	[-0.18, 1.01]	.1656
Topic	2.91	.74	+2.17	[1.51, 2.82]	<.0001

Notes. D denotes additional coauthor coverage. Difference = observed D minus the mean benchmark from 200 draws per paper. Both observed and replacement teams use a reselected maximum-covering anchor. Donors are unique, exclude focal-team members, and use histories from the focal window. Confidence intervals and p values use organization-cluster CR1 mean contrasts, without multiplicity adjustment. This exploratory benchmark is not an exact randomized test.

4.5 Reference contrasts and internal entry

Baseline-uncovered references are on average 3.722 years older than covered references within the same paper (95% CI 2.762 to 4.622; 299 papers in 135 organizations). Their review share is 0.193 percentage points lower, with an interval from -0.909 to 0.560 (Appendix Table A24). The age contrast therefore does not coincide with clear evidence of a higher review share and cannot by itself indicate unfamiliarity.

Among 300 papers with baseline-uncovered references, the mean share subsequently represented in intervening author histories is 12.89% (95% CI 10.28 to 15.81). Removing the cited works themselves from those histories reduces it to 10.83% (8.51 to 13.55). The unmodified share equals G divided by E_{pre} for each paper. Both versions depend on the same bibliographic records as the coverage measure, so neither supplies independent evidence of learning.

4.6 Second confirmation: selection and estimates

The second confirmation began with 832 selected grants, equally divided between AI-use and non-AI groups (Table 5). Of these, 363 were eligible for evidence collection, 330 had ambiguous NRF attribution, 84 had no eligible first article, 48 failed focal-metadata requirements, and seven overlapped earlier papers. These mutually exclusive categories partition grants; they are not consecutive exclusions. The linked articles comprise 748 grant–paper rows and 745 unique publications.

The baseline-history gate retains 202 observations, of which 194 meet all primary-model requirements: 112 AI-use and 82 non-AI grants in 97 organizations. The paired-history branch retains 198 before the remaining requirements and 190 afterward, in 95 organizations. Overall primary retention is 23.32%, with 26.92% for AI-use grants and 19.71% for controls. Failed first articles were not replaced (Appendix Table A4).

Table 5 Second confirmation selection and retention

Selection or retention stage	AI use	Non-AI	Total
Selected cohort and article linkage			
Fixed selection	416	416	832
First article linked	381	367	748
Mutually exclusive eligibility outcomes			
Eligible for evidence collection	203	160	363
Ambiguous NRF attribution	141	189	330
No first article	35	49	84
Focal metadata failure	31	17	48
Overlap with an earlier focal paper	6	1	7
Baseline-history branch			
Passes history gate	113	89	202
Included in primary model	112	82	194
Paired-history branch			
Passes history gate	110	88	198
Included in paired model	109	81	190

Notes. The five eligibility outcomes partition the 832 selected grants; they are not consecutive exclusions. Article linkage comprises 748 grant–paper rows and 745 unique papers. Baseline-history and paired-history retention form separate branches. The primary and paired models contain 97 and 95 organizations, respectively.

Attribution is the largest source of exclusion, and retention changes field composition. Engineering rises from 58.7% to 70.5% among AI-use grants and from 31.5% to 43.9% among controls. Budget and headcount distributions also change (Appendix Table A10). These differences describe selection into the observed sample; the acknowledgment literature does not identify the cause of each ambiguous case (Paul-Hus et al., 2016; Grassano et al., 2017).

Inverse-retention weighting changes the primary E_{pre} estimate little: -0.227 points (95% CI -7.673 to 7.218 ; $p=.9518$), with an effective sample size of 163.9 before clustering. The retention model uses 825 grants with complete covariates. It adjusts for measured predictors of observation but cannot correct unobserved selection or define outcomes for grants without an attributable paper. Appendix Tables A9–A11 report retention contrasts, probability ranges, and weight diagnostics.

The second-confirmation AI-use coefficient for E_{pre} is $+0.455$ percentage points under parent-field adjustment (95% CI -6.831 to 7.742 ; $p=.9015$; $N=194$). Detailed-field adjustment gives $+0.517$ points (-7.395 to 8.428 ; $p=.8972$). The restricted-null wild cluster bootstrap gives $p=.9001$ for the primary model. Neither specification meets the positive primary criterion; both intervals allow differences of substantive size in either direction (Table 6).

The failed primary test leaves the sequential A test inactive. The paired-history estimate is -1.621 points with parent-field adjustment (95% CI -4.633 to 1.391 ; $p=.2880$; $N=190$) and -2.423 with

detailed-field adjustment (-5.273 to 0.428 ; $p=.0948$). Table 6 reports these as descriptive estimates. Neither a positive association nor equivalence is established.

Table 6 Primary and sequential estimates

Field adjustment	AI-use b (pp)	SE (pp)	95% CI (pp)	p
Preproject uncovered share E_{pre} · N = 194; 97 organizations				
Parent field (primary)	+0.455	3.671	[−6.831, 7.742]	.9015
Detailed field	+0.517	3.986	[−7.395, 8.428]	.8972
Coverage change A · N = 190; 95 organizations				
Parent field	−1.621	1.517	[−4.633, 1.391]	.2880
Detailed field	−2.423	1.436	[−5.273, 0.428]	.0948

Notes. b is the AI-use coefficient in percentage points; SE is the organization-cluster CR1 standard error, with inference based on $t(\text{organizations} - 1)$. The parent-field E_{pre} model is primary; detailed-field adjustment is a sensitivity. The sequential test for A is inactive, so its estimates are descriptive. E_{pre} models include 112 AI-use and 82 non-AI observations; A models include 109 and 81. The primary-model wild cluster bootstrap uses 9,999 draws ($p = .9001$).

4.7 Other outcomes and robustness

The complete 15-outcome matrix for the common paired sample appears in Appendix Table A25. E_{pre} and E_{pub} coefficients are +0.07 and +1.69 points, whose difference gives $A=-1.62$. Gross gain and loss coefficients are -1.09 and $+0.53$. The coefficient for any gain is -17.09 points (95% CI -33.67 to -0.50 ; $p=.0436$), reversing the discovery sign. This comparison is unadjusted for multiplicity and cannot replace the failed primary endpoint.

AI-use status is associated with an 11.11-point higher AI-reference share and a 43.02-point higher probability of any AI reference. These compositional differences do not imply a baseline-coverage effect. The full matrix satisfies $A=E_{pre}-E_{pub}=G-L$, the zero-sum relation for $M+D+E_{pub}$, and the additive AI/other-reference partition, providing an internal accounting check.

Sensitivity estimates do not reveal a stable positive baseline association (Appendix Table A27). Requiring two publications per covered subfield gives -5.32 points (95% CI -15.23 to 4.58); topic-level classification gives -4.69 (-13.62 to 4.25). Estimates are -1.12 points in the prior-author-disjoint sample and -4.42 on common field-year support. Inverse-allocation weighting gives $+2.06$ points, and excluding discordant DOI groups leaves the primary estimate unchanged.

Across 97 organization-deletion fits, the primary coefficient ranges from -0.73 to $+2.09$ points, with no p-value below .532. The organization-disjoint subset contains only nine observations, below the fixed estimation minimum. Shared organizations therefore limit comparisons across stages. Nonengineering and baseline AI-history interaction tests are also imprecise ($p=.1127$ and $.6057$).

Coverage-change sensitivities remain imprecise (Appendix Table A16). Removing referenced works from intervening histories gives $A=-0.492$ points ($p=.7657$); excluding conference papers gives -1.629 ($p=.2886$), and combining the rules gives -0.496 ($p=.7651$). The two exact-date variants concern gross gain G, not A; each gives -2.330 points ($p=.0377$). These unadjusted exploratory

results do not activate the sequential test. Adjustment for initial coverage, publication lag, and intervening output leaves conditional A estimates between -1.838 and -1.310 points.

4.8 Earlier stages and field composition

The first confirmation failed its prespecified test of positive A: the estimate was $+1.525$ points (95% CI -1.134 to 4.184 ; $p=.2584$; $N=240$). Discovery yielded $+3.706$ points ($p=.0216$), and pooling the earlier stages yielded $+2.218$ ($p=.0285$). That pooled result includes the discovery observations and is a post-result synthesis. Appendix Table A26 reports all 15 earlier outcomes and stage interactions.

The earlier confirmation E_{pre} coefficient was $+7.417$ points ($p=.0038$), and the pooled coefficient was $+5.872$ (95% CI 0.851 to 10.892 ; $p=.0222$). These post-result findings motivated the prospective second-confirmation endpoint. The earlier $+7.417$ lies within the new primary interval, -6.831 to 7.742 ; their different significance does not establish attenuation. A stage comparison would need joint covariance accounting for shared authors and organizations.

Detailed-field adjustment reduces the earlier confirmation E_{pre} coefficient from $+7.417$ to $+5.441$ points (95% CI -0.228 to 11.111 ; $p=.0598$), and the pooled coefficient from $+5.872$ to $+4.933$ ($p=.0749$). Appendix Table A28 compares both field adjustments for each earlier stage. These analyses followed inspection of the earlier outcomes; the corresponding second-confirmation sensitivity was specified before its outcomes were observed.

Earlier field-specific patterns are also unstable. In first confirmation, the detailed-field E_{pre} coefficient is $+10.049$ points in engineering ($N=124$; $p=.0283$) and -0.621 elsewhere ($N=116$; $p=.8830$). The nonengineering interaction p -value changes from $.0135$ to $.0816$ when detailed fields replace parent fields; the pooled interaction changes from $.2822$ to $.4970$. These results show sensitivity to disciplinary composition without identifying a stable subgroup effect.

In first confirmation, the $+5.892$ -point publication-window E_{pub} coefficient comprises $+5.038$ points from non-AI references and $+0.854$ from AI-classified references (Appendix Table A26). These publication-window components neither decompose E_{pre} nor establish knowledge acquisition. Conditioning E_{pub} on E_{pre} gives -0.0318 points ($p=.9762$), the sign-reversed A coefficient under the same adjustment. The two regressions contain the same conditional information.

The first-confirmation interaction with preproject AI history has $p=.4765$. The separate pooled interaction, $p=.793$, uses prepublication history and addresses a different window. Mean AI-reference shares among AI-use projects are 20.26% in discovery and 11.17% in first confirmation; the source of this sample difference remains unresolved.

5 Discussion

The framework links a paper's references to the publication records of its authors. It separates the reference share represented in a single individual's history from the additional share represented across the team and the remainder absent from those records. This output-specific comparison complements reference diversity and between-author profile differences. Keeping the references fixed across observation windows also separates gains from losses in recorded coverage.

The within-paper analysis shows why citation weight matters. At fixed histories, cited subfields, diversity, and field totals, actual portfolios place less weight on uncovered subfields than the structural benchmark. The contrast persists outside the focal paper's own primary subfield, where simple overlap between a paper's main subject and its authors' records cannot account for it. Ordinary subfield citation frequencies nevertheless explain much of the full-portfolio difference, and the remaining discovery contrast is imprecise. The structural and frequency comparisons therefore answer distinct questions about the observed portfolios.

The regressions show a related but narrower result. Shannon adds .174 and .202 to R^2 after author count, reference count, and history volume are controlled; it also lowers organization-exclusion prediction error. These are improvements in explaining recorded E_{pre} , not evidence that E_{pre} outperforms established measures against an external criterion. The within-paper analysis addresses what the regressions cannot: differences in citation emphasis when diversity and histories are unchanged.

The decomposition changes what can be reported about collaboration. In first-confirmation AI-use papers, additional coauthor coverage is 10.89% for the average paper but 3.45% for the reference corpus. Both summarize the same records, with different weights. Stating the denominator thus changes the substantive account of the coauthors' marginal coverage. Anchor selection matters as well: the small subfield D has no clear excess over the symmetrically evaluated replacement benchmark, whereas topic-level excess is positive. Reference weighting, category resolution, and anchor selection should be reported together when comparing collaboration profiles.

The examples make this use concrete. Several uncovered categories may account for only a small fraction of a paper's citations, while one uncovered category can dominate another paper's bibliography. Coverage distinguishes those cases without assigning a cognitive interpretation to them. Its accuracy nevertheless depends on the indexed histories, field assignments, and author identities. Neither the matched-case search nor the within-paper benchmark resolves errors in those records.

The AI application does not confirm the proposed association. The first confirmation failed its coverage-change test. Once E_{pre} was designated primary for the second confirmation, its AI-use coefficient was close to zero under both field adjustments. AI-use projects cite more AI literature and have more observed prepublication AI history, but these compositional differences do not establish a higher overall baseline uncovered share. A project can draw heavily on AI literature already represented in its eventual authors' records.

Precision limits that conclusion. With a primary standard error of 3.671 points, the fixed-precision 80%-power threshold is 10.39 points, and power against the earlier +7.417-point estimate is 51.6%. Detailed-field adjustment gives 45.3% against that same benchmark and 27.2% against its own earlier +5.441-point estimate. The new interval includes the earlier primary point estimate. The failed test therefore establishes neither equivalence nor a statistically demonstrated decline across stages.

The A estimates of +3.706, +1.525, and -1.621 points correspond to discovery, the unsuccessful first confirmation, and the inactive second sequential test. They should not be combined as three prospective tests of the same primary endpoint. Shared authors and organizations, changing selection, and the switch in primary outcome also preclude interpreting their sequence as an identified attenuation mechanism.

A descriptive use of E requires reporting history counts, the cited categories, and the benchmark. Actual and field-conditioned coverage distinguish the number of uncovered categories from their citation weight; the frequency comparison shows how much of that allocation is shared with other teams. These summaries support comparisons of recorded coverage, but they do not provide a general normalization across fields or a basis for ranking researchers.

External validity remains bounded by selection: 194 of 832 fixed selected grants enter the primary model. Attribution, metadata, history, and quality requirements determine that subset. Allocation weights address the selection of controls; post-result retention weights address measured predictors of observation. Similar estimates under these adjustments do not recover missing outcomes or correct unobserved selection. Generalization should remain tied to the conservatively linked, observed population.

Independent assessments of familiarity or verified research tasks would be needed for stronger expertise claims. Future work can also compare history windows, taxonomies, and verified project-start rosters. The present evidence supports the narrower use of coverage as a reproducible account of citation emphasis relative to recorded publication histories.

6 Conclusion

Reference coverage connects a paper's cited fields to individual and collective publication histories. Actual references place more weight in recorded fields than under reallocations that preserve histories, occupied subfields, diversity, and broader-field totals. The contrast persists after removing the focal paper's own primary subfield; it is not confined to overlap in the paper's main subject. Common subfield citation frequencies explain much of the full-portfolio difference. The decomposition also clarifies why aggregation matters: coauthors' additional coverage can be appreciable for the average paper while accounting for a small share of pooled reference occurrences.

The three-stage AI application does not confirm a positive association with baseline uncovered share, and the final sequential test of coverage change remains inactive. The measurement contribution is a reproducible account of which cited fields appear in whose publication records and how heavily those fields are cited. Interpreting an uncovered reference as unfamiliar knowledge requires evidence beyond those records.

References

- Abramo, G., D'Angelo, C. A., & Zhang, L. (2018). A comparison of two approaches for measuring interdisciplinary research output: The disciplinary diversity of authors vs the disciplinary diversity of the reference list. *Journal of Informetrics*, 12(4), 1182–1193. <https://doi.org/10.1016/j.joi.2018.09.001>
- Ávila-Robinson, A., Mejia, C., & Sengoku, S. (2021). Are bibliometric measures consistent with scientists' perceptions? The case of interdisciplinarity in research. *Scientometrics*, 126, 7477–7502. <https://doi.org/10.1007/s11192-021-04048-0>
- Bianchini, S., Di Girolamo, V., Ravet, J., & Arranz, D. (2026). AI in science: When and where it makes a difference. *Research Policy*, 55(6), 105478. <https://doi.org/10.1016/j.respol.2026.105478>

- Bianchini, S., Müller, M., & Pelletier, P. (2022). Artificial intelligence in science: An emerging general method of invention. *Research Policy*, 51(10), 104604. <https://doi.org/10.1016/j.respol.2022.104604>
- Cantone, G. G. (2024). How to measure interdisciplinary research? A systemic design for the model of measurement. *Scientometrics*, 129, 4937–4982. <https://doi.org/10.1007/s11192-024-05085-1>
- Céspedes, L., Kozłowski, D., Pradier, C., Sainte-Marie, M. H., Shokida, N. S., Benz, P., Poitras, C., Ninkov, A. B., Ebrahimi, S., Ayeni, P., Filali, S., Li, B., & Larivière, V. (2025). Evaluating the linguistic coverage of OpenAlex: An assessment of metadata accuracy and completeness. *Journal of the Association for Information Science and Technology*, 76(6), 884–895. <https://doi.org/10.1002/asi.24979>
- Cohen, W. M., & Levinthal, D. A. (1990). Absorptive capacity: A new perspective on learning and innovation. *Administrative Science Quarterly*, 35(1), 128–152. <https://doi.org/10.2307/2393553>
- Culbert, J. H., Hobert, A., Jahn, N., Haupka, N., Schmidt, M., Donner, P., & Mayr, P. (2025). Reference coverage analysis of OpenAlex compared to Web of Science and Scopus. *Scientometrics*, 130(4), 2475–2492. <https://doi.org/10.1007/s11192-025-05293-3>
- Du, A., Head, M., & Brede, M. (2025). Integration vs segregation: Network analysis of interdisciplinarity in funded and unfunded research on infectious diseases. *Journal of Informetrics*, 19(1), 101634. <https://doi.org/10.1016/j.joi.2024.101634>
- Duede, E., Dolan, W., Bauer, A., Foster, I., & Lakhani, K. (2024). Oil & water? Diffusion of AI within and across scientific fields [Preprint]. *arXiv*. <https://doi.org/10.48550/arXiv.2405.15828>
- Foster, J. G., Rzhetsky, A., & Evans, J. A. (2015). Tradition and innovation in scientists' research strategies. *American Sociological Review*, 80(5), 875–908. <https://doi.org/10.1177/0003122415601618>
- Gao, J., & Wang, D. (2024). Quantifying the use and potential benefits of artificial intelligence in scientific research. *Nature Human Behaviour*, 8(12), 2281–2292. <https://doi.org/10.1038/s41562-024-02020-5>
- Grassano, N., Rotolo, D., Hutton, J., Lang, F., & Hopkins, M. M. (2017). Funding data from publication acknowledgments: Coverage, uses, and limitations. *Journal of the Association for Information Science and Technology*, 68(4), 999–1017. <https://doi.org/10.1002/asi.23737>
- Guimerà, R., Uzzi, B., Spiro, J., & Amaral, L. A. N. (2005). Team assembly mechanisms determine collaboration network structure and team performance. *Science*, 308(5722), 697–702. <https://doi.org/10.1126/science.11106340>
- Haeussler, C., & Sauermann, H. (2020). Division of labor in collaborative knowledge production: The role of team size and interdisciplinarity. *Research Policy*, 49(6), 103987. <https://doi.org/10.1016/j.respol.2020.103987>
- Hao, Q., Xu, F., Li, Y., & Evans, J. (2026). Artificial intelligence tools expand scientists' impact but contract science's focus. *Nature*, 649(8099), 1237–1243. <https://doi.org/10.1038/s41586-025-09922-y>

- Jones, B. F. (2009). The burden of knowledge and the ‘death of the renaissance man’: Is innovation getting harder? *The Review of Economic Studies*, 76(1), 283–317. <https://doi.org/10.1111/j.1467-937X.2008.00531.x>
- Kim, J., & Owen-Smith, J. (2021). ORCID-linked labeled data for evaluating author name disambiguation at scale. *Scientometrics*, 126(3), 2057–2083. <https://doi.org/10.1007/s11192-020-03826-6>
- Lee, Y.-N., Walsh, J. P., & Wang, J. (2015). Creativity in scientific teams: Unpacking novelty and impact. *Research Policy*, 44(3), 684–697. <https://doi.org/10.1016/j.respol.2014.10.007>
- Leydesdorff, L., & Rafols, I. (2011). Indicators of the interdisciplinarity of journals: Diversity, centrality, and citations. *Journal of Informetrics*, 5(1), 87–100. <https://doi.org/10.1016/j.joi.2010.09.002>
- Leydesdorff, L., Wagner, C. S., & Bornmann, L. (2019). Diversity measurement: Steps towards the measurement of interdisciplinarity? *Journal of Informetrics*, 13(3), 904–905. <https://doi.org/10.1016/j.joi.2019.03.016>
- Luncheon, A., Huang, Y., & Contractor, N. S. (2014). Understanding the assembly of interdisciplinary teams and its impact on performance. *Journal of Informetrics*, 8(1), 59–70. <https://doi.org/10.1016/j.joi.2013.10.006>
- Luncheon, A., Whalen, R., Wu, Y. J., DeChurch, L. A., & Contractor, N. S. (2023). Diversity, networks, and innovation: A text analytic approach to measuring expertise diversity. *Network Science*, 11(1), 36–64. <https://doi.org/10.1017/nws.2022.34>
- Messeri, L., & Crockett, M. J. (2024). Artificial intelligence and illusions of understanding in scientific research. *Nature*, 627(8002), 49–58. <https://doi.org/10.1038/s41586-024-07146-0>
- Milojević, S. (2014). Principles of scientific research team formation and evolution. *Proceedings of the National Academy of Sciences*, 111(11), 3984–3989. <https://doi.org/10.1073/pnas.1309723111>
- Nakhoda, M., Whigham, P., & Zwanenburg, S. (2025). How relevant is collaboration in interdisciplinary research? Bibliometric evidence from an institution’s publications. *Scientometrics*, 130, 6833–6846. <https://doi.org/10.1007/s11192-025-05492-y>
- Paul-Hus, A., Desrochers, N., & Costas, R. (2016). Characterization, description, and considerations for the use of funding acknowledgement data in Web of Science. *Scientometrics*, 108(1), 167–182. <https://doi.org/10.1007/s11192-016-1953-y>
- Paul-Hus, A., Mongeon, P., Sainte-Marie, M., & Larivière, V. (2017). The sum of it all: Revealing collaboration patterns by combining authorship and acknowledgements. *Journal of Informetrics*, 11(1), 80–87. <https://doi.org/10.1016/j.joi.2016.11.005>
- Priem, J., Piwowar, H., & Orr, R. (2022). OpenAlex: A fully-open index of scholarly works, authors, venues, institutions, and concepts [Preprint]. *arXiv*. <https://doi.org/10.48550/arXiv.2205.01833>

- Rafols, I., & Meyer, M. (2010). Diversity and network coherence as indicators of interdisciplinarity: Case studies in bionanoscience. *Scientometrics*, 82, 263–287. <https://doi.org/10.1007/s11192-009-0041-y>
- Salatino, A., Osborne, F., Reforgiato Recupero, D., Angioni, S., & Motta, E. (2026). Does diversity of expertise drive citation impact? Evidence from computer science. *Scientometrics*, 131(2), 1119–1146. <https://doi.org/10.1007/s11192-026-05560-x>
- Sourati, J., & Evans, J. A. (2023). Accelerating science with human-aware artificial intelligence. *Nature Human Behaviour*, 7(10), 1682–1696. <https://doi.org/10.1038/s41562-023-01648-z>
- Stirling, A. (2007). A general framework for analysing diversity in science, technology and society. *Journal of the Royal Society Interface*, 4(15), 707–719. <https://doi.org/10.1098/rsif.2007.0213>
- Tahamtan, I., & Bornmann, L. (2018). Creativity in science and the link to cited references: Is the creative potential of papers reflected in their cited references? *Journal of Informetrics*, 12(3), 906–930. <https://doi.org/10.1016/j.joi.2018.07.005>
- Tekles, A., & Bornmann, L. (2020). Author name disambiguation of bibliometric data: A comparison of several unsupervised approaches. *Quantitative Science Studies*, 1(4), 1510–1528. https://doi.org/10.1162/qss_a_00081
- Uzzi, B., Mukherjee, S., Stringer, M., & Jones, B. (2013). Atypical combinations and scientific impact. *Science*, 342(6157), 468–472. <https://doi.org/10.1126/science.1240474>
- Wagner, C. S., Roessner, J. D., Bobb, K., Klein, J. T., Boyack, K. W., Keyton, J., Rafols, I., & Börner, K. (2011). Approaches to understanding and measuring interdisciplinary scientific research (IDR): A review of the literature. *Journal of Informetrics*, 5(1), 14–26. <https://doi.org/10.1016/j.joi.2010.06.004>
- Wang, H., Fu, T., Du, Y., Gao, W., Huang, K., Liu, Z., Chandak, P., Liu, S., Van Katwyk, P., Deac, A., Anandkumar, A., Bergen, K., Gomes, C. P., Ho, S., Kohli, P., Lasenby, J., Leskovec, J., Liu, T.-Y., Manrai, A., ... Zitnik, M. (2023). Scientific discovery in the age of artificial intelligence. *Nature*, 620(7972), 47–60. <https://doi.org/10.1038/s41586-023-06221-2>
- Wu, L., Wang, D., & Evans, J. A. (2019). Large teams develop and small teams disrupt science and technology. *Nature*, 566(7744), 378–382. <https://doi.org/10.1038/s41586-019-0941-9>
- Wuchty, S., Jones, B. F., & Uzzi, B. (2007). The increasing dominance of teams in production of knowledge. *Science*, 316(5827), 1036–1039. <https://doi.org/10.1126/science.1136099>
- Yang, A. J. (2024). On the temporal diversity of knowledge in science. *Journal of Informetrics*, 18(4), 101594. <https://doi.org/10.1016/j.joi.2024.101594>
- Zeng, A., Fan, Y., Di, Z., Wang, Y., & Havlin, S. (2021). Fresh teams are associated with original and multidisciplinary research. *Nature Human Behaviour*, 5(10), 1314–1322. <https://doi.org/10.1038/s41562-021-01084-x>
- Zeng, A., Fan, Y., Di, Z., Wang, Y., & Havlin, S. (2022). Impactful scientists have higher tendency to involve collaborators in new topics. *Proceedings of the National Academy of Sciences*, 119(33), e2207436119. <https://doi.org/10.1073/pnas.2207436119>

- Zeng, A., Shen, Z., Zhou, J., Fan, Y., Di, Z., Wang, Y., Stanley, H. E., & Havlin, S. (2019). Increasing trend of scientists to switch between topics. *Nature Communications*, 10, 3439. <https://doi.org/10.1038/s41467-019-11401-8>
- Zhang, L., Rousseau, R., & Glänzel, W. (2016). Diversity of references as an indicator of the interdisciplinarity of journals: Taking similarity between subject fields into account. *Journal of the Association for Information Science and Technology*, 67(5), 1257–1265. <https://doi.org/10.1002/asi.23487>
- Zwanenburg, S., Nakhoda, M., & Whigham, P. (2022). Toward greater consistency and validity in measuring interdisciplinarity: A systematic and conceptual evaluation. *Scientometrics*, 127, 7769–7788. <https://doi.org/10.1007/s11192-022-04310-z>

Supplementary Files

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

- [ScientometricsSupplementAUTHORDETAILS20260919.docx](#)