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.

# Does Conceptual Centrality Pay Off in Citations Everywhere? Topic-Dependent Evidence from Library and Information Science

**Kumar Sumit**
Mizoram University
**Manendra Singh**
manendra@mzu.edu.in
Mizoram University

**Research Article**

**Keywords:** conceptual centrality, citation attention, semantic embeddings, topic modeling, bibliometrics, scientometrics

**Posted Date:** November 6th, 2026

**DOI:** https://doi.org/10.21203/rs.3.rs-11332556/v1

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

**Additional Declarations:** No competing interests reported.

# Does Conceptual Centrality Pay Off in Citations Everywhere? Topic-Dependent Evidence from Library and Information Science

Kumar Sumit[1] and Manendra Singh[2*†]

[1*]Department of Library and Information Science, Mizoram University, Tanhrill, Aizawl, 796004, Mizoram, India.

*Corresponding author(s). E-mail(s): manendra@mzu.edu.in;
Contributing authors: aryan.sumit121@gmail.com;
†These authors contributed equally to this work.

**Abstract**

Citation counts are widely used as indicators of scholarly recognition, but a paper's citation performance need not correspond closely to its position within the conceptual structure of its research field. This study explores that distinction in Library and Information Science by introducing a topic-relative measure of conceptual centrality, defined as the leave-one-out cosine similarity between a paper's embedding and the centroid of the other papers in its topic. Using a Scopus corpus of 14,720 papers organized into 52 empirically derived research topics, centrality shows a positive but modest aggregate association with citation rate (Spearman's $r_s = .107$, $p < .001$). The association becomes stronger after accounting for the journal citation environment (standardized $\beta = .095$ for centrality versus **.312** for journal environment), consistent with suppression rather than confounding. More notably, its strength is not uniform across topics. Among 45 topics with at least 40 regression observations, 26 show a statistically distinguishable positive slope, 19 are inconclusive, and none shows a reliable negative slope. This heterogeneity is not an artifact of topic sample size ($p = .844$). The pattern persists when centrality is constructed using only prior publications ($F = 4.79$, $p < .001$), and a residual analysis shows that review articles are modestly overrepresented among papers cited more than expected (7.7% versus 5.6%, $p = .008$). The findings support a multidimensional view of scholarly recognition in which conceptual embeddedness and citation attention are related but not interchangeable, and their relationship depends on the topical context in which a paper sits.

**Keywords:** conceptual centrality, citation attention, semantic embeddings, topic modeling, bibliometrics, scientometrics

# 1 Introduction

Today's citation counts take a central position in quantitative research evaluation, as they provide an observable record of subsequent scholarly interaction and recognition. Citation accumulation reflects multiple processes operating at the paper, author, venue, field, and temporal levels. Publication venue, collaboration, document type, publication age, accessibility, and other characteristics have all been associated with differences in citation counts. Prior work has repeatedly cautioned that citations are informative but imperfect indicators of the contribution of an individual article [1–3]. Empirical studies have shown that citation counts can correspond reasonably well with assessments of scientific contribution at an aggregate level while remaining unreliable as stand-alone indicators for individual articles [1]. Collaboration, journal, document, and other publication characteristics jointly contribute to differences in citation counts [4, 5]. These findings motivate the use of multivariable specifications when evaluating a new semantic correlate of citation performance. Overall, a citation count is better understood as an outcome of scholarly recognition than as a direct measurement of intellectual value. This distinction matters most when comparing papers in heterogeneous research environments where citation opportunities differ across fields and communities.

A complementary way of understanding scholarly recognition begins with the semantic organization of the literature itself. Research papers occupy positions within configurations of concepts, problems, methods, applications, and terminology. Transformer-based document embeddings make it possible to represent these configurations as high-dimensional semantic spaces, and embedding-based topic modeling pipelines can identify empirically derived research communities without requiring the researcher to specify a fixed thematic taxonomy in advance [6]. BERTopic, for example, combines document embeddings, dimensionality reduction, density-based clustering, and class-based TF-IDF topic representations [7]. Recent scientometric work has already demonstrated that textual representations can be associated with citation outcomes. Deep representations of scientific papers have been used to construct topicality and originality indicators related to subsequent scholarly impact [8], while GPT and other text embeddings have been used to predict citation-impact categories directly from article text [9]. More recent work has likewise examined semantic representations in relation to citation impact, including evidence of heterogeneous relationships across research contexts [10, 11]. These studies establish the feasibility of using semantic representations in citation analysis, but they do not resolve a different question: whether a paper's position relative to the conceptual structure of its research topic has a consistent relationship with citation attention. This study differs in emphasis from that predictive tradition. It derives an interpretable relational quantity, specifically the focal paper's position relative to the semantic centroid of its topic, instead of treating the embedding itself as a large set of predictive features. It then tests whether

the relationship between that position and citation attention is homogeneous across topics. The contribution is therefore less about maximizing prediction accuracy and more about identifying the conditions under which semantic embeddedness translates into citation recognition.

This study accordingly distinguishes conceptual position from citation prominence. The concept of centrality has a well-established foundation in relational analysis, where it refers to the position of an element within a structured system of relationships [12]. In scientometrics, related relational approaches have been used to identify scientific specialties and to examine topic-dependent patterns of scholarly impact [13–16]. Here, however, centrality is formulated in semantic rather than citation-network space. Conceptual centrality is defined as a paper's semantic proximity to the conceptual centroid of the other publications within its assigned research topic. The construct is deliberately topic-relative: it does not seek to identify a single conceptual center of the discipline but rather asks how deeply a paper is embedded in the conceptual configuration of related topics. The operational measure used here is leave-one-out cosine similarity between the focal document embedding and the centroid of the remaining embeddings in the same topic. The leave-one-out construction is important because if the focal paper were included in its centroid, the similarity to that centroid would be mechanically increased by its contribution to the reference point, and removing the focal paper avoids this direct self-inclusion. Conceptual centrality should be distinguished from intellectual quality, originality, importance, and theoretical value. It is a measure of semantic embeddedness within an empirically derived research topic, nothing more.

The theoretical rationale for treating semantic embeddedness as a meaningful property of scholarly work follows from the view that scientific communities organize research around shared conceptual frameworks, problems, exemplars, and modes of inquiry [17–19]. Scientometric and sociological accounts have similarly treated scientific specialties as structured intellectual communities rather than homogeneous collections of publications [14, 19]. The concept–symbol perspective further suggests that cited documents can function as symbols for particular concepts, methods, findings, or theoretical positions, linking citation practices to the conceptual organization of scientific work [20]. However, citation accumulation also reflects social processes of scholarly recognition and cumulative advantage rather than conceptual position alone [21, 22]. From this perspective, a paper's semantic embeddedness within a research topic and its subsequent citation recognition are theoretically related but non-equivalent properties. A paper may occupy a recognizable position within a shared conceptual structure without becoming a highly cited contribution, while another paper may receive substantial attention for other reasons, including a methodological, empirical, or practical contribution that is relatively peripheral to that structure.

This distinction creates a non-trivial empirical problem. If citation recognition is partly related to conceptual embeddedness, centrality should exhibit some positive association with citation performance. However, there is no theoretical reason to expect that the association will be homogeneous. Citations perform multiple functions. A paper may be cited as a conceptual foundation, methodological resource,

empirical precedent, review, data source, or convenient reference, and these functions need not correspond to the paper's location near the semantic core of a topic [2, 20, 23, 24]. A highly central paper can therefore accumulate modest citation attention if its conceptual content is widely shared but its individual contribution is not uniquely required. Conversely, a semantically peripheral paper may attract substantial attention because it contributes a transferable method, application, dataset, synthesis, or bridge to another research community. Citation functions may therefore not be reducible to semantic position. Moreover, a weak aggregate association between centrality and citations is not necessarily evidence that semantic position is irrelevant, because the association may simply be attenuated when heterogeneous citation regimes are pooled together. The critical empirical question is consequently whether the strength of the association changes across topics, not merely whether it exists on average.

Research topics plausibly differ in conceptual maturity, audience breadth, methodological dependence, interdisciplinarity, practical orientation, and citation norms. Research-area differences have been shown to matter for citation impact and for how citation relationships are modeled, while topic-sensitive scientometric studies demonstrate that scholarly impact can vary substantially across topical contexts [15, 16, 19, 25]. Recent work using semantic representations likewise suggests that relationships between textual characteristics and citation impact may be heterogeneous across research contexts [11]. In a topic with a strongly cumulative conceptual vocabulary, semantic embeddedness may translate more directly into citation recognition because subsequent work repeatedly builds on shared concepts. In a method-driven or application-oriented topic, citation attention may instead accrue to contributions that are useful outside the local conceptual core. These latter mechanisms are proposed as theoretical possibilities rather than directly tested explanations. This study therefore makes a deliberately narrower claim. The strength of the centrality-citation association varies across topic contexts. Topic size is evaluated as an alternative explanation because slope estimates from small topics can be unstable. If topic size does not explain the slope gradient, the observed heterogeneity cannot be attributed simply to differences in topic sample size and may instead reflect topic-specific or other unmeasured features of topic context.

The framework guiding this study is therefore deliberately non-causal. Conceptual centrality represents a semantic dimension of a paper's position within a research topic, while citation rate represents a subsequent recognition outcome. Publication and journal environments operate alongside semantic position as additional sources of citation opportunity, and topic is treated as a moderator of the centrality-citation relationship. Residual mismatch analysis then examines individual cases in which realized citation attention departs substantially from the level the model would predict. The present study builds on an earlier topic-stratified analysis of citation recognition in Library and Information Science [26]. It retains a topic-sensitive perspective while introducing an independent semantic construct and explicitly testing whether its relationship with citation performance varies across topics. Whereas the earlier analysis examined whether the association between open-access status and citation attention

varied across LIS topics, the present study asks whether the relationship between semantic embeddedness and citation attention is similarly topic-dependent.

The present study identifies three major gaps. First, existing citation-impact models have increasingly incorporated textual and semantic information [8–11], but comparatively little attention has been given to a topic-relative measure of semantic centrality that explicitly excludes the focal paper from the centroid used as its reference point. Second, semantic characteristics are often treated as broadly applicable predictors, leaving open the question of whether their citation consequences differ systematically among research topics. Evidence that citation relationships vary across fields and topics suggests that this assumption warrants explicit testing [15, 16, 19, 25]. Third, the use of aggregate regression coefficients provides information on averages but does not identify which papers show citation behavior that differs from what their conceptual position and contextual characteristics would lead one to expect. In this paper, these questions are addressed by combining a leave-one-out measure of semantic centrality with topic-adjusted regression, an antecedent-only robustness design, and a residual-based mismatch analysis. This is not another attempt to develop a universal citation predictor. The purpose of the paper is to examine whether conceptual embeddedness constitutes a distinct dimension of scholarly recognition.

This study pursues five objectives, corresponding to five research questions.

- **RO1:** To determine whether conceptual centrality is associated with citation rate at the discipline-wide level.
- **RO2:** To determine whether this association varies systematically across the discipline's research topics and whether any such variation is explained by topic size rather than topic identity.
- **RO3:** To determine whether the association remains stable, strengthens, or weakens after accounting for publication year, authorship, international collaboration, document type, journal citation environment, open access status, and topic.
- **RO4:** To determine whether the topic–dependent structure of the association persists when conceptual centrality is measured using only information available prior to publication.
- **RO5:** To determine whether papers whose citation performance diverges sharply from that predicted by their conceptual position exhibit identifiable characteristics.

Each of these objectives is examined through a corresponding research question, stated below in the same order.

- **RQ1:** Is a paper's conceptual centrality within its research topic associated with its citation rate?
- **RQ2:** Does this association vary systematically across the discipline's constituent research topics and is any such variation explained by topic size rather than topic identity?
- **RQ3:** Does the aggregate association remain stable after accounting for publication year, authorship, international collaboration, document type, journal citation environment, open access status, and topic?
- **RQ4:** Does the topic-dependent structure persist when conceptual centrality is measured using only information available prior to publication?

- **RQ5:** Do papers whose citation performance diverges sharply from that predicted by their conceptual position exhibit identifiable characteristics?

This study makes four contributions. First, it makes a methodological contribution by introducing a topic-relative semantic measure of conceptual centrality based on leave-one-out embedding centroids. The measure captures a paper's semantic embeddedness within its empirically derived research topic while avoiding direct self-inclusion in the reference centroid. Second, it makes an empirical contribution by showing that the centrality-citation relationship is substantially heterogeneous across research topics and that a single discipline-wide coefficient does not adequately capture this variation. The analysis further shows that this heterogeneity is not explained by topic size. Third, it makes a conceptual contribution by distinguishing semantic embeddedness from citation recognition as related but non-equivalent dimensions of scholarly recognition, with their relationship varying across topical contexts. Fourth, it makes a methodological and diagnostic contribution by combining an antecedent-only robustness design with a citation-centrality mismatch framework that identifies papers receiving substantially more or less citation attention than expected from their conceptual position and measured publication environment.

# 2 Materials and Methods

## 2.1 Data source and corpus construction

Bibliographic records were retrieved from Scopus (Elsevier) in 2026, which served as the citation reference year for all citation counts in this study. The search was restricted to the Social Sciences subject area, with document type limited to articles and reviews, language restricted to English, and publication years restricted to 2010–2024. Title-based keyword anchoring using Library and Information Science terminology was adopted after an initial broader bibliometric-methodology query produced substantial retrieval contamination from unrelated fields. The final search returned 15,254 records.

Records were checked for duplicate identifiers, and none were found. Bracketed non-English original-language titles were stripped prior to text processing. During the topic-modeling procedure described below, a distinct cluster of 534 records consisting exclusively of records with missing or placeholder abstract text was identified as a data-quality artifact and excluded from subsequent analysis, yielding a final analytic corpus of 14,720 records.

Author count was derived from semicolon-separated author entries. Records with zero parseable authors (17 records, 0.1% of the corpus) were coded as missing for models using author count as a covariate. International collaboration was coded as a binary indicator when author affiliations spanned more than one country. Document type was represented as a binary review indicator.

## 2.2 Topic discovery and assignment

Topic discovery used BERTopic, combining transformer-based document embeddings, UMAP dimensionality reduction, HDBSCAN density-based clustering, and class-based TF-IDF topic representation [7]. Cleaned titles and abstracts were concatenated to

form the document text. The primary embedding model was `all-MiniLM-L6-v2` (384 dimensions), a sentence-transformer model designed to generate semantically meaningful dense representations [6]. UMAP was used for dimensionality reduction with 15 neighbors, five components, minimum distance 0.0, cosine metric, and a fixed random seed [27]. HDBSCAN was used for density-based clustering with excess-of-mass cluster selection [28], and topic representations used English stopword removal with unigram-to-bigram tokenization.

The HDBSCAN minimum cluster size was selected through a sweep from 15 to 100, using topic coherence and the proportion of unclassified documents as joint criteria. A minimum cluster size of 40 provided the most favorable balance. The initial configuration classified 4,475 documents (29.3% of the pre-cleaning corpus) as outliers; these were reassigned using BERTopic's class-based TF-IDF outlier reduction strategy [7]. The resulting solution contained 53 topics. The 534-record artifact topic was then excluded, leaving 52 substantive topics with a final coherence score of 0.631. Topics were labeled manually based on each topic's top keywords and most representative documents.

Two semantic robustness checks were conducted. First, the topic pipeline was rerun using a higher-capacity embedding model (`all-mpnet-base-v2`, 768 dimensions), yielding 79 topics with a mean best-match Jaccard similarity of 0.714 against the primary solution. Second, a citation-graph-trained scientific-document embedding model was evaluated but not used as the primary representation. SPECTER2 was explicitly trained to encode scholarly papers using citation-based information [29]; it therefore provides a useful robustness comparison but was excluded from the primary analysis because it produced markedly coarser topic granularity and its citation-informed training objective could introduce circularity into a study relating semantic position to citation outcomes.

## 2.3 Conceptual centrality construction

For paper $i$ assigned to topic $j$, retrospective conceptual centrality was defined as the cosine similarity between its normalized embedding vector and a leave-one-out centroid constructed from all other papers assigned to topic $j$:

$$CC_i = \cos\left(E_i,\ \frac{1}{n_j - 1}\sum_{k\in j,\ k\neq i} E_k\right) \tag{1}$$

where $E_i$ denotes paper $i$'s normalized embedding vector and $n_j$ is the number of papers assigned to topic $j$. The leave-one-out construction prevents the focal paper from contributing to the centroid against which it is evaluated, avoiding the mechanical correlation that would otherwise arise from every paper being partially compared to itself. This measure was computed for all 14,720 papers and ranged from 0.078 to 0.927 ($M = 0.692$, $SD = 0.100$).

A prospective version of conceptual centrality was constructed using only papers from the same topic published in years strictly prior to a focal paper's own publication year $t$:

$$CC_i^{pre} = \cos\left(E_i,\ \frac{1}{n_{j,<t}} \sum_{k \in j,\ \mathrm{Year}(k)<t} E_k\right) \tag{2}$$

where $n_{j,<t}$ denotes the number of papers in topic $j$ published before year $t$.

The prospective measure was restricted to topics with at least 20 such antecedent papers, with the threshold fixed prior to inspection of the study results. This threshold was met for 13,195 of 14,720 papers (89.6 percent). Coverage was lowest in the corpus's earliest years and in the smallest topics, reaching near-complete coverage from 2015 onward. This prospective measure tests whether the centrality reference point depends on post-publication semantic information. Because the topic structure was discovered using the complete corpus and subsequently reviewed and verified by the authors, the prospective measure does not constitute a fully recursive, time-evolving topic model. Instead, it restricts the construction of the centrality reference point to literature published before the focal paper. This scope limitation is noted explicitly in the Discussion. Sensitivity to the antecedent threshold was assessed at minimum values of 10, 20, and 30 papers. The prospective centrality-citation correlation was stable across this range.

## 2.4 Citation rate and covariates

Citation rate was calculated as raw citation count divided by citation age, where citation age represents the number of full calendar years between publication year and the 2026 retrieval year:

$$\text{Citation rate}_i = \frac{\text{Citation count}_i}{\text{Citation age}_i} \tag{3}$$

The resulting rate was log transformed using $\ln(1 + \text{citation rate})$ to reduce right skew while retaining zero citation observations.

Covariates included author count, international collaboration, review status, open access status, and a journal citation environment measure. The journal measure was calculated as the leave-one-out mean citation rate of other papers published in the same source title, among journals contributing at least 10 papers:

$$\text{Journal citation environment}_i = \frac{1}{N_j - 1} \sum_{k \neq i} \text{Citation rate}_k \tag{4}$$

where $j$ denotes the source title containing paper $i$ and $N_j$ is the number of papers from that source title in the analytic sample. This restriction excluded 1,471 records and yielded 11,094 observations for models containing the journal measure. The measure represents realized citation performance in the observed sample and time window rather than an independently validated journal prestige indicator such as CiteScore or SJR.

## 2.5 Statistical analysis

The aggregate association between conceptual centrality and citation rate was evaluated using Spearman rank correlation, given the right-skewed distribution of citation

data. Nested ordinary least squares models predicted log citation rate, with HC3 heteroscedasticity-consistent standard errors throughout. Conceptual centrality was entered first, followed sequentially by publication year; author count and international collaboration; document type; journal citation environment; open access status; and topic fixed effects. Because predictors are measured on substantively different natural scales, standardized coefficients are reported wherever the relative practical magnitude of different predictors is compared.

Topic heterogeneity was tested by estimating the full interaction between conceptual centrality and topic, yielding 51 interaction terms for 52 topics, tested jointly with an F test. Topic-specific marginal slopes were calculated as the sum of the main centrality coefficient and the corresponding interaction term, with standard errors derived from the full coefficient covariance matrix rather than by treating the two terms as independent. An inverse-variance-weighted meta-regression of topic-specific slopes on log topic sample size tested whether slope heterogeneity was attributable to topic size. Topics with at least 40 regression observations were designated standard-sample topics for individual interpretation; smaller topics remained in the formal joint test but were flagged accordingly.

The prospective analysis was restricted to papers with citation ages of four to five years, corresponding to a 2021 to 2022 publication cohort given the corpus window and antecedent coverage requirements described above.

Standardized residuals from the fully specified topic-interaction model were used to identify papers whose citation performance diverged substantially from prediction:

$$z_i = \frac{y_i - \hat{y}_i}{\hat{\sigma}_{\text{residual}}} \tag{5}$$

where $y_i$ is paper $i$'s observed log citation rate, $\hat{y}_i$ is the value predicted by the fully specified model, and $\hat{\sigma}_{\text{residual}}$ is the residual standard deviation across the analytic sample. Papers with $z_i \geq 1.5$ were classified as high positive-residual attention outliers; papers with $z_i \leq -1.5$ were classified as high negative-residual cases. Thresholds of $\pm 1.0$ and $\pm 2.0$ were examined for sensitivity. Review status was compared across mismatch categories using a chi-square test. High negative-residual cases were additionally inspected qualitatively for recurring topic-specific structural characteristics; these observations are descriptive and do not constitute a formal statistical test unless noted as such.

The design is observational and cross-sectional. Conceptual centrality is not manipulated or randomly assigned; consequently, all reported coefficients are interpreted as conditional statistical associations rather than causal effects.

## 2.6 Software and reproducibility

All analyses were conducted in Python. Table 1 summarizes the components of the analysis pipeline, together with the specific tools and configuration settings used at each stage.

**Table 1** Software and tools used in the analysis pipeline

| Component | Tool | Configuration |
|---|---|---|
| Data retrieval | Scopus (Elsevier) | 2010–2024; articles and reviews; English |
| Primary embeddings | sentence-transformers | all-MiniLM-L6-v2; 384 dimensions |
| Robustness embeddings | sentence-transformers | all-mpnet-base-v2; 768 dimensions |
| Dimensionality reduction | UMAP | 15 neighbors; 5 components; minimum distance = 0.0; cosine metric; fixed random seed |
| Clustering | HDBSCAN | min_cluster_size = 40; excess-of-mass |
| Topic modeling | BERTopic | class-based TF-IDF representation |
| Data manipulation | pandas, numpy | — |
| Statistical modeling | statsmodels, scipy.stats | OLS with HC3; WLS meta-regression |
| Visualization | matplotlib | — |

# 3 Results

## 3.1 Descriptive overview

The final analytic corpus comprised 14,720 papers organized into 52 empirically derived research topics. Retrospective conceptual centrality ranged from 0.078 to 0.927 ($M = 0.692$, $SD = 0.100$, median = 0.706); prospective centrality, available for 13,195 papers, showed a closely comparable distribution ($M = 0.688$, $SD = 0.100$, median = 0.702; Table 2, Figure 1). Citation rate was strongly right skewed, with a mean of 1.841 citations per year against a median of 0.857, a gap wide enough on its own to motivate the log transformation used throughout the regression analysis rather than treating the mean as representative of a typical paper's citation experience.

**Table 2** Descriptive statistics for conceptual centrality and citation rate

| Variable | N | Mean | SD | Min | Median | Max |
|---|---|---|---|---|---|---|
| Conceptual centrality (retrospective) | 14,720 | 0.692 | 0.100 | 0.078 | 0.706 | 0.927 |
| Conceptual centrality (prospective) | 13,195 | 0.688 | 0.100 | 0.073 | 0.702 | 0.932 |
| Citation rate (per year) | 14,720 | 1.841 | 3.857 | 0.000 | 0.857 | 114.333 |
| Log citation rate | 14,720 | 0.759 | 0.649 | 0.000 | 0.619 | 4.748 |

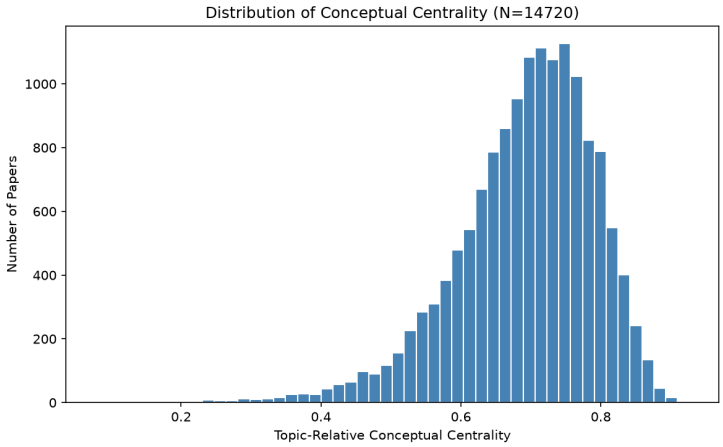



**Fig. 1** Distribution of topic-relative conceptual centrality across all 14,720 papers in the analytic corpus.

## 3.2 Does the centrality-citation relationship vary across research topics?

Across the full corpus, conceptual centrality was positively but weakly associated with citation rate (Spearman $r_s = .107$, $p < .001$, $N = 14{,}720$). Read in isolation, this figure would support a fairly deflationary conclusion: that a paper's semantic position within its topic tells a reader almost nothing about how often that paper will be cited, and that conceptual centrality is a scientifically interesting but practically negligible correlate of scholarly attention.

The remainder of this section shows why that reading would be premature, and specifically why. A single discipline wide correlation coefficient implicitly assumes that the centrality citation relationship is a stable, transferable quantity across the 52 topics comprising the corpus. If instead this relationship differs sharply from one topic to another, a pooled correlation does not describe a real, shared phenomenon so much as an average over topics that may have almost nothing in common with one another statistically. Establishing which of these two readings is correct, rather than treating the aggregate figure as the headline result, is the central empirical task of this study.

## 3.3 Is topic size responsible for the observed heterogeneity?

A joint test of the interaction between conceptual centrality and topic, estimated within the fully specified regression model, was highly significant ($F(51, 10985) = 4.15$, $p = 3.21 \times 10^{-21}$, $N = 11{,}094$). This result directly falsifies the implicit assumption behind the pooled correlation: the centrality citation relationship is not a single, shared quantity across the discipline, but a family of topic specific relationships that differ from one another far more than sampling variation would allow.

Among the 45 topics meeting the prespecified minimum sample size of 40 regression observations, 26 showed statistically significant positive marginal slopes and 19 were inconclusive, with 95 percent confidence intervals spanning zero. No standard sample

topic showed a statistically significant negative association (Table 3; Figure 2). This asymmetry is itself worth pausing on: if centrality's relationship with citations varied purely at random across topics, a comparable number of significant negative slopes would be expected alongside the significant positive ones. Their complete absence indicates that the observed heterogeneity is concentrated in the range from positive to statistically indistinguishable from zero associations, rather than extending to reliable negative relationships.

The topics carrying the strongest associations, Health Information Seeking Behavior ($\beta = 2.414$), Embedded and Distance Learning Librarianship ($\beta = 1.755$), and Public Libraries: Community and Value ($\beta = 1.448$), sit alongside topics such as Information Retrieval Systems, whose point estimate was the most negative in the standard sample tier ($\beta = -1.008$, 95% CI $[-2.261, 0.246]$) but did not reach significance. This particular contrast is worth naming explicitly, since it illustrates the substantive content behind the statistical pattern: topics oriented toward broad, practice facing audiences, health information behavior, embedded instructional librarianship, public library value, show the tightest coupling between a paper's conceptual position and its citation attention, while a narrower, more technically self contained topic like information retrieval shows no detectable coupling at all. One plausible reading is that in topics with a wide, applied readership, a paper's proximity to the field's dominant conceptual vocabulary genuinely tracks how legible and citable it is to that broad audience, whereas in a more specialized, self referential technical topic, citation attention may be driven by narrower factors, such as a specific algorithmic contribution or dataset, that are orthogonal to conceptual centrality as measured here. This account is offered as a plausible interpretation consistent with the pattern, not as a mechanism this study tested directly.

This heterogeneity is not attributable to topic size. An inverse variance weighted meta regression of topic specific slopes on log regression sample size found no significant relationship ($b = -0.019$, $SE = 0.094$, $p = .844$, $R^2 = .001$, 52 topics). This null result matters more than its small effect size might suggest, because it rules out the most parsimonious, least interesting explanation for Figure 2's pattern, namely that large, well sampled topics simply produce more precise, and therefore more often significant, estimates while small topics produce noisier, less often significant ones by construction. The analysis provides no evidence that topic sample size systematically explains the observed variation in slope magnitude.

## 3.4 Does the association persist after accounting for publication and journal characteristics?

The nested regression sequence clarifies not only whether the centrality citation association survives standard controls, but why the aggregate correlation reported above was so weak to begin with. The centrality coefficient was stable across the first several specifications: 0.920 in the bivariate model, 0.897 after adding year, and approximately 0.90 after further adding authorship, international collaboration, and document type. Introducing journal citation environment produced a sharp increase, to 1.029, the largest single change anywhere in the sequence; the estimate remained elevated after adding open access status (1.030) and topic fixed effects (1.011; Figure 3, Panel A).

**Table 3** Selected topic-specific marginal slopes (standard-sample topics; full ranking across all 52 topics shown in Figure 2)

| Topic | N | Slope | p | Direction |
|---|---|---|---|---|
| Health Information Seeking Behavior | 385 | 2.414 | $< .001$ | Positive |
| Embedded/Distance Learning Librarianship | 290 | 1.755 | $< .001$ | Positive |
| Public Libraries – Community & Value | 704 | 1.448 | $< .001$ | Positive |
| Academic Librarians & Faculty Status | 636 | 1.389 | $< .001$ | Positive |
| LIS Bibliometrics & Scholarly Communication | 574 | 1.132 | $< .001$ | Positive |
| LIS Education & Practice (Developing Regions) | 701 | 1.045 | $< .001$ | Positive |
| Information Seeking Behavior (General) | 599 | 0.892 | $< .001$ | Positive |
| Library Collections & E-books | 416 | 0.898 | .002 | Positive |
| Information Literacy Instruction | 1,728 | 0.303 | .045 | Positive |
| Digital Libraries & Preservation | 297 | 0.317 | .217 | Inconclusive |
| Health Sciences & Clinical Librarianship | 263 | 0.321 | .250 | Inconclusive |
| Information Retrieval Systems | 261 | −1.008 | .115 | Inconclusive |

This pattern is the opposite of what a conventional confound would produce. If journal environment were confounding the centrality citation relationship, that is, if some journals simultaneously attracted both conceptually central papers and inherently high citation rates for unrelated reasons, then controlling for journal environment should shrink the centrality coefficient, not enlarge it. Instead, accounting for journal environment reveals a stronger relationship than was visible beforehand, a pattern consistent with statistical suppression: variation in venue level citation opportunity was previously obscuring part of the genuine association between a paper's conceptual position and its citation performance, rather than manufacturing a spurious one. This interpretation is corroborated directly rather than inferred solely from the coefficient's movement: conceptual centrality and journal citation environment are themselves essentially uncorrelated ($r = -.037$), which is inconsistent with the alternative explanation that the two measures are simply entangled with each other in a way that distorts both estimates.

The practical implication is that the weak pooled correlation reported at the start of this section is, in a specific and demonstrable sense, an underestimate of the paper level relationship between conceptual position and citation attention, precisely because

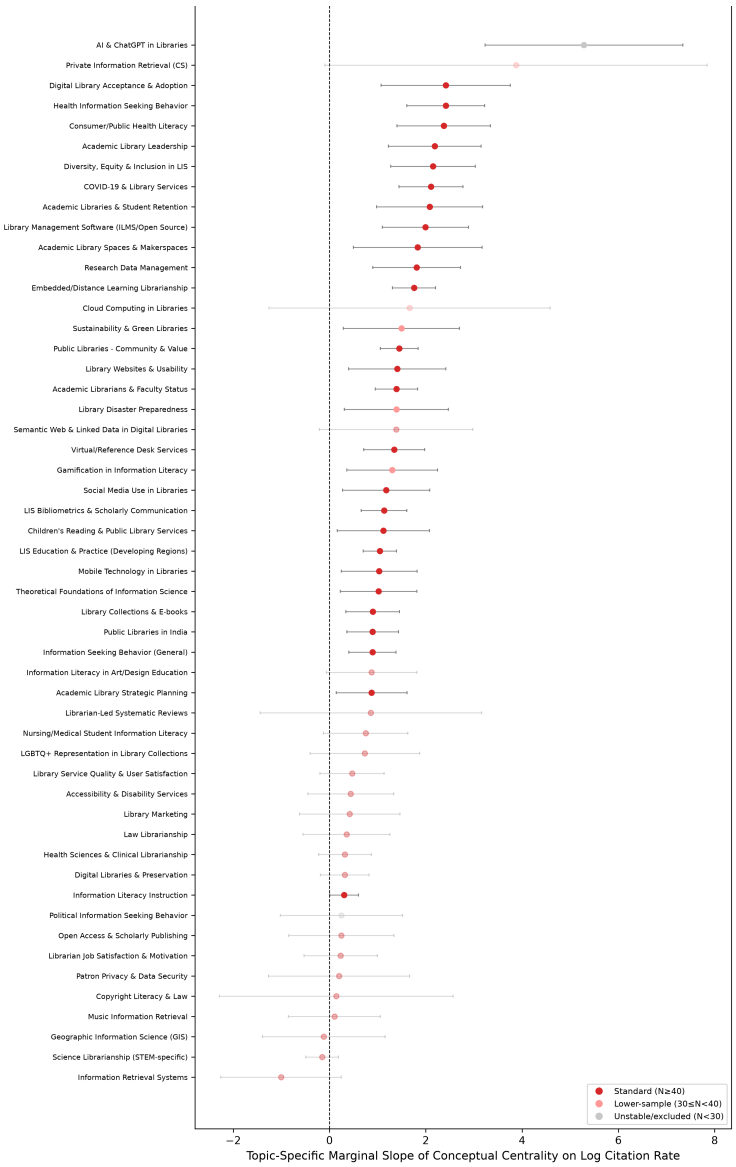



**Fig. 2** Topic-specific marginal slopes of conceptual centrality on log citation rate, with 95% confidence intervals, across all 52 research topics. Markers are colored by regression-sample-size tier (Standard, $N \geq 40$; Lower-sample, $30 \leq N < 40$; Unstable/excluded, $N < 30$) and shaded by statistical significance. The dashed line marks zero.

it does not hold venue level citation opportunity constant. A weak unconditional correlation and a materially stronger conditional relationship are not in tension; the difference between them is the empirical content of this finding.

Having established that centrality's association survives and strengthens under controls, a separate and equally important question is how large that association is in practical terms relative to the discipline's other known citation correlates. Because predictors in the fully specified model are measured on different natural scales, this comparison requires standardized coefficients rather than the raw coefficients discussed above (Figure 3, Panel B). Journal citation environment showed a standardized association with citation rate roughly three times larger than conceptual centrality ($\beta = .312$ versus $\beta = .095$). This is an important corrective to any reading of the preceding paragraphs as suggesting centrality rivals venue in importance: it does not. What the standardized comparison shows instead is a clear ordering, with journal environment dominant, conceptual centrality a distant but still statistically robust second, and international collaboration, review status, author count, and publication year clustered together in a materially smaller, roughly comparable tier below both. Conceptual centrality therefore occupies a genuine but modest place in the discipline's citation landscape: too small to be mistaken for the primary driver of citation performance, too large and too robust across specifications to be dismissed as negligible.

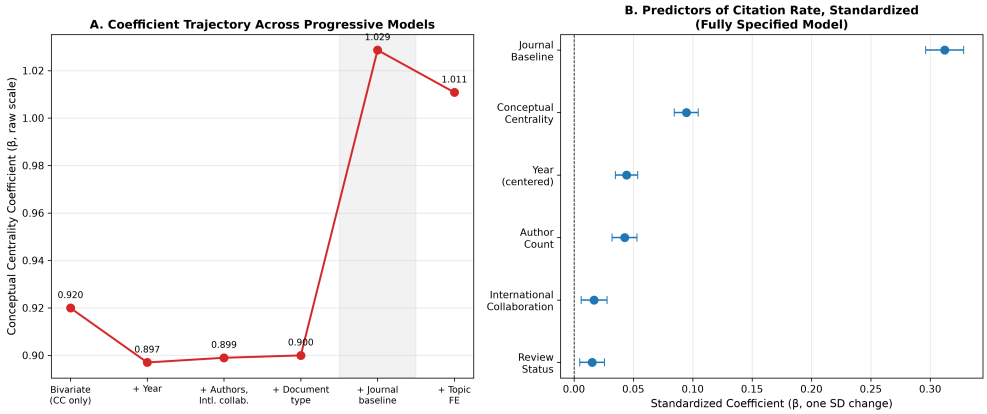



**Fig. 3** (A) Trajectory of the conceptual centrality coefficient across progressively controlled regression models. (B) Standardized coefficients for all predictors in the fully specified model, with 95% confidence intervals, permitting direct comparison of relative practical magnitude.

## 3.5 Does the topic-dependent relationship persist when centrality is based on antecedent literature?

A natural objection to every result reported so far is that conceptual centrality, as constructed, could partly reflect a paper's own influence on the field that came after it, since the topic centroid a paper is compared against includes papers published later, some of which the focal paper may itself have shaped. This would make the observed association at least partly circular: a paper's centrality could be inflated by citing

descendants that used its own vocabulary, rather than reflecting the paper's position relative to a genuinely prior conceptual landscape.

The prospective analysis directly tests this possibility by restricting the centroid used to calculate centrality to papers published strictly before the focal paper. On a fixed exposure subsample of papers aged four to five years, corresponding to a 2021 to 2022 publication cohort, the topic interaction test remained highly significant, and if anything numerically stronger than the retrospective analysis ($F(51, 1945) = 4.79$, $p = 4.73 \times 10^{-25}$, $N = 2{,}055$, against $F = 4.15$ in the main analysis). This is a meaningful result precisely because it goes in the opposite direction from what the circularity objection would predict: if retrospective self influence were inflating the main finding, the prospective, antecedent only version should show a weaker pattern, not a comparable or stronger one. Its persistence indicates that the topic dependent structure documented above reflects a paper's position relative to a genuinely pre existing body of work, not an artifact of measuring centrality against a centroid the paper itself helped shape.

This conclusion carries one explicit qualification. Topic assignments themselves were generated once from the complete 2010 to 2024 corpus rather than reconstructed recursively at each point in time, so while the reference centroid used to calculate prospective centrality is genuinely restricted to earlier papers, the topical categories into which papers were sorted in the first place could in principle still reflect later developments in the field. The prospective check therefore rules out one specific and important form of circularity without ruling out every conceivable one, a distinction addressed further in the Discussion.

## 3.6 Which papers receive substantially more or less citation attention than their conceptual position prediction?

The regression analysis establishes that conceptual centrality relates to citation rate on average, conditional on topic; it does not by itself say anything about individual papers whose citation performance departs sharply from that average pattern. Standardized residuals from the fully specified topic interaction model were used to identify such cases.

At the primary threshold of 1.5 residual standard deviations, 853 papers (7.7 percent) were classified as high positive residual attention outliers, cited substantially more than their conceptual position and publication environment predicted, while 355 papers (3.2 percent) were classified as high negative residual cases, cited substantially less than predicted (Table 4).

**Table 4** Sensitivity of mismatch classification to the standardized-residual threshold

| Threshold | Attention outliers | High negative-residual cases | % of sample |
|---|---|---|---|
| ±1.0 SD | 1,609 | 1,349 | 26.7% |
| ±1.5 SD (primary) | 853 | 355 | 10.9% |
| ±2.0 SD | 455 | 103 | 5.0% |

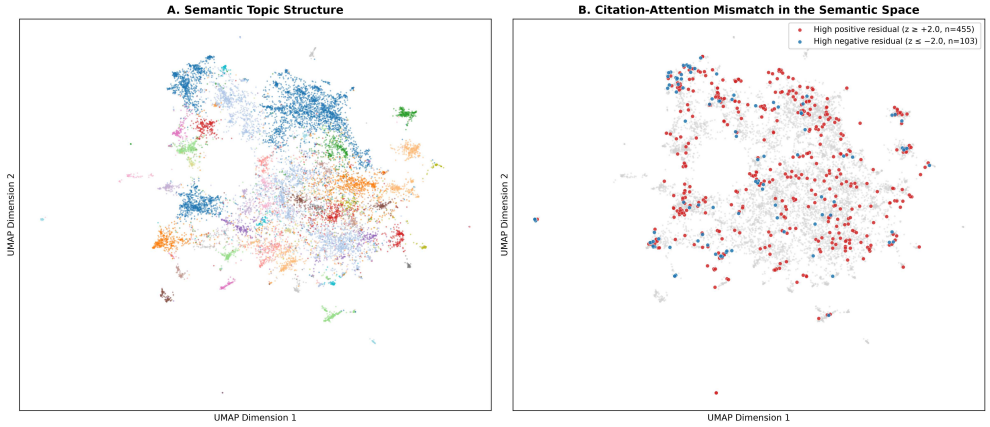



**Fig. 4** (A) Two-dimensional projection of the corpus embedding space, colored by topic assignment, illustrating the corpus's semantic clustering structure. Because more topics exist than are visually distinguishable by color, some unrelated topics share similar hues; this panel establishes that distinct semantic clustering exists, not that every color identifies a unique topic. (B) The same projection with papers showing high positive residuals ($z \geq +2.0$) and high negative residuals ($z \leq -2.0$) highlighted against the background corpus. The projection is provided for visualization only; conceptual centrality and citation residuals were computed in the original, unprojected embedding and regression space, and spatial proximity in this two-dimensional layout should not be interpreted as evidence for either measure.

The ratio of over performing to under performing papers widens considerably at stricter thresholds, from roughly 1.2 to 1 at ±1.0 SD to nearly 4.4 to 1 at ±2.0 SD. This is not, on its own, evidence that citation over performance is intrinsically more common than under performance as a feature of scholarly recognition; it is more parsimoniously explained as an inherited property of the citation count distribution itself, which is bounded at zero on the low end but unbounded on the high end, so that residuals from a model of log citation rate will mechanically show a longer, thinner right tail than left tail as the threshold tightens. This distinction matters for how the asymmetry should be interpreted and reported.

Review type publications were modestly but significantly overrepresented among attention outliers relative to the full sample (7.7 percent versus 5.6 percent, $\chi^2 = 9.56$, $p = .008$). This is a real, statistically confirmed pattern consistent with the general citation function of review articles as broad synthesizing entry points into a literature, a function plausibly somewhat independent of a review's own conceptual centrality. At the same time, the effect size is modest and the majority of attention outlier papers were not reviews, so review status functions as a partial, contributing explanation rather than the dominant mechanism behind citation over performance in this corpus.

The high negative residual cases, papers that are conceptually central to their topic yet cited less than that centrality would predict, are in some respects the more theoretically interesting group, since they represent the clearest evidence that conceptual embeddedness and citation recognition are genuinely separable rather than merely imperfectly correlated. Qualitative inspection of these cases within specific

topics identified two distinct, nameable patterns rather than a single unifying explanation. Within Digital Library Acceptance and Adoption, essentially all identified high negative residual cases explicitly invoked technology acceptance or adoption framework terminology, despite most showing high conceptual centrality in the 0.70 to 0.88 range (Figure 5). This suggests a specific mechanism: centrality in this topic is driven substantially by shared adherence to a common theoretical framework rather than by conceptual originality or distinctiveness, and individual applications of that shared framework, often single site or single institution case studies, do not automatically inherit broad citation reach simply because they use the field's dominant vocabulary.

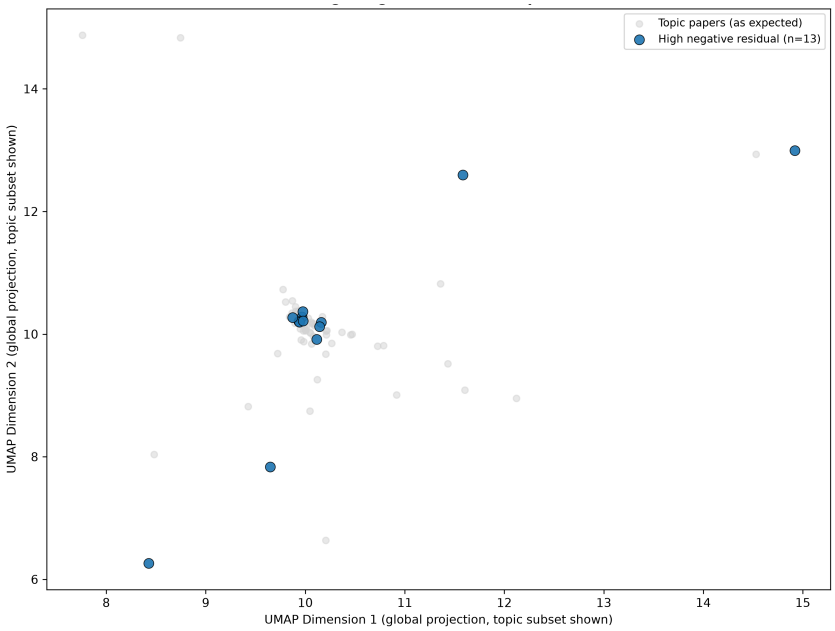



**Fig. 5** Papers within the Digital Library Acceptance & Adoption topic classified as high negative-residual cases (blue), shown against the topic's full paper set (gray).

Within Consumer and Public Health Literacy, a different but related pattern involved a long running serialized report series describing regional trends in health science librarianship, whose individual installments were markedly under cited relative to their strong thematic and stylistic similarity to one another. Here the mechanism is plausibly about audience fragmentation rather than framework sharing: a reader interested in this series is likely to cite only the specific regional installment relevant to their own work, spreading citation attention thinly across many conceptually similar entries rather than concentrating it on any one.

Both patterns illustrate the same underlying point from different angles: high conceptual centrality, as measured here, can arise from genuine substantive importance to a topic's core problems, but it can equally arise from surface level similarity, whether through a shared theoretical template or a shared publication format, that carries

no automatic citation premium. This is an important interpretive boundary on the centrality construct itself, and is treated as such in the Discussion rather than as a limitation to be explained away.

### 3.7 Are the principal findings robust to alternative methodological choices?

The principal findings were stable across the sensitivity analyses conducted. Recomputing conceptual centrality using only the 10,779 documents originally and confidently classified by HDBSCAN, while excluding rather than reassigning documents initially left unclassified, produced a comparable aggregate association ($r_s = .095$, $p < .001$) to the primary estimate ($r_s = .107$, $p < .001$), indicating that the reported association is not driven by the outlier reassignment step in topic construction. The prospective centrality-citation correlation remained stable across antecedent thresholds of 10, 20, and 30 minimum papers ($r_s = .083$, .076, and .071, respectively; all $p < .001$), indicating that the specific threshold choice does not drive the result. The alternative MPNet embedding model produced a substantively comparable topic structure to the primary MiniLM solution (mean best-match Jaccard = .714), indicating that the topic-level findings are not an artifact of a single embedding architecture. Taken together, these checks indicate that the principal findings are robust to the tested methodological alternatives, including outlier handling, the antecedent threshold, and the embedding model.

## 4 Discussion

### 4.1 Conceptual centrality and citation recognition appear related but distinct

The central takeaway is not that conceptual centrality determines citation impact. The evidence is more consistent with a partial coupling interpretation. Papers that are more embedded in the conceptual configuration of their topic tend, on average, to receive somewhat more citation attention, but semantic embeddedness accounts for only a modest share of citation variation. The modest size of the standardized association is itself informative. It suggests that conceptual position contributes information beyond conventional bibliometric and journal-level variables, without displacing them or claiming a larger role than the data support. This interpretation is consistent with the broader scientometric literature, in which citation counts are understood as multifactorial indicators shaped by characteristics of the document, authorship, collaboration, publication venue, and scholarly communication [1–4].

This distinction helps separate two processes that are often blurred in impact analysis. Conceptual centrality describes where a paper sits within a semantic knowledge structure at a given point in time. Citation attention records what happens afterward, as other researchers decide which contributions to use, discuss, extend, or cite. This distinction is consistent with the view that scientific communities organize research around shared conceptual frameworks, problems, and exemplars [17, 19], while cited

works can function as concept symbols that connect citation practices with the conceptual organization of scientific knowledge [20]. At the same time, citation accumulation is shaped by processes of scholarly recognition and cumulative advantage that are not reducible to conceptual position alone [21, 22]. These processes clearly overlap, but they are not the same. A central paper can be broadly foundational without being singled out for citation, while a peripheral paper can become highly visible because it solves a specific methodological or practical problem. Semantic centrality should therefore not be treated as a substitute for citation impact. Its more modest, but still useful, contribution is to add a different descriptive dimension to how scholarly recognition is understood.

## 4.2 Why the relationship may vary across topics

The most consequential finding is the heterogeneity of the centrality-citation relationship. The aggregate coefficient indicates a single, weak association, but the interaction model suggests that this average conceals materially different topic-level patterns. In some research communities, semantic embeddedness tracks citation recognition relatively closely, while in others the relationship is weak or statistically indeterminate.

The present analysis identifies where the relationship differs, but not the mechanisms responsible for those differences. The discussion below should therefore be read with this limitation in mind. One plausible explanation, although the present data cannot adjudicate it directly, is that topics differ in the extent to which citation accumulation depends on a shared conceptual vocabulary. In more cumulative areas, later papers may repeatedly draw on established constructs, making conceptual centrality more consequential for recognition. In other areas, citation attention may be generated more by methodological novelty, application, interdisciplinary reach, or external audiences than by proximity to the local conceptual core. The study does not establish which of these candidate mechanisms produces the observed slope differences. Audience composition, topic maturity, methodological dependence, interdisciplinarity, and citation intent were not directly measured. The empirical contribution is therefore best understood as identifying topic-conditioned heterogeneity, with the underlying mechanisms left as an agenda for future research rather than as questions this study can settle.

## 4.3 Topic size does not seem to be a sufficient explanation

The absence of a relationship with topic size narrows, but does not fully resolve, the interpretation of this heterogeneity. Smaller topics naturally produce less precise slope estimates, which is why a minimum sample-size rule was adopted for individual interpretation in the first place. Precision and substantive slope magnitude are nevertheless different questions, and the weighted meta-regression provided no indication that larger or smaller topics systematically produce stronger or weaker associations. This finding supports retaining topic as a substantive conditioning structure in future analyses rather than treating it simply as a source of sampling noise. It does not,

however, establish that topic identity is a causal mechanism. Topic labels summarize complex research environments, and identifying which specific properties of those environments alter citation recognition would require data not available here.

## 4.4 Journal environment and conceptual position seem to capture different things

The regression sequence also clarifies how semantic position relates to venue-level citation opportunity. Journal citation environment has a substantially larger standardized association than conceptual centrality (.312 versus .095), so the paper should not be read as claiming that semantic position rivals journal environment in importance. Instead, the two variables appear to capture different, only weakly correlated dimensions. Journal environment reflects citation conditions associated with a publication venue, whereas conceptual centrality reflects a paper's semantic position within its topic. The modest increase in the centrality coefficient after journal environment is added is consistent with a mild suppression pattern, in which variation in venue-level citation opportunity may have partly obscured the conditional association between semantic position and citation performance. This pattern should not be treated as a general conclusion. It is offered as one plausible interpretation of the regression sequence shown in Figure 3, Panel A.

## 4.5 Prospective centrality and the temporal interpretation of semantic position

The antecedent-only analysis provides some support for interpreting centrality as a property of a paper's position relative to an existing knowledge structure rather than merely as a retrospective description of where the paper eventually sits. The persistence of the topic-dependent pattern, in broad outline, when the centroid is constructed from earlier publications suggests that the main moderation pattern does not depend entirely on the focal paper's own publication shaping its reference point.

The scope of this conclusion is deliberately limited. Topic assignments were retained from the full-corpus model rather than reconstructed recursively at each point in time. The prospective analysis therefore removes one direct source of circularity without establishing that every part of the semantic representation is free of later information. Its value lies in serving as a targeted validity check rather than as a complete reconstruction of how the field's knowledge structure evolved over time.

## 4.6 Understanding citation centrality mismatch

The residual analysis shifts attention from average relationships to individual departures from expectation, and this part of the study is more exploratory than the regression results presented above. Attention outliers show that citation attention can exceed what semantic position and the measured publication environment would predict. Review articles are modestly overrepresented in this group, consistent with their synthesis function and generally higher citation rates, although review status explains only a small share of the outlier population and clearly cannot serve as a general explanation on its own.

Overlooked core cases are, if anything, more interesting because they are harder to explain with the variables available to the model. A paper can be deeply embedded in a topic's semantic structure without receiving the citation attention that its embeddedness might lead one to expect. The two illustrative patterns identified in the mismatch analysis, shared methodological frameworks that produce semantic similarity without making each individual application uniquely citable, and serialized reports that may divide citation attention across installments, are offered as candidate explanations for a subset of cases rather than as general mechanisms. These cases are useful mainly because they make the study's central distinction more concrete. High conceptual centrality should not be interpreted as evidence of realized scholarly influence, just as low centrality should not be interpreted as evidence of irrelevance.

# 5 Implications

## 5.1 Implications for scientometric theory

Taken as a whole, the study fits a multidimensional account of scholarly recognition in which citation indicators and semantic indicators capture different stages and properties of the research process. Citation counts reflect subsequent recognition and use, whereas conceptual centrality describes a paper's semantic position relative to its topic. Neither, on its own, provides a complete measure of scholarly value. Scholarly recognition may therefore be better understood as multidimensional rather than reduced to a single citation-based measure. Conceptual embeddedness and citation attention appear empirically related but conceptually distinct, and the citation consequences of semantic position seem to depend on research-topic context, at least within this corpus.

## 5.2 Implications for bibliometric methodology

The results carry a modest implication for model specification. When a semantic characteristic is expected to operate differently across research communities, a single discipline-wide coefficient may not provide an adequate summary, and topic-conditioned interactions can reveal structure that would otherwise disappear in aggregate analysis. This seems particularly relevant in a heterogeneous field such as LIS, where library science, information behavior, information retrieval, scholarly communication, health information, and computational topics plausibly operate under somewhat different epistemic and citation regimes. Confirming whether this pattern generalizes beyond LIS would require applying the same design in other disciplines. More generally, leave-one-out semantic centroids offer a way to avoid mechanical self-inclusion when measuring topic-relative centrality. Topic-by-predictor interactions are worth considering whenever a discipline contains meaningfully heterogeneous research communities. Antecedent-only semantic measures provide a useful, although partial, temporal robustness check against one form of retrospective construction. Residual-based mismatch analysis can also complement average-effect models, cautiously, by identifying papers whose observed recognition differs substantially from model expectations.

### 5.3 Implications for research evaluation

Conceptual centrality is not recommended as a replacement for citation indicators. Its value lies in distinguishing semantic position from realized citation recognition, allowing an evaluation framework to ask two separate questions. How embedded is a contribution within an existing conceptual structure, and how much subsequent scholarly attention did it receive? A research evaluation process could, in principle, keep these questions separate without assuming that the two measures are interchangeable.

### 5.4 Implications for citation culture and scholarly recognition

Beyond individual evaluation exercises, the findings speak in a limited way to how citation counts function as a currency within academic culture more broadly. Citation counts are convenient precisely because they reduce recognition to a single number, but the analysis suggests that this number can represent different processes depending on where, topically, a paper sits. In some research communities, a high citation count may plausibly track conceptual centrality fairly closely, whereas in others it may reflect methodological utility, timeliness, or reach into an adjacent audience. This does not argue for abandoning citation counts as a recognition signal, since they remain informative. It does, however, suggest caution when treating citation counts as though they had the same meaning across topical contexts, particularly in cross-disciplinary or cross-topic comparisons within a single evaluation exercise.

### 5.5 Implications for scholarly discovery

The mismatch framework points to a more targeted, and admittedly more speculative, use in scholarly discovery. Rather than ranking papers only by citation counts, analysts could identify papers whose citation trajectories appear unusually high or low relative to their semantic position and publication environment, and treat these papers as candidates for qualitative review or closer reading. Such an approach would function as a screening aid rather than as an automated judgment about a paper's quality, influence, or neglect. It could plausibly support research synthesis or expert review, particularly when the goal is to surface atypical recognition trajectories rather than simply rank highly cited papers. This is a reasonable direction to explore, but the present study establishes only that such mismatches exist and are somewhat interpretable in a limited number of cases. It does not establish that a screening tool based on this framework would perform well in practice.

## 6 Limitations and Future Research

The study is observational and cannot establish that conceptual centrality causes citation attention. The regression coefficients throughout should therefore be interpreted as conditional associations. Longitudinal designs, explicit out-of-sample prediction, or quasi-experimental settings would be needed to support stronger causal or predictive claims.

The prospective analysis uses antecedent publications to construct the centrality reference point but retains topic assignments from the full-corpus model. A fully

dynamic topic model estimated recursively through time would provide a stronger test of how conceptual position changes as topics themselves develop, something the present design cannot fully capture.

Semantic centrality depends on the geometry induced by the particular embedding model used. The MPNet robustness check suggests that the broad semantic organization of the corpus is reasonably stable across embedding choices, but alternative architectures, domain-specific models, or full-text rather than title-and-abstract representations could still shift fine-grained centrality estimates in ways this study cannot rule out.

Citation rate remains an imperfect proxy for scholarly recognition. Citation intent, self-citation, citation context, negative citations, and non-citation forms of scholarly use are not modeled here. Combining conceptual centrality with citation-context classification or with other indicators of scholarly use therefore represents a natural next step.

The review-status finding in the mismatch analysis has reasonably direct statistical support, but the explanations offered for overlooked core cases are topic-specific qualitative observations drawn from a small number of examples. These mechanisms would need to be tested more systematically across a wider set of topics and disciplines before being treated as general patterns rather than suggestive leads.

# 7 Conclusion

This study asked whether a paper's position within the conceptual structure of its research topic bears any relationship to the citation attention it subsequently receives. The evidence supports a fairly nuanced answer. Conceptual centrality is positively associated with citation performance at the aggregate level, but that association is weak when LIS is treated as a single, undifferentiated population. Once publication and journal level characteristics are taken into account the association remains evident, and the topic specific analysis suggests its strength varies meaningfully across research areas.

The central contribution is therefore not a claim that conceptual centrality is a powerful, general predictor of citation impact; it plainly is not. The more defensible claim is that semantic embeddedness constitutes a distinct, if modest, dimension of citation recognition, and that its relationship with citation attention appears to depend on topical context. Topic size does not seem to explain the observed slope gradient, and the prospective analysis suggests, with appropriate caveats, that this topic dependent structure holds even when the centrality reference point is restricted to antecedent publications.

The mismatch analysis reinforces why this distinction seems worth preserving. Some papers receive noticeably more citation attention than their semantic and publication characteristics would predict, while others receive noticeably less. Conceptual position, in other words, cannot be treated as a stand in for influence, and citation count cannot be assumed to capture everything meaningful about a paper's place in the knowledge structure it belongs to.

The broader suggestion, offered tentatively, is that scientometric models might do well to distinguish between semantic embeddedness and realized scholarly recognition as related but separable properties. The more useful question going forward may not be whether conceptual centrality determines citation impact, since the present evidence suggests it does not, in any strong sense, but under what topical and publication conditions a contribution's position within a field's conceptual structure tends to translate into citation attention, and under what conditions it does not.

**Acknowledgements.** NA

## Declarations

- Funding This research received no specific grant from any funding agency in the public, commercial, or not-for-profit sectors.
- Conflict of interest/Competing interests (check journal-specific guidelines for which heading to use) The authors declare that they have no known competing financial interests or personal relationships that could have influenced the work reported in this paper.
- Ethics approval and consent to participate This study relied exclusively on publicly available, previously published bibliographic and citation metadata retrieved from Scopus. No human participants, animal subjects, or personal or identifiable data were involved, and no ethics approval was required.
- Consent for publication Not applicable.
- Data availability The bibliographic data used in this study were retrieved from Scopus (Elsevier) under a licensed institutional subscription and cannot be publicly redistributed due to third-party licensing restrictions. Derived data (topic assignments, centrality scores, and analysis code) will be made available by the corresponding author upon reasonable request.
- Materials availability NA
- Code availability NA
- Author contribution K.S. conceptualized the study, conducted the analysis, developed the methodology, prepared the main draft, conducted the literature review and wrote the manuscript. M.S. supervised the project and contributed to copyediting and critical revision of the manuscript. All authors read and approved the final manuscript.

If any of the sections are not relevant to your manuscript, please include the heading and write 'Not applicable' for that section.

## References

[1] Aksnes, D.W.: Citation rates and perceptions of scientific contribution. Journal of the American Society for Information Science and Technology **57**(2), 169–185 (2006) https://doi.org/10.1002/asi.20262

[2] Bornmann, L., Daniel, H.-D.: What do citation counts measure? a review of studies on citing behavior. Journal of Documentation **64**(1), 45–80 (2008) https://doi.org/10.1108/00220410810844150

[3] Tahamtan, I., Afshar, A.S., Ahamdzadeh, K.: Factors affecting number of citations: A comprehensive review of the literature. Scientometrics **107**(3), 1195–1225 (2016) https://doi.org/10.1007/s11192-016-1889-2

[4] Didegah, F., Thelwall, M.: Which factors help authors produce the highest impact research? collaboration, journal and document properties. Journal of Informetrics **7**(4), 861–873 (2013) https://doi.org/10.1016/j.joi.2013.08.006

[5] Bornmann, L., Schier, H., Marx, W., Daniel, H.-D.: What factors determine citation counts of publications in chemistry besides their quality? Journal of Informetrics **6**(1), 11–18 (2012) https://doi.org/10.1016/j.joi.2011.08.004

[6] Reimers, N., Gurevych, I.: Sentence-bert: Sentence embeddings using siamese bert-networks. In: Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing, pp. 3982–3992 (2019). https://doi.org/10.18653/v1/D19-1410

[7] Grootendorst, M.: Bertopic: Neural topic modeling with a class-based tf-idf procedure. arXiv preprint arXiv:2203.05794 (2022) https://doi.org/10.48550/arXiv.2203.05794

[8] Jiang, Z., Lin, T., Huang, C.: Deep representation learning of scientific paper reveals its potential scholarly impact. Journal of Informetrics **17**(1), 101376 (2023) https://doi.org/10.1016/j.joi.2023.101376

[9] Vital Junior, A., Silva, F.N., Oliveira Junior, O.N.d., Amancio, D.R.: Predicting citation impact of research papers using gpt and other text embeddings. Physica A: Statistical Mechanics and its Applications **674**, 130789 (2025) https://doi.org/10.1016/j.physa.2025.130789

[10] Suzen, N., Gorban, A., Levesley, J., Mirkes, E.: Predicting the impact of scientific articles through semantic analysis of abstracts. Scientometrics (2026) https://doi.org/10.1007/s11192-026-05647-5

[11] Liu, L., Rong, Y.: Beyond citation-based metrics: Measuring interdisciplinarity via sbert semantic embeddings and its heterogeneous effects on citation impact. PLOS ONE **21**(7), 0354129 (2026) https://doi.org/10.1371/journal.pone.0354129

[12] Freeman, L.C.: Centrality in social networks: Conceptual clarification. Social Networks **1**(3), 215–239 (1978) https://doi.org/10.1016/0378-8733(78)90021-7

[13] Small, H.: Co-citation in the scientific literature: A new measure of the relationship between two documents. Journal of the American Society for Information Science **24**(4), 265–269 (1973) https://doi.org/10.1002/asi.4630240406

[14] Small, H., Griffith, B.C.: The structure of scientific literatures i: Identifying and graphing specialties. Science Studies **4**(1), 17–40 (1974) https://doi.org/10.1177/030631277400400102

[15] Ding, Y.: Topic-based pagerank on author cocitation networks. Journal of the American Society for Information Science and Technology **62**(3), 449–466 (2011) https://doi.org/10.1002/asi.21467

[16] Zhang, Y., Ma, J., Wang, Z., Chen, B., Yu, Y.: Collective topical pagerank: A model to evaluate the topic-dependent academic impact of scientific papers. Scientometrics **114**(3), 1345–1372 (2018) https://doi.org/10.1007/s11192-017-2626-1

[17] Kuhn, T.S.: The Structure of Scientific Revolutions. University of Chicago Press, Chicago (1962)

[18] Price, D.J.d.S.: Little Science, Big Science. Columbia University Press, New York (1963). https://doi.org/10.7312/pric91844

[19] Whitley, R.: The Intellectual and Social Organization of the Sciences. Oxford University Press, Oxford (2000). https://doi.org/10.1093/oso/9780199240531.001.0001

[20] Small, H.G.: Cited documents as concept symbols. Social Studies of Science **8**(3), 327–340 (1978) https://doi.org/10.1177/030631277800800305

[21] Merton, R.K.: The matthew effect in science: The reward and communication systems of science are considered. Science **159**(3810), 56–63 (1968) https://doi.org/10.1126/science.159.3810.56

[22] Merton, R.K.: The matthew effect in science, ii: Cumulative advantage and the symbolism of intellectual property. Isis **79**(4), 606–623 (1988) https://doi.org/10.1086/354848

[23] Moravcsik, M.J., Murugesan, P.: Some results on the function and quality of citations. Social Studies of Science **5**(1), 86–92 (1975) https://doi.org/10.1177/030631277500500106

[24] Case, D.O., Higgins, G.M.: How can we investigate citation behavior? a study of reasons for citing literature in communication. Journal of the American Society for Information Science **51**(7), 635–645 (2000) https://doi.org/10.1002/(SICI)1097-4571(2000)51:7⟨635::AID-ASI6⟩3.0.CO;2-H

[25] Zhang, F., Wu, S.: Predicting citation impact of academic papers across research

areas using multiple models and early citations. Scientometrics **129**, 4137–4166 (2024) https://doi.org/10.1007/s11192-024-05086-0

[26] Kalita, A., Sumit, K., Singh, M., et al.: Does Open Access Matter for Citations? A Topic-Stratified Analysis of Library and Information Science Discipline. Research Square. Preprint, Version 1, 30 August 2026 (2026). https://doi.org/10.21203/rs.3.rs-10855106/v1

[27] McInnes, L., Healy, J., Melville, J.: Umap: Uniform manifold approximation and projection. Journal of Open Source Software **3**(29), 861 (2018) https://doi.org/10.21105/joss.00861

[28] Campello, R.J.G.B., Moulavi, D., Zimek, A., Sander, J.: Hierarchical density estimates for data clustering, visualization, and outlier detection. ACM Transactions on Knowledge Discovery from Data **10**(1), 5–1551 (2015) https://doi.org/10.1145/2733381

[29] Cohan, A., Feldman, S., Beltagy, I., Downey, D., Weld, D.S.: Specter: Document-level representation learning using citation-informed transformers. In: Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, pp. 2270–2282 (2020). https://doi.org/10.18653/v1/2020.acl-main.207