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.

Before Measuring Epistemic Change: The Dependence of Scientific Antecedent Retrieval on Text Representation, Retrieval Models, and Temporal Windows

Mardochée Ogécime

mardocheeogecime@gmail.com

Universidade Federal de Minas Gerais

Research Article

Keywords: epistemic change, novelty indicators, bibliometric validity, scientific antecedents, semantic retrieval, BM25

Posted Date: November 19th, 2026

DOI: https://doi.org/10.21203/rs.3.rs-11460349/v1

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

Additional Declarations: No competing interests reported.

Before Measuring Epistemic Change: The Dependence of Scientific Antecedent Retrieval on Text Representation, Retrieval Models, and Temporal Windows

Mardochée Ogécime[1,2*]

[1*]School of Information Science, Federal University of Minas Gerais (UFMG), Belo Horizonte, Minas Gerais, Brazil.
[2]Department of Computing, Institute of Exact and Biological Sciences, Federal University of Ouro Preto (UFOP), Ouro Preto, Minas Gerais, Brazil.

Corresponding author(s). E-mail(s): mardocheeogecime@gmail.com;

Abstract

Purpose. Measures of scientific novelty and epistemic change frequently operationalize a document's relationship to prior knowledge. Yet the identification of that prior knowledge is itself a measurement operation whose robustness to documentary representation, retrieval architecture, and temporal eligibility is rarely made explicit. This study examines these dependencies rather than treating retrieved antecedents as an invariant reference set. Methods. We investigated a harmonized corpus of 993,086 Brazilian postgraduate theses and dissertations from 2013–2024. A second-stage balanced sample contained 108 focal documents, twelve from each of nine broad disciplinary areas, dated 2016–2024. For each focal, eligible antecedents were earlier documents within the same harmonized specific disciplinary area. We constructed four retrieval configurations: BM25 using titles and keywords or titles, abstracts, and keywords, and two corresponding multilingual sentence-embedding representations (M2 and M4). We compared complete available historical windows with rolling three-year windows, assessing Top 10/30/50 overlap, uniform-selection chance references, stratified bootstrap intervals, and semantic-score geometry. A separate year-matched reference design compared same-area and other-area similarities. Results. Across 108 focal documents, the semantic experiment evaluated 1,640,271 focal–antecedent pairs in the historical window and 756,979 in the three-year window. In the historical Top 30, overlap was 20.25% between M2 and M4 and 22.41% between short- and full-text

BM25. Agreement between text-aligned lexical and semantic configurations was 29.17% (BM25 short–M2) and 26.98% (BM25 full–M4), compared with 17.07% and 16.20% for crossed-field comparisons. The average paired advantage of text-aligned over crossed comparisons was 11.44 percentage points (95% stratified bootstrap interval, 10.19–12.67). Excluding 2016, for which the available historical and three-year windows coincide by construction, temporal Top-30 stability ranged from 51.31% for short BM25 to 55.28% for M4. Year-matched same-area minus other-area cosine contrasts were 0.01196 for M2 and 0.01425 for M4. Conclusions. Identification of potential scientific antecedents is sensitive to seemingly upstream measurement decisions. The findings establish operational uncertainty in the construction of reference sets for novelty-oriented bibliometrics, but neither demonstrate influence relationships nor validate the epistemic relevance of retrieved documents. Robustness analysis of antecedent retrieval should precede substantive interpretation of similarity-derived evidence of epistemic change.

Keywords: epistemic change; novelty indicators; bibliometric validity; scientific antecedents; semantic retrieval; BM25

1 Introduction

Epistemic change refers not merely to fluctuations in publication volume or the occurrence of unusual combinations of terms, but to transformations in the problems, concepts, methods, and knowledge claims through which research communities produce and develop knowledge. Bibliometric research has sought observable traces of such transformations in citation relations, emerging topics, recombinations, originality, and disruption (Small et al., 2014; Rotolo et al., 2015; Shibayama & Wang, 2020; Park et al., 2023). These operationalizations enable large-scale comparison, while raising the question of how much of the underlying epistemic process their observable signals actually capture.

The distinction between epistemic processes and single-document events is particularly important. A document may introduce a proposition, method, or combination, but the significance and uptake of that contribution may become visible only through subsequent research trajectories. In addition, an observed difference between a focal document and its predecessors depends on which predecessors were selected for comparison. The construction of the reference set is therefore not a neutral preprocessing detail. It is part of the measurement model.

Consider a focal thesis whose novelty is assessed through its semantic distance from preceding theses and dissertations. One retrieval pipeline searches for lexical correspondence in titles and keywords; another uses fuller text metadata; a third represents similar fields through dense embeddings; and a fourth restricts the historical horizon. Each specification could select a different set of documents against which the focal work appears close, distant, conventional, or unusual. Consequently, a change in the indicator might reflect a change in the research contribution, a change in the antecedent set, or both. Without an assessment of reference-set stability, these possibilities are difficult to distinguish.

This issue is connected to a broader validity debate. Work on scientific topic reconstruction has stressed the need to distinguish operational consistency from correspondence with knowledge structures meaningful to researchers (Held et al., 2021). Related research on originality has combined bibliometric operationalization with independent assessment, illustrating the added value of criterion-related validation (Shibayama & Wang, 2020). The methodological implication is straightforward: agreement between computational systems is not equivalent to validity, but substantial disagreement can reveal measurement dependence that requires explanation.

We therefore investigate the antecedent-identification step itself. Rather than proposing another composite novelty score, we ask how sensitive the recovered antecedents are to choices that frequently precede indicator calculation: textual field coverage, retrieval mechanism, and temporal reference window. Our empirical setting is the Brazilian postgraduate corpus of theses and dissertations, whose heterogeneous disciplinary composition enables a controlled stress test across research domains.

Four questions guide the analysis: (RQ1) How strongly does text representation change top-ranked potential antecedents? (RQ2) How does lexical–semantic concordance depend on matched versus crossed textual fields? (RQ3) How sensitive are the rankings to the historical reference window? and (RQ4) How do the distributions and disciplinary contrasts of semantic similarities condition interpretation of these rankings? We intentionally delimit the estimand to **operational retrieval stability**, not epistemic transformation itself.

2 Conceptual framework: antecedents as a measurement object

2.1 Three distinct relations

Three relations should be analytically separated. **Temporal precedence** establishes that a candidate document appeared before a focal document. **Computational proximity** indicates similarity according to a specified lexical or embedding-based score. **Epistemic antecedence** would require a substantively meaningful relation in research problems, conceptual developments, methodological dependence, or knowledge claims. The first is an eligibility rule; the second is an observed property of a model; the third is an interpretive and potentially historical claim. None logically entails the next.

This separation matters because novelty-oriented measures often define a document relative to its nearest historical neighbors or to an inferred prior frontier. If the identity of those neighbors is contingent on input fields and retrieval rules, the resulting novelty signal may incorporate an unreported source of uncertainty. Operational robustness cannot establish criterion validity, but it can identify fragile measurement choices before stronger claims are made.

2.2 Field composition and retrieval paradigm

Titles and keywords typically emphasize compact descriptions and selected descriptors, whereas abstracts add objectives, theoretical framing, methodological details, and contextual information. Their inclusion can alter both which documents appear similar

and the textual cues driving their ranking. Sparse lexical retrieval assigns weight to shared words according to their frequencies and document statistics; dense retrieval maps texts into learned semantic representations. These approaches differ in what constitutes a match. Their outputs are most interpretable in a controlled comparison when the information fields are held constant.

Accordingly, two contrasts are conceptually distinct: within-paradigm changes from shorter to fuller field coverage, and between-paradigm changes from lexical to embedding-based retrieval using aligned fields. Crossed-field comparisons intentionally change both aspects and should not be interpreted as isolated algorithmic effects.

2.3 Reference windows and historical truncation

A complete observed history and a recent three-year frontier correspond to different hypotheses about relevant preceding knowledge. Long windows permit older antecedents; short windows focus on near-term research activity. However, a corpus beginning in 2013 is left-truncated: for focal documents in 2016, the available full history (2013–2015) exactly equals the preceding three years. This structural identity must be excluded from substantive comparisons of temporal-window stability.

2.4 Three levels of validity and a conditional reference-set problem

Our design distinguishes **procedural reproducibility**, **operational robustness**, and **criterion-related epistemic validity**. Procedural reproducibility asks whether the same pipeline and archived inputs recover the same rankings; cross-stage audits can evaluate this property. Operational robustness asks whether results are stable across defensible choices of fields, retrieval architecture, and temporal eligibility. Epistemic validity asks whether the retrieved relations correspond to substantive predecessors in the development of scientific problems, concepts, or techniques; this requires evidence independent of ranking agreement. Success at one level does not demonstrate success at the next.

This distinction exposes a conditional measurement problem. Let $R(\mathrm{d}; \mathrm{f}, \mathrm{m}, \mathrm{w})$ denote the ranked antecedent set of document d obtained with field representation f, retrieval mechanism m, and temporal window w. A novelty statistic based on proximity to predecessors can be written generically as $N(\mathrm{d} \mid R(\mathrm{d}; \mathrm{f}, \mathrm{m}, \mathrm{w}))$. Consequently, the apparent novelty of the same document may vary even while the focal document is held fixed, because the reference relation has changed. We do **not** estimate this induced variability in downstream novelty scores here; instead, we measure the instability of the antecedent sets on which such scores could depend. This is a sensitivity analysis of the **measurement substrate**, not direct measurement of epistemic change.

The conceptual contrast with event-based measures is therefore qualified. We replace neither one-paper novelty scores nor their empirical validation with a complete model of epistemic processes. Rather, we demonstrate that the candidate relational environment against which a document is evaluated is specification dependent. A process-oriented extension would additionally require following the subsequent use, transformation, and recontextualization of knowledge claims over sequences of publications.

3 Data and methods

3.1 Corpus and analytical sample

The source was a harmonized longitudinal collection of **993,086 distinct Brazilian postgraduate theses and dissertations dated 2013–2024**, derived from CAPES data. Document identifiers, year, disciplinary classification, and available titles, abstracts, and keywords were harmonized before retrieval. The corpus size describes the source collection, not the number of candidate documents involved in every focal-level comparison.

An initial year-stratified master sample of 1,440 documents was constructed for methodological development. A subsequent balanced analytical sample selected **108 focal documents** dated 2016–2024, with twelve in each of nine broad CAPES areas, using additional area-year restrictions (at most two focals per broad area-year). This second-stage selection was designed for cross-area methodological comparison and is **not a self-weighted probability sample of national outputs**. Estimates and bootstrap intervals describe variability in the balanced analytical sample; they should not be interpreted as population-representative national effects.

For each focal document, candidate antecedents were restricted to documents from the same harmonized specific CAPES disciplinary area whose year strictly preceded that of the focal. Two candidate universes were specified: all eligible years available since 2013 ("historical"), and years $t-3$ through $t-1$ ("frontier3y"). Across all 108 focals, there were **1,640,271 historical focal–candidate comparisons** and **756,979 frontier3y comparisons**. Candidate counts varied by focal and window.

3.2 Textual representations and retrieval systems

Four ranked retrieval configurations were compared. **BM25_short** represented both focal queries and candidate documents with titles and keywords. **BM25_full** used titles, abstracts, and keywords. **M2** and **M4** used corresponding previously computed multilingual-E5-based embeddings, respectively. The embedding representations contained 768 dimensions in the archived outputs; cosine similarities were computed using the frozen document vectors. The precise embedding checkpoint, preprocessing prompts, and truncation configuration must be verified from the upstream embedding-generation provenance before submission rather than inferred solely from embedding dimensionality.

The lexical corpus contained **642,779 distinct documents required for the specified focal and historical candidate universes**; it was not a new or competing estimate of the entire source corpus. Text underwent Unicode normalization, case folding, and tokenization on the normalized whitespace-delimited content, without stemming or stopword filtering. BM25 was computed independently for each focal-specific eligible candidate corpus and each field configuration, with $k1 = 1.5$ and $b = 0.75$, positive smoothed inverse-document-frequency weights $\log[1 + (N - df + 0.5)/(df + 0.5)]$, and unique query terms. Thus BM25 numerical scores are corpus-dependent and not directly comparable with cosine values or across the two candidate windows.

Table 1 Controlled retrieval configurations.

Configuration	Documentary fields	Ranking score	Reference corpus / specification
BM25_short	Title and keywords	BM25 ($k_1 = 1.5$, $b = 0.75$)	IDF fitted to each eligible focal-specific corpus
BM25_full	Title, abstract, keywords	BM25 (same parameters)	Window- and focal-specific IDF
M2	Title and keywords	Cosine similarity (768-dimensional E5 vectors)	Frozen embeddings; identical candidate universe
M4	Title, abstract, keywords	Cosine similarity (768-dimensional E5 vectors)	Frozen embeddings; identical candidate universe

All four systems ranked the **same candidate set within each focal and window**. The lexical experiment produced 432 top-50 rankings (108 focals × two windows × two textual configurations); the semantic rankings were reconstructed from the archived full-score tables. No selected BM25 top-50 entry had a zero score. A cross-stage audit reproduced all **1,296** previously archived M2/M4 stability values with exactly zero discrepancy.

Note. The labels "short" and "full" refer to the available bibliographic textual fields, **not** the full text of the thesis or dissertation. Embedding-model checkpoint and encoding details require independent provenance verification before submission.

3.3 Retrieval concordance

For two ranked lists A and B of depth k, set-overlap was defined as

$$O_k(A, B) = \frac{\left|\mathrm{Top}_k(A) \cap \mathrm{Top}_k(B)\right|}{k}, \qquad k \in \{10, 30, 50\}.$$

This measure describes agreement on identity, not rank-order correlation, criterion validity, or a causal effect of retrieval design. We compared six method pairs within each reference window and examined historical–frontier concordance separately for each of the four configurations. We also computed the Jaccard index descriptively. For a simple benchmark, we used the expected overlap k/N obtained from two independent uniform selections of k documents from a candidate universe of size N. The associated adjusted overlap was $(O_k - k/N)/(1 - k/N)$. This benchmark is not a plausible null model of trained retrieval systems, but contextualizes the magnitude of observed overlap relative to uniform random selection.

To quantify variability across the analytical sample, **3,000 bootstrap draws** resampled focal documents within each of the nine broad-area strata while preserving the stratum sizes. Paired contrasts were computed within the same focal documents. The intervals represent exploratory sample-based uncertainty, not design-based confidence intervals for the full CAPES population. Unless otherwise indicated, the primary reporting depth was Top 30, with Top 10 and Top 50 available as sensitivity analyses.

3.4 Semantic-score geometry and disciplinary reference contrast

To avoid conflating high absolute cosine similarities with epistemic equivalence, we examined within-focal score distributions, including mean, median, upper percentiles, and extrema. Separately, for each of the 108 focals, 200 antecedents from the same specific disciplinary area and 200 from other specific areas were sampled under exact antecedent-year matching, giving **43,200 scored reference pairs**. Same-minus-other contrasts were computed separately for M2 and M4, with paired, broad-area-stratified bootstrap intervals. These are discriminative diagnostics, not independent ground-truth judgments of research influence.

3.5 Reproducibility and scope of validation

The workflow preserved focal-level score checkpoints, top-ranked candidate lists, provenance, and cross-stage auditing files. Following repeated integrity checks, the final quantitative comparison was frozen after the robustness stage. A separate exploratory protocol for blinded human judgments was prepared using 18 focal documents (two per broad area), 108 selected focal–antecedent pairs, and two independent evaluation instruments. **No human ratings were available at the time of this analysis**, and no human-validation findings are reported. This distinction is essential: a prepared assessment protocol cannot substitute for an observed validation result.

4 Results

4.1 Text-field sensitivity within retrieval paradigms

Within the historical window, M2 and M4 shared an average of **20.25**% of their Top-30 candidates (95% bootstrap interval [17.93%, 22.72%]). BM25_short and BM25_full shared **22.41**% ([19.94%, 25.16%]). In the three-year frontier, the corresponding overlaps were **21.20**% ([18.98%, 23.52%]) and **23.21**% ([20.86%, 25.74%]). Consequently, expanding the documentary representation from title-and-keywords to title-abstract-keywords replaced approximately four in five items of the Top 30 in either retrieval paradigm. This is substantial configuration dependence even before comparing lexical with semantic systems.

The within-paradigm difference between BM25 field configurations (22.41%) and M2 versus M4 (20.25%) was **2.16 percentage points** in the historical window (95% paired bootstrap interval [−0.25, 4.51]); the interval included zero. A similar contrast in frontier3y was **2.01 points** ([−0.40, 4.41]). There is therefore no sufficiently precise evidence, on this comparison alone, that field expansion leads to greater Top-30 agreement within either paradigm.

Note. Intervals are the 2.5th and 97.5th percentiles of 3,000 bootstrap replicates stratified by the nine broad CAPES areas. They quantify uncertainty in the balanced analytical sample, not design-based national population inference. All method contrasts use identical focal-specific candidate universes within each temporal window.

Table 2 Top-30 set overlap by retrieval configuration and temporal window.

Comparison	Historical: mean (95% CI)	Three-year: mean (95% CI)
BM25_short × M2 (aligned)	0.2917 (0.2627–0.3210)	0.2932 (0.2660–0.3216)
BM25_full × M4 (aligned)	0.2698 (0.2441–0.2966)	0.2648 (0.2423–0.2886)
BM25_short × BM25_full	0.2241 (0.1994–0.2516)	0.2321 (0.2086–0.2574)
M2 × M4	0.2025 (0.1793–0.2272)	0.2120 (0.1898–0.2352)
BM25_full × M2 (crossed)	0.1707 (0.1491–0.1941)	0.1701 (0.1503–0.1907)
BM25_short × M4 (crossed)	0.1620 (0.1398–0.1861)	0.1651 (0.1444–0.1864)

4.2 Agreement between lexical and semantic retrieval

Lexical–semantic agreement was higher when the configurations received matched textual fields. In the historical window, **BM25_short versus M2** yielded **29.17**% Top-30 overlap ([26.27%, 32.10%]); **BM25_full versus M4** yielded **26.98**% ([24.41%, 29.66%]). Crossed-field comparisons yielded **17.07**% for BM25_full–M2 ([14.91%, 19.41%]) and **16.20**% for BM25_short–M4 ([13.98%, 18.61%]). The analogous matched-field results under frontier3y were 29.32% and 26.48%, respectively.

A focal-paired comparison of the average matched-field overlap against the average crossed-field overlap showed an advantage of **11.44 percentage points** in the historical window (95% interval [10.19, 12.67]) and **11.14 points** in frontier3y ([9.98, 12.28]). Field alignment thus materially improves agreement when contrasting lexical and semantic architectures. However, even the aligned comparisons share fewer than nine of the Top-30 antecedents on average. Shared input fields do not make the two retrieval paradigms interchangeable.

All observed overlaps were much larger than the expected uniform-selection benchmark. For example, in the historical Top 30, the mean k/N reference was **0.00325**, compared with observed overlap values ranging from **0.16204** to **0.29167** across method pairs. The chance-adjusted historical overlap was **0.28942** for BM25_short–M2 and **0.19993** for M2–M4. The descriptive gap is large relative to the uniform-selection reference, but is not a formal independence test and does not demonstrate which retrieval strategy is more epistemically relevant.

4.3 Temporal-window sensitivity

Across all 108 focals, historical–frontier Top-30 agreement was 57.62% for BM25_short, 58.55% for BM25_full, 60.83% for M2, and 61.08% for M4. However, fourteen focal documents dated 2016 had identical observed historical and three-year universes because the source corpus began in 2013. Including these cases mechanically inflates temporal stability.

After excluding 2016, the **94 focals from 2017–2024** showed Top-30 agreement of **51.31**% for BM25_short (95% interval [47.80%, 54.68%]), **52.38**% for BM25_full

Table 3 Temporal top-30 overlap excluding the structurally identical 2016 windows (94 focal documents).

Method	Historical vs 3Y: mean (95% CI)	Difference from lexical counterpart (percentage points; 95% CI)
BM25_short	0.5131 (0.4780–0.5468)	Reference for M2
M2	0.5500 (0.5106–0.5872)	+3.69 (+1.45 to +5.82)
BM25_full	0.5238 (0.4886–0.5592)	Reference for M4
M4	0.5528 (0.5156–0.5890)	+2.91 (+0.74 to +5.18)

([48.86%, 55.92%]), **55.00**% for M2 ([51.06%, 58.72%]), and **55.28**% for M4 ([51.56%, 58.90%]). Paired semantic-minus-lexical differences were **3.69 percentage points** for M2 minus BM25_short ([1.45, 5.82]) and **2.91 points** for M4 minus BM25_full ([0.74, 5.18]). The corresponding within-paradigm differences, M4–M2 and BM25_full–BM25_short, had intervals crossing zero.

These estimates support a modest advantage of embedding-based configurations in temporal *set stability* within the sampled focals. They do not imply higher scientific relevance. Furthermore, changing a BM25 temporal window changes not only the eligible candidates but also corpus-dependent inverse-document-frequency statistics; the temporal contrast is accordingly a pipeline-level robustness comparison, not an isolated causal estimate of candidate-window width.

Note. Ninety-four focal documents from 2017–2024; 14 focals dated 2016 were excluded because observed full historical and three-year candidate universes coincide by construction. Temporal agreement combines changes in eligible documents with corpus-dependent IDF re-estimation in BM25.

4.4 Semantic similarity geometry and disciplinary discrimination

Absolute cosine similarity was high across candidate distributions. In the historical window, mean focal-level cosine values were approximately **0.87967** for M2 and **0.92037** for M4; median values were approximately **0.88003** and **0.92159**, respectively. The corresponding upper 99th percentiles were **0.91090** and **0.94513**, and observed maxima were **0.94002** and **0.96174**. These descriptive differences should not be used to argue that M4 is intrinsically more accurate than M2 because raw scores across representations are not calibrated on a common validity scale.

The year-matched disciplinary reference test found higher cosine similarity for same-area than other-area documents. The mean same-minus-other contrast was **0.01196** for M2 (95% interval [0.01087, 0.01306]) and **0.01425** for M4 ([0.01292, 0.01561]). The difference between these contrasts was **0.00229** ([0.00138, 0.00322]). Standardized contrasts were **0.87439** for M2 and **1.02402** for M4, a difference of **0.14963** ([0.07774, 0.22158]). Hence the embeddings contain a measurable disciplinary signal despite the

Table 4 Cosine geometry and disciplinary discrimination in the historical reference window.

Diagnostic	M2	M4
Mean cosine across eligible pairs	0.87967	0.92037
Median cosine	0.88003	0.92159
99th percentile cosine	0.91090	0.94513
Maximum cosine	0.94002	0.96174
Same-area minus other-area cosine	0.01196	0.01425
95% bootstrap CI for area contrast	0.01087–0.01306	0.01292–0.01561
Standardized area contrast	0.87439	1.02402

concentration of their absolute cosine values. Disciplinary separability is not equivalent to recovery of epistemically meaningful antecedents.

Note. The disciplinary reference analysis uses 200 same-specific-area and 200 other-specific-area documents per focal, matched on antecedent publication year (43,200 pairs in total). Cosines across M2/M4 are not calibrated as interchangeable scales. The paired between-representation difference in the raw disciplinary contrast is +0.00229 (95% interval 0.00138–0.00322), not a criterion-validity estimate.

5 Discussion

5.1 Antecedent selection is part of the measurement model

The central result is that a purportedly preliminary step—identifying prior documents for comparison—is far from invariant. Even when all methods draw on the same temporally eligible disciplinary candidate universe, changing textual fields or retrieval paradigms markedly changes which documents appear among the most similar. For novelty measures that depend on similarity to preceding documents, these differences can propagate into the apparent location of a focal work relative to a prior knowledge frontier. The exact downstream propagation is a hypothesis implied by the measurement architecture; it was **not directly tested here through a reconstructed novelty index**.

The distinction is important in the context of epistemic-change research. A stable reference set is not automatically epistemically valid; conversely, a changing reference set is not automatically erroneous. Different retrieval architectures may capture complementary dimensions of proximity. The evidence instead establishes that the operational definition of "the prior literature most related to a focal document" requires explicit justification and sensitivity analysis.

5.2 The role of metadata and field composition

The markedly higher agreement between text-aligned lexical and semantic retrieval strengthens the argument that field composition must be treated as an analytical decision, not incidental data availability. Titles and keywords select a different informational view from fuller representations that include abstracts. The findings therefore offer direct empirical evidence concerning the special collection's question of how much

signal about scientific content can be extracted from bibliographic metadata: the signal is structured differently under different representations.

Nevertheless, retrieval overlap does not quantify *how much epistemic meaning* each field contains. That question requires independent criteria—expert judgments, verified intellectual dependencies, or other substantively justified reference data. It cannot be answered by selecting the model with a larger cosine, stronger pairwise overlap, or modestly more stable rankings.

5.3 The temporal frontier is a substantive assumption

Short and complete historical windows operationalize different notions of relevant knowledge. A three-year reference universe captures a near-term frontier, whereas a longer observed history permits older sources and enduring conceptual traditions. The reduced concordance after removing 2016 shows that temporal eligibility meaningfully changes the candidate set in later years. Restricting the time horizon should therefore be defended with reference to the temporal dynamics of the domain, rather than treated as an arbitrary computational parameter.

In this corpus, year availability begins in 2013 and coverage of earlier intellectual histories is absent. For fields with long citation half-lives or enduring foundational works, a short historical frontier may underrepresent highly relevant conceptual antecedents. Conversely, a long window may disproportionately retrieve shared generic vocabulary. Our analysis shows instability but cannot decide which substantive interpretation is appropriate for every disciplinary area.

5.4 Implications for bibliometric indicators of epistemic change

The special collection calls attention to the fact that many existing indicators compress epistemic change into a single scalar at the level of one document, often emphasizing degree over type. Our findings address a logically prior requirement: the reference relations that support document-level novelty claims must be operationally examined. Three methodological recommendations follow from the evidence. First, report input field composition and retrieval model as part of the indicator specification. Second, examine reference-window sensitivity, carefully distinguishing structural overlap caused by corpus truncation. Third, treat retrieval-system agreement and score geometry as robustness diagnostics while reserving validity claims for independent assessments of substantive knowledge relations.

This is a contribution to the **methodology of measuring epistemic change**, not itself a complete measure of change as a process. A full process-oriented research design would need to follow sequences of documents, conceptual or methodological developments, and subsequent uptake, possibly integrating citations and domain-specific expert evidence. Potential antecedents offer one component of such a design, not a substitute for it.

5.5 Operational robustness is not criterion-related validation

The present article estimates operational robustness under specified and reproducible retrieval conditions, not criterion-related validity. Although a separate blinded human

assessment protocol was prepared, it had generated no ratings at the analytical freeze. Accordingly, no inter-rater reliability, expert relevance scores, precision, or recall estimates are reported. The study is not described as a validation experiment that has been completed; its evidentiary claim is narrower and explicit.

This is not a reason to disregard validation; rather, it prevents a category error between *stability* and *truth*. Future work should ask domain-qualified reviewers whether pairs are merely thematically related, conceptually or methodologically connected, or plausible as meaningful research antecedents. Because the pilot pairs were deliberately stratified by retrieval profile, their judgments would support an exploratory comparison rather than a national precision or recall estimate without a redesigned probability sample and appropriate weighting.

5.6 Toward an empirically grounded validation design

The operational findings motivate a subsequent two-level validation strategy. At the first level, independently blinded judges can assess whether individual focal–antecedent pairs exhibit topical, conceptual, or methodological relevance, with separate indication of insufficient textual evidence. Sampling should cover both retrieval-specific and jointly retrieved pairs, but agreement from a purposive pilot must not be presented as population precision. At the second level, verified traces of intellectual uptake—such as explicit conceptual appropriation, methodological transmission, or later research trajectories—are needed to test claims about *epistemic change as process*. This second level goes beyond matching abstracts or inferring influence from same-area membership.

These are proposals, not observations in the present dataset. Crucially, an eventual evaluation should compare the contribution of retrieved antecedents to well-defined scientific tasks, not merely reward agreement with the most familiar vocabulary. This follows from the distinction between a thematically similar document and a scientifically consequential antecedent.

6 Limitations

Several constraints govern interpretation. First, the 108 focals form a balanced, second-stage analytical sample, not a national probability sample with design-based inferential weights. Second, eligible antecedents are limited to preceding documents within the same harmonized specific CAPES disciplinary area, excluding potentially crucial cross-disciplinary influences. Third, the observation window begins in 2013; historical measures are therefore left-truncated, and the coincidence of temporal windows for 2016 is a structural artifact. Fourth, the study analyzes thesis and dissertation metadata rather than full research texts, and quality and completeness of abstracts and keywords vary. Fifth, BM25 and dense similarities measure different functions and their raw scores are not commensurate. Sixth, the simple uniform chance benchmark is descriptive and is not a model of retrieval-system independence. Seventh, neither citation pathways nor qualitative evidence of historical intellectual influence were established. Finally, no completed blinded human judgments are available to test the criterion validity of retrieved antecedents. None of these limitations invalidates the operational-stability

findings, but together they restrict their extension to epistemic validity or population-level claims.

7 Conclusion

Potential scientific antecedents are not an invariant input to bibliometric measurement. Across Brazilian postgraduate documents, substantial changes in the top-ranked historical reference set arose from differences in text-field composition, lexical versus semantic retrieval, and temporal eligibility rules. Agreement was systematically higher when compared retrieval systems used matched textual fields, and embedding-based approaches exhibited modestly greater historical-window stability in this balanced analytical sample. Yet neither agreement nor stability establishes that retrieved documents are epistemically relevant.

The contribution is therefore a measurement principle: **before treating proximity to prior research as evidence of novelty or epistemic change, the robustness and validity of the antecedent-identification procedure must be investigated explicitly**. This refocuses attention from the numerical novelty score alone to the representational and relational assumptions that make the score possible.

Declarations

Funding.

This research received no specific funding.

Competing interests.

The author declares no competing interests.

Author contributions.

Mardochée Ogécime: conceptualization, methodology, software, formal analysis, data curation, investigation, writing—original draft, and writing—review and editing.

Ethics and consent.

The reported research is a secondary analysis of bibliographic metadata. No participants were recruited, no human assessments were performed or analyzed, and no identifiable human-subject data are reported. Consent to participate and consent for publication are not applicable.

Data availability.

The underlying Brazilian CAPES data can be obtained subject to the original provider's access and reuse terms through https://dadosabertos.capes.gov.br/. Derived records, intermediate embeddings, and checkpoints are not necessarily available in the linked code notebook; no unrestricted replication-data release is claimed.

Code availability.

Analytical notebook snapshot: https :// github .com / mardocheeogecime -gif / Rede _Colabora -o _Acad -mica _ECI _UFMG / blob / 4677686d6b053d80507446457de3ac3183962fb3 / LAB _BIBLIOMETRIC _MEASURES_OF_EPISTEMIC_CHANGE.ipynb. Interactive environment: https:// colab.research.google.com/drive/1Fyo5YZrR6hQtTJ87hZ3sw8TOATyUnRdH?usp= sharing. The ability to rerun the analyses depends on separately acquired CAPES records and generated checkpoints.

Acknowledgements.

None.

ORCID.

Mardochée Ogécime: https://orcid.org/0000-0001-7846-6941.

References

Held, M., Laudel, G., & Gläser, J. (2021). Challenges to the validity of topic reconstruction. *Scientometrics, 126*, 4511–4536. https://doi.org/10.1007/s11192-021-03920-3

Leydesdorff, L., & Rafols, I. (2011). Local emergence and global diffusion of research technologies: An exploration of patterns of network formation. *Journal of the American Society for Information Science and Technology, 62* (5), 846–860.

McCain, K. W. (2010). Core journal literatures and persistent research themes in an emerging interdisciplinary field: Exploring the literature of evolutionary developmental biology. *Journal of Informetrics, 4*, 157–165.

Park, M., Leahey, E., & Funk, R. J. (2023). Papers and patents are becoming less disruptive over time. *Nature, 613*, 138–144.

Rotolo, D., Hicks, D., & Martin, B. R. (2015). What is an emerging technology? *Research Policy, 44* (10), 1827–1843.

Shibayama, S., & Wang, J. (2020). Measuring originality in science. *Scientometrics, 122*, 409–427. https://doi.org/10.1007/s11192-019-03263-0

Small, H., Boyack, K. W., & Klavans, R. (2014). Identifying emerging topics in science and technology. *Research Policy, 43* (8), 1450–1467.

Wang, J., Veugelers, R., & Stephan, P. (2017). Bias against novelty in science: A cautionary tale for users of bibliometric indicators. *Research Policy, 46* (8), 1416–1436.

Wang, L., Yang, N., Huang, X., Yang, L., Majumder, R., & Wei, F. (2024). Multilingual E5 text embeddings: A technical report. *arXiv*, 2402.05672. https://arxiv.org/abs/2402.05672

Supplementary Files

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

- ScientometricsTables1to4.docx
- supplementaryinformation.tex
- supplementaryinformation.pdf