# Supplementary Information

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

Mardochée Ogécime

## S1. Scope and linked materials

This document describes the completed computational robustness analysis and its interpretive boundaries. The study does not include completed blinded human relevance evaluations. The full notebook is referenced as Supplementary Code 1, but it is not a self-contained executable replication of all historical data and embedding stages.
GitHub: `https://github.com/mardocheeogecime-gif/Rede_Colabora-o_Acad-mica_ECI_UFMG/blob/4677686d6b053d80507446457de3ac3183962fb3/LAB_BIBLIOMETRIC_MEASURES_OF_EPISTEMIC_CHANGE.ipynb`.
Colab: `https://colab.research.google.com/drive/1Fyo5YZrR6hQtTJ87hZ3sw8TOATyUnRdH?usp=sharing`.

## S2. National corpus and sampling scope

The harmonized CAPES corpus contains 993,086 Brazilian postgraduate theses and dissertations (2013–2024). The controlled ranking benchmark uses 108 focal documents, balanced with 12 observations in each of nine broad areas. It comprises 1,640,271 historical and 756,979 rolling-three-year focal–antecedent similarity pairs. These are not independent observations in a probability sample of the national document population. Eligible antecedents are restricted to preceding documents in the same specific CAPES area.

## S3. Retrieval settings and checks

Four conditions were evaluated: BM25 over title plus keywords, BM25 over title plus abstract plus keywords, and the corresponding M2/M4 frozen multilingual-E5 embeddings. BM25 uses $k_1 = 1.5$, $b = 0.75$ and a positive inverse document frequency $\log(1+(N-\mathrm{df}+0.5)/(\mathrm{df}+0.5))$. Each focal/window has its own lexical reference corpus. Top 10, 30, and 50 are compared by intersection size divided by $k$. The national BM25 benchmark comprises 432 rankings, 21,600 Top-50 records and no zero-score documents inside the Top 50. Exact archived embedding checkpoint, chunking, prefix and truncation specifications must be verified against the original embedding-generation experiment before asserting full reproducibility.

## S4. Results summary

Matched minus crossed lexical–semantic Top-30 agreement is 0.1144 (95% stratified bootstrap interval 0.1019–0.1267) in the full historical window. After excluding 2016, temporal Top-30 overlap is 0.5131 for BM25 short, 0.5238 for BM25 full, 0.5500 for M2, and 0.5528 for M4 (94 focal documents).

Table 1: Mean Top-30 overlap in the historical window.

| Comparison | Overlap |
|---|---|
| BM25 short vs M2 | 0.2917 |
| BM25 full vs M4 | 0.2698 |
| BM25 short vs full | 0.2241 |
| M2 vs M4 | 0.2025 |
| BM25 full vs M2 | 0.1707 |
| BM25 short vs M4 | 0.1620 |

## S5. Operational and epistemic validity

These diagnostics assess measurement robustness rather than a human-validated ground truth. Sampling of candidate pairs for blinded review (108 pairs across 18 focals) and preparation of two reviewer instruments do not constitute completed evaluations. No kappa statistics, human relevance labels or evidence of scientific influence are reported.

## S6. Replication and disclosure limits

The notebook uses year-partitioned CAPES records, frozen embeddings, intermediate checkpoints and Drive-specific paths. Publishing this notebook alone does not make those files available. Redistribution of source abstracts and restricted derived records must respect provider and institutional rules. No credentials or raw records are included in this submission package.