Science, Engineering and Technology Vol. 7, No. 1, pp. 1-13
www.setjournal.com https://doi.org/10.54327/set2026/v7.i1.421

Auditing Archive-Size Effects in Transformer-Based Semantic Monitoring of AI Research Literature

Omar Alniemi[1], Alaa Saadoon[1], Saif Alhusseny[2], Hanaa F. Mahmood[1]

[1]*Department of Computer Science, College of Education for Pure Science, University of Mosul, Mosul, Iraq*
[2]*Department of Physics, College of Education for Pure Science, University of Mosul, Mosul, Iraq*

Abstract

Field-level semantic distinctiveness can be estimated by measuring the semantic distance between a focal paper and previous papers. The methodological background comes from previous research on transformer-based document representation, semantic novelty, and topic dynamics, and the challenges of longitudinal nearest-neighbor monitoring arise from the growth of archives, the choice of representation, and the definition of the neighborhood. This study audits these sources of sensitivity in historical semantic distinctiveness. We establish that, for nested prior-paper sets, adding candidates cannot increase the mean distance to the k nearest prior papers. The audit is based on the title-abstract text of a text-audited, relevance-ranked OpenAlex sample of 5,429 artificial intelligence (AI)-related records from 2018–2025. MiniLM, SciBERT, SPECTER and past-only term frequency–inverse document frequency (TF-IDF) representations are evaluated using exact cosine k-nearest-neighbor search (k=5; k=10 sensitivity) under cumulative memory, and MiniLM is also tested under three-year rolling and fixed-capacity memory, while BERTopic-based topic drift and retrieval provenance are explored as a form of exploratory context. The primary comparison is for 2020–2025. Under cumulative memory, endpoint distinctiveness changed by −12.85% for MiniLM, −8.11% for SciBERT, −22.09% for SPECTER, and −2.27% for TF-IDF. A fixed capacity MiniLM memory, by contrast, resulted in a +9.84% endpoint change. This reversal is considered a sensitivity finding, not evidence of a trend in creativity, research quality, or the whole AI field. The study demonstrates that archive dependence needs to be audited prior to the interpretation of semantic distinctiveness as novelty.

Keywords: Artificial Intelligence, Nearest Neighbors, OpenAlex, Semantic Monitoring, Scientific Text Analytics, Transformer Embeddings

1. Introduction

Artificial intelligence (AI) has emerged as a subject of study and a part of scientific infrastructure. AI systems can facilitate literature synthesis, experimental design, coding, model development, and parts of the research cycle [1] and recent evidence even suggests that increased individual visibility may coexist with a narrower collective focus [2]. It is not possible to tell from the number of publications if a field is opening up new semantic space or if it is focusing its efforts on existing paths. The use of AI in science is evaluated in terms of acceleration, accountability, and epistemic risk. Meanwhile, there has been a resurgence of interest in measuring novelty in papers and patents longitudinally [3]. The field of AI is a very informative test domain because of its rapidly evolving methods, application domains and interdisciplinary connections [4], [5] and large language models can also affect the lexical surface

Corresponding author: Omar Alniemi (omaralniemi@uomosul.edu.iq)
Received: 9 July 2026; Revised: 29 October 2026; Accepted: 3 November 2026; Published: 13 November 2026
© 2026 The Author(s). This work is licensed under a Creative Commons Attribution 4.0 International License

of scientific writing [6]. AI is now playing a role in scientific discovery, interdisciplinary research, and scholarly writing in rapidly evolving fields around the world, and the demand for trustworthy semantic monitoring of evolving research directions is growing [1], [4], [5], [6]. Prior-paper archives can greatly affect longitudinal nearest-neighbor distance estimates, depending on the size and composition of the archive. A semantic-change system needs to differentiate between conceptual closeness, vocabulary change, and measurement artifacts. This paper will discuss one measurement threat. A paper-level k-nearest-neighbor (kNN) score depends on the candidate set available for comparison. As a cumulative archive grows, the opportunity to find a close precedent can only increase. If the document-generating process does not change, there can thus be a deterministic archive-size component in an apparent temporal decline in distance. There are existing novelty, topic and science-mapping methods that offer valuable, but differently scoped, evidence. The historical reference sets are not always considered as part of the estimand. However, in the case of longitudinal nearest-neighbor systems, the scope, capacity, geometry of the encoder, neighborhood rule and construction of the corpus need to be specified and tested separately. The contribution is not a new encoder or archive-invariant novelty metric, but an audit framework for archive-size effects in longitudinal semantic-distance measurements.

- It establishes non-increasing behavior of k-nearest-neighbor distance under nested archive expansion for a fixed paper.
- Tests temporal patterns across three transformer representations, a lexical baseline, two neighborhood sizes, corpus repair, and three memory regimes.
- Presents representation-conditional provenance by using nearest-prior-paper retrieval. The empirical study is clearly restricted to the OpenAlex relevance-ranked capped sample. The analysis is additionally constrained by inputs of title–abstract and tested representation and memory configurations.

The study contributes to AI-based scientific text analytics by demonstrating the impact of archive construction on transformer-based semantic monitoring systems. The empirical sensitivity analysis showed that the endpoint change of the MiniLM was −12.85% for cumulative memory and +9.84% for fixed-capacity memory, which highlights that the trajectory measured is sensitive to archive construction. The study is guided by three questions: How strongly do temporal patterns depend on representation and neighborhood choice? What effect does the capacity of the archive have on the measured trajectory? What additional exploratory structure is visible in topic-centroid movement and local retrieval evidence?

2. Related Work

2.1. AI-assisted science and document representation

AI science research infrastructure, collaboration, prior capability, and organizational conditions influence AI adoption [7]. A complementary literature uses dense document representations, which can recover conceptual proximity beyond keywords and citation counts [8]. Sentence-BERT (SBERT) provides efficient cosine-based semantic retrieval [9], MiniLM provides compact transformer representations [10] and scientific-text encoders like SciBERT and SPECTER include domain and citation-informed inductive biases [11], [12], [13]. While these models can be helpful measurement tools, none of them should be considered to be semantic ground truth [14].

Embedding proximity is not self-interpreting. A paper can be influential and be near to known semantic neighborhoods or far away and not be scientifically valuable. For longitudinal analysis to be possible, an explicit unit of comparison, historical boundary, encoder, distance function and neighborhood rule must be specified.

2.2. Semantic distinctiveness, topic dynamics, and measurement validity

Content-sensitive novelty models are based on embeddings, language models, graphs, and knowledge recombination [15]. Together, this literature demonstrates that novelty indicators vary in terms of the analytical unit, temporal baseline, and the target of the validation. The present study therefore employs the more restricted term

historical semantic distinctiveness: Distance from the documents observed earlier, under a stated representation and memory.

Topic modeling is concerned with another construct. The BERTopic algorithm is a fusion of transformer representations, dimensionality reduction, density-based clustering, and class-based term frequency–inverse document frequency (TF-IDF) [16], [17], [18]. Sample-local thematic organization is described by topic prevalence and centroid movement, but cannot validate a document-level distinctiveness score.

Historical distinctiveness and same-year semantic compactness are two distinct measures. Compactness measures same-year local density; historical distinctiveness measures distance from a prior archive. It is possible to have distant papers and dense current clusters in a field sample. If these quantities are kept separate, then a lesser historical distance will not be automatically labeled as convergence.

Convergence research offers a more structural view on the problem of interdisciplinary integration [19] while the operational score used here only allows for a more limited description of the local compactness within the year. More network or field level evidence is needed to substantiate claims of disciplinary convergence [20].

2.3. Comparative synthesis and research gap

The studies reviewed here do not consider archive capacity as a longitudinal measurement variable, but rather look at semantic representation, retrieval/recommendation, and novelty measurement [8], [13], [15]. High-dimensional distance concentration and high-dimensional hubness have also been studied in classical nearest-neighbor research [21], [22]. Beyond archive capacity, corpus construction is also important. Research on semantic-interoperability reveals that heterogeneous data may require syntactic and semantic alignment, and that mapping with the support of an ontology and AI can support this alignment [23], [24]. The principles are applicable to semantic-monitoring pipelines where the scholarly metadata and/or the vocabularies are heterogeneous and need to be harmonized prior to representation and distance analysis. OpenAlex provides open scholarly metadata that differs from proprietary databases in coverage and completeness [25], [26].

The default ranking of its search results is based on a relevance score which is a combination of text similarity and citation count, and in the case of a limit on retrieval per year, the resulting corpus is a non-probability, relevance-ranked sample. The alignment of identifiers, text repair, and clear exclusions are therefore part of measurement validity, not peripheral preprocessing [26], [27].

Data reliability can be enhanced by rule-based natural language processing (NLP) cleaning, and automated processing tools can improve consistency and scalability in complex data pipelines [28], [29]. Although these principles are reflected in the text quality controls and reproducible preprocessing in the present audit, they cannot eliminate selection bias that may have been introduced by relevance-ranked, non-probability sampling.

The threats to validity are summarized in Table 1, along with the controls used in this study and the inferences they permit and the remaining limits. The audit aims to avoid an aggregate temporal line having more meaning than the data and measurement design it represents. Table 2 shows a comparison of the representative prior studies, and places the present archive-audit protocol in the context of the surrounding semantic-analysis literature.

Explicit archive auditing is the methodological gap that is tackled here. There are few longitudinal embedding studies that focus on the monotonic effect of nested candidate sets and at the same time report representation sensitivity and paper-level provenance. The suggested protocol transforms these choices into measurable parts, instead of being wrapped in a single novelty label.

Table 1. Measurement threats, implemented controls, permitted inference, and residual limitations.

Threat	Control implemented	Permitted inference	Residual limitation
Archive capacity	Cumulative, rolling 3-year, and fixed-capacity memories; 250-draw fixed-capacity diagnostic; formal monotonicity result	Raw kNN trajectories are conditional on the declared memory	Fixed capacity isolates cardinality, not age or topic composition
Representation	MiniLM, SciBERT, SPECTER, and TF-IDF	Agreement or disagreement can be reported as sensitivity	Raw scales are not comparable across encoders
Neighborhood and text quality	k=5/k=10; 400-record audit; immutable identifiers	Main pattern is not attributable to one k or flagged-text subset	Abstract-only, English-oriented, relevance-ranked sample
Topic module	Prevalence separated from centroid drift	Exploratory sample-local thematic structure	One BERTopic configuration and 23.41% outliers
Retrieval provenance	Nearest prior papers retrieved per encoder	Scores are inspectable within representation	Retrieval is not causal explanation or construct validation

Table 2. Comparison of representative semantic-analysis studies and the present archive-audit protocol.

Ref.	Corpus / domain	AI or semantic method	Primary analytical objective	Relevance to the present audit
[8]	Scientific abstracts; multidisciplinary Web of Science categories	Meaning Space; Doc2Vec, SBERT, SciBERT	Semantic representation and citation-impact prediction	Shows that representation choice materially shapes scholarly-text semantic analysis.
[13]	Research-paper recommendation	Paper embeddings; network/NLP novelty-diversity framework	Novel and diverse recommendations for interdisciplinary knowledge transfer	Shows that document representation changes downstream recommendation behavior.
[15]	Academic-paper novelty literature	Review of novelty measures, data sources, units, and validation strategies	Synthesize approaches to measuring scholarly novelty	Supports separating historical distinctiveness from broader novelty constructs and auditing design choices.
[20]	Science-of-science; scholarly metadata	Co-occurrence networks with BERTopic	Identify interdisciplinary field emergence	Illustrates field-level topic/network evidence that is distinct from document-level historical distance.
[23]	Healthcare ontologies	Syntactic/semantic similarity and ontology matching	Transform heterogeneous healthcare data toward HL7 FHIR	Motivates explicit semantic alignment before analysis of heterogeneous data.
[24]	Heterogeneous public-policy data	AI/NLP and Semantic Web translation, annotation, mapping, transformation	Enhanced semantic interoperability	Supports semantic harmonization when heterogeneous metadata or vocabularies are integrated.
[26]	16.8M records shared by OpenAlex, Web of Science, and Scopus	Large-scale reference/metadata coverage comparison	Assess database coverage and metadata completeness	Supports treating database coverage and metadata quality as measurement conditions.
Present study	5,429 AI-related OpenAlex records, 2018–2025	MiniLM, SciBERT, SPECTER, past-only TF-IDF; kNN; cumulative/rolling/fixed memories	Audit archive-size sensitivity of longitudinal semantic distinctiveness	Makes archive scope/capacity part of the estimand and adds representation-conditional provenance.

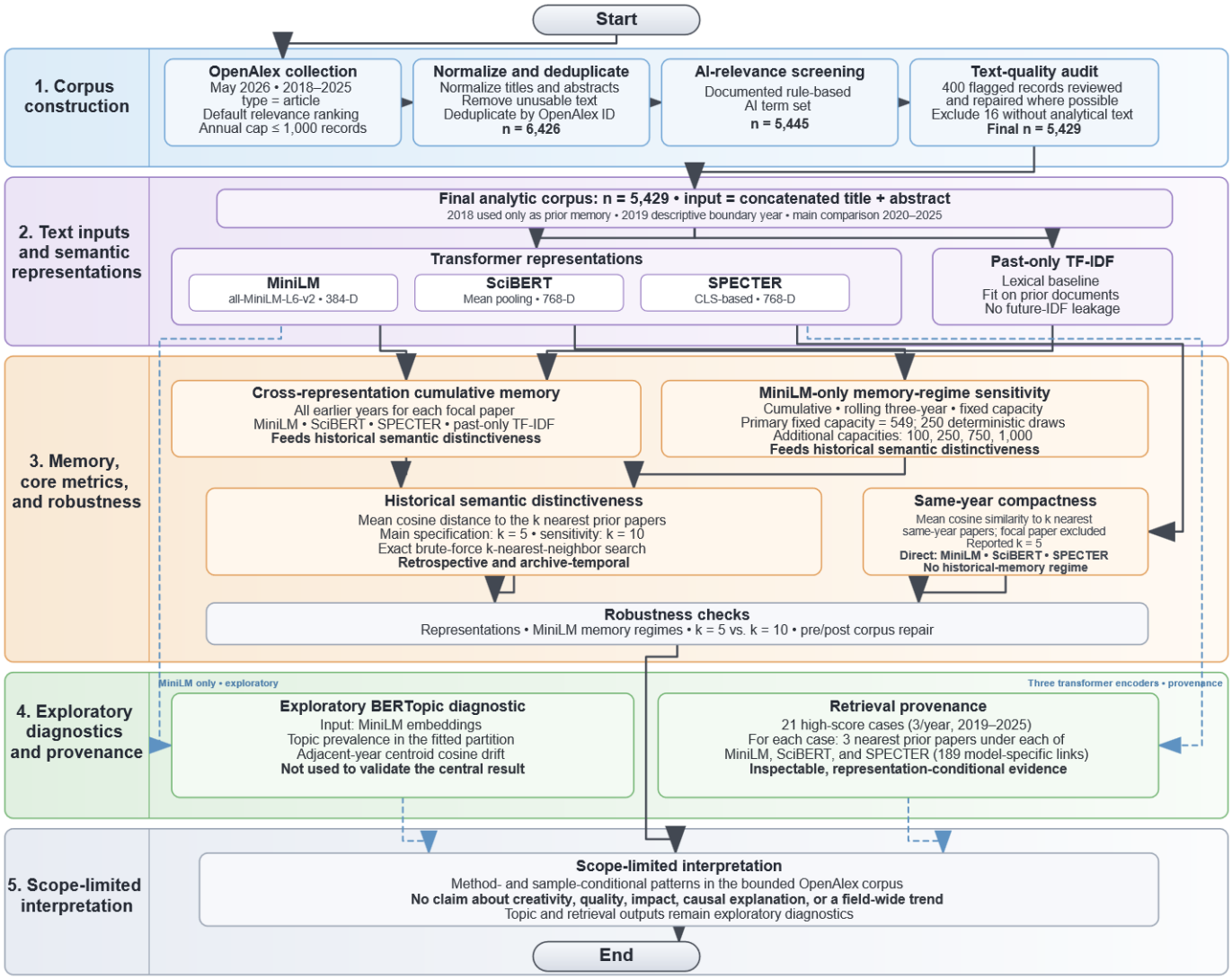


Figure 1. Unified Modeling Language (UML) activity diagram of the proposed semantic-monitoring audit and replication workflow.

3. Methodology

The methodology was designed as an audit of measurement instead of an estimate of the trend at the population level. To create a bounded, relevance-ranked temporal sample, a consistent annual OpenAlex retrieval rule was used; relevance screening and text-quality auditing ensured that semantic representations were calculated from usable, AI-relevant text; representation, neighborhood size and memory construction were evaluated separately, as each of them can affect the historical-distance estimates derived from kNN (Figure 1).

3.1. Corpus construction and text preparation

OpenAlex records of articles from 2018 to 2025 were retrieved in May 2026 with the search string: “artificial intelligence machine learning deep learning neural network transformer large language model”. The OpenAlex Works application programming interface (API) (https://api.openalex.org/works) was used with annual filter: from_publication_date:{YEAR}-01-01,to_publication_date:{YEAR}-12-31,type:article, per_page=200, and page=1,...,5, with a maximum of 1000 records per year (no sort parameter was used). OpenAlex sorted the results by relevance, which is a combination of text similarity and citation count. The corpus is a bounded, non-probability temporal sample rather than a census.

Titles and abstracts were normalized, unusable text was removed, and OpenAlex identifiers were used for deduplication, leaving 6,426 records.

Titles and abstracts of articles were concatenated and screened for AI relevance with a well-documented rule set that included terms related to AI, machine learning, neural models, transformers and attention, language models and NLP, computer vision, reinforcement learning, recommender systems, knowledge graphs, and common supervised-learning terms. This kept 5,445 records for the text-quality audit and semantic pipeline. A rule-based audit identified 400 records that lacked text, were very short, had bibliographic or front-matter-like text, or had missing text. All available evidence from ACL Anthology, OpenAlex, Crossref and publisher was collected to fix or determine the flagged text. There were 5,429 records after excluding 16 records that did not contain any analytical text. Year counts were 77, 472, 725, 785, 749, 851, 872, and 898 for 2018–2025. After the retrieval and screening, 77 records were eligible for the 2018 cohort. All seven records marked in the last text quality audit were retained, four of which had recovered abstracts. The 77 papers from 2018 were used only as prior memory, and 2019 was treated as a descriptive boundary year. The initial comparison starts in 2020 when the prior archive of observed papers was 549. The final 5,429-record corpus was used for a text audit and all outputs reported in the manuscript were regenerated.

3.2. Semantic representations

The final 5,429-record corpus was encoded with sentence-transformers/all-MiniLM-L6-v2 (384D), allenai/scibert_scivocab_uncased (768D, mean pooling), and allenai/specter (768D, classification-token pooling). The maximum sequence lengths were 256, 256, and 512 tokens, respectively, and longer inputs were cut by the corresponding tokenizer. For SciBERT and SPECTER, tokenization used padding=True and truncation=True. The observed truncation rates were 50.76% for SciBERT and 2.82% for SPECTER. This SciBERT truncation rate is significant, and SciBERT is not used as a basis for the central archive-size conclusion, but is used only as a check on the representation-sensitivity. Vectors were L2-normalized. A temporal TF-IDF baseline was used, which did not include any future-document IDF leakage, but only previous documents for the historical comparison in each year. Pooling and truncation are implementation options and will be part of the scope of the audit. Embedding was done with Python 3.12.13, PyTorch 2.11.0+cu128, Transformers 5.13.0, Sentence-Transformers 5.6.0 and a Tesla T4; the seed was recorded as 42.

3.3. Historical semantic distinctiveness and within-year compactness

The historical semantic distinctiveness of the focal paper i (with representation zi and publication year t) is the average of the cosine distances 1 - cos (zi, zj) to its k most similar papers in the historical archive M<t (which only contains papers published before year t). Consider the set of k nearest neighbors and let Pk (i; M<t) be this set. The main specification adopts the following: k=5, exact brute-force cosine search:

$$N_{i(k;M_{<t})} = \frac{1}{k} \sum_{j \in P^{k(i;M<t)}} \left[1 - cos(z_i, z_j)\right] \quad (1)$$

The higher the score, the greater the mean distance to the k nearest earlier papers under the specified representation and archive. It is not a direct measure of creativity, quality, impact and epistemic contribution. The local compactness was computed separately from the k nearest same-year papers, excluding the focal paper:

$$C_{i(k)} = \frac{1}{k} \sum_{j \in S^{k(i)}} cos(z_i, z_j) \quad (2)$$

The two quantities are used to answer different questions. Distinctiveness is retrospective and archive-temporal, compactness is contemporaneous and only refers to the local semantic density. The neighborhood sensitivity check is not an independent confirmatory analysis, but rather is used as k=10 in annual summaries. Proposition 1 (Monotonicity of the archive size at the pointwise level). Let M_a ⊆ M_b, |M_a| ≥ k and let P_k(i; M) be the k closest neighbors of a fixed representation z_i in memory M.

$$N_{i(k;M)} = \frac{1}{k} \sum_{j \in P^{k(i;M)}} d(z_i, z_j) \quad (3)$$

where

$$d(z_i, z_j) = 1 - \cos(z_i, z_j) \quad (4)$$

then

$$N_{i(k;M_b)} \leq N_{i(k;M_a)} \quad (5)$$

If a candidate set is nested, then enlarging the set cannot increase the ordered k-nearest distances and so cannot increase their mean. The order-statistics property is an elementary property used in the longitudinal archive audit and presented in Proposition 1, not as a new general theorem.

This effect is also present for a fixed focal paper, even in the presence of a stationary document-generating process. Annual means are not necessarily monotone as they are comparing different focal cohorts. The memory-regime comparison is an empirical sensitivity analysis of practical dimensions and is not a proof nor a causal estimate of a field trend.

3.4. Memory-regime controls

There are three memory regimes that probe this design dependence. All previous papers recorded are stored in the cumulative memory. Rolling memory contains papers from the last three publication years, and capacity may still change. The numbers of prior papers in cumulative/rolling-three-year memory were 77/77 (2019), 549/549 (2020), 1,274/1,274 (2021), 2,059/1,982 (2022), 2,808/2,259 (2023), 3,659/2,385 (2024), and 4,531/2,472 (2025). The archives were the same up to 2021, and the archives differed by 77 papers in 2022, with larger differences thereafter, so rolling memory is a sensitivity comparison, but not an independent control. The main capacity was 549, which was the same as the full previous archive available in 2020. Each year, 250 samples were taken without replacement for each capacity level m (default_rng(240613 + 10000×year + m)) and the reported annual value is the mean over draws. To illustrate the capacity effect shape, additional fixed capacities of 100, 250, 750 and 1000 were tested. This control only prevents the cardinality of candidate-set papers from changing, the age and topical composition of eligible papers still change. The fixed-capacity curve is thus a sensitivity diagnostic and not a population-calibrated estimate.

3.5. Robustness and sensitivity analyses

The main evidence is descriptive (2020-2025 annual means, endpoint changes, and slopes). Six annual observations, in a temporal sequence, have overlapping memories and do not merit confirmatory independent-error inferences. For this reason, p values are not included in the main interpretation. The robustness is assessed based on direction and magnitude across representations, k=5 vs k=10, pre/post-repair results, and memory regimes. There are no claims of multiple testing.

Uncertainty in cumulative and rolling MiniLM annual means and endpoint changes was quantified by 10,000 nonparametric bootstrap resamples within each year (seed=240613) to give descriptive 95% intervals conditional on the observed relevance-ranked sample.

3.6. Topic modeling and drift

MiniLM vectors were used as an exploratory module for BERTopic. For Uniform Manifold Approximation and Projection (UMAP) [17], the vectors were reduced to 5 dimensions (15 neighbors, min_dist=0, cosine geometry, random_state=42) and for Hierarchical Density-Based Spatial Clustering of Applications with Noise (HDBSCAN) [18], the min_cluster_size was set to 35, the min_samples to 10, the Euclidean distance in the reduced space, and prediction data were enabled. The topic labels were constructed using the class-based TF-IDF using the English stop words, 1–2-grams, and min_df=5. Adjacent-year drift was cosine distance between topic centroids, and a prevalence-weighted mean was a summary of sample-level drift. These topic outputs are an extension of the audit protocol only, and are not used to validate the central archive-size result, but are used to describe one fitted partition.

3.7. Retrieval-based provenance layer

Each year's MiniLM, SciBERT, and SPECTER distinctiveness scores were transformed to percentile ranks within each year. The minimum percentile and the mean percentile were calculated for each paper, and the

top three papers for each year (2019–2025) were chosen based on the minimum percentile, followed by the mean percentile, both in descending order. Under three different search methods (MiniLM, SciBERT, and SPECTER), the three closest previous papers were retrieved for each of the 21 papers selected. The results are traceable for these cases due to similarity scores, gaps in years, and overlap between neighbor sets. They don't establish causation, provide expert validation, or demonstrate agreement between encoders in the broader population.

3.8. Validity safeguards and scope conditions

The corpus repairs were done before the metric reconstruction and immutable identifiers were checked after the merges. Four constructs are maintained: distance from previous texts, compactness same year, topic prevalence, centroid movement. No semantic ground truth or construct validity claim is made. Agreement is a sign of robustness to a tested choice; disagreement is kept as representation sensitivity. The external validity is limited by the relevance-ranked sample, the left boundary of the year 2018, the text of title+abstract, the English-oriented preprocessing, the document-type indexing, and the single BERTopic configuration.

3.9. Scale interpretation and claim discipline

The raw distances are not comparable between encoders due to the different concentration and anisotropy in the vector space. The within-model temporal change, indexed trajectories, rank patterns, and retrieval overlap are thus used in cross-model comparison, rather than equality of raw scores. Distinctiveness is the operational quantity defined above under a declared memory, and not creativity or scientific value.

4. Results

4.1. Cross-model historical distinctiveness and compactness

Each representation is indexed to the mean of 2020 in Figure 2 to allow for comparison of direction, but not scale. The indexed numbers for MiniLM, SciBERT, SPECTER and temporal TF-IDF baseline were 87.15, 91.89, 77.91 and 97.73, respectively, by 2025. Raw endpoints, endpoint changes and descriptive slopes are reported in Table 3.

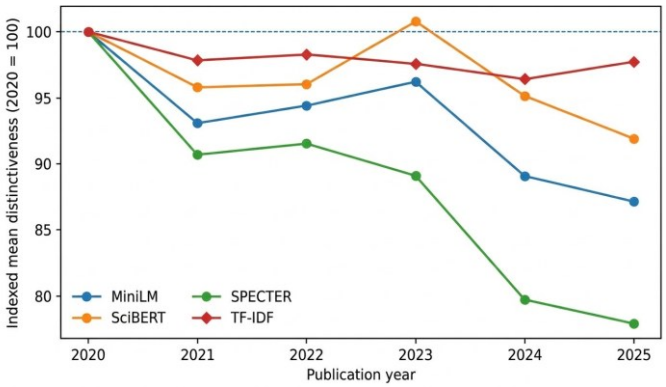


Figure 2. Indexed annual mean historical semantic distinctiveness across MiniLM, SciBERT, SPECTER, and TF-IDF, with 2020 set to 100 within each representation.

The MiniLM, SPECTER, and baseline (temporal TF-IDF) move downward over the main period, whereas SciBERT has a smaller and less monotonous decrease.

Table 3. Cross-representation historical semantic-distinctiveness summaries, 2020–2025.

Model	2020	2025	Change	Slope/year
MiniLM	0.3902	0.3401	−12.85%	−0.00830
SciBERT	0.0654	0.0601	−8.11%	−0.00071
SPECTER	0.1660	0.1293	−22.09%	−0.00692
TF-IDF	0.7729	0.7554	−2.27%	−0.00360

This illustrates that a pattern is representation- and memory-conditioned rather than universal across encoders, and that it does not mean that creativity has diminished in the field.

The distinctiveness of the endpoints of MiniLM, SciBERT, SPECTER and temporal TF-IDF decreased by −12.85%, −8.11%, −22.09% and −2.27%, respectively. The encoders are not compatible with a universal magnitude or trajectory and raw scales are still model dependent. Since 50.76% of the inputs to SciBERT were truncated, the trajectory is considered to be representation-sensitivity evidence rather than central evidence to support the archive-size conclusion.

Within-year compactness produced a different pattern: MiniLM changed from 0.6270 in 2020 to 0.6248 in 2025 (−0.35%), SciBERT from 0.9360 to 0.9380 (+0.21%), and SPECTER from 0.8434 to 0.8541 (+1.26%). These endpoint changes are less distinct and smaller in size than historical distinctiveness. A decrease in distance from previous work can therefore not be taken as an automatic increase in contemporaneous convergence.

4.2. Neighborhood and cleaning robustness

The annual ordering was almost maintained for the entire range of k values available, with k=10 maintaining the k=5 ordering. The annual distinctiveness (k=5, 10) had Pearson r=0.9998 and Spearman ρ=1.000, with a difference of 0.53 percentage points between the endpoint changes of 2020 and 2025. Original and repaired annual scores were highly correlated (r=0.9973), as were the original and repaired paper-level distinctiveness and compactness scores (r=0.9739 and r=0.9729, respectively). These correlations are agreement diagnostics, on related measurements, not independent validation of the construct.

4.3. Historical-memory dependence

The cumulative MiniLM trajectory is compared to the fixed capacity reference memories of various sizes in Figure 3.

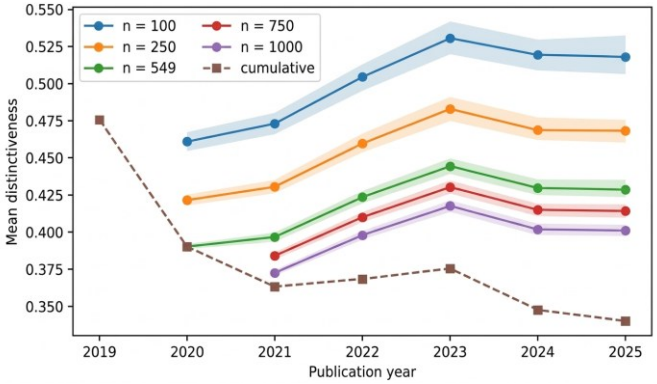


Figure 3. MiniLM historical semantic distinctiveness under cumulative and fixed-capacity prior-memory regimes; 2019 is a boundary point and 2020–2025 is the main period. Fixed-capacity bands show ±1 standard deviation (SD) across 250 draws. For m = 549, the 2020 band has zero width (all 250 draws use the same 549-paper prior archive).

The absolute distance value for the cumulative summary decreases during the time period while the fixed-capacity summary increases after 2020, where the larger capacities have lower absolute distance values.

Results for fixed capacities greater than the available prior archive are only displayed starting from the first year with enough previous papers; the cumulative 2019 point is a boundary-year diagnostic. The empirical audit result is the separation between the curves. Proposition 1 does not define annual cohort means, but only sets the pressure to be non-increasing for a specific document in a nested set of candidates. However, the sign reversal at the endpoint observed in this sample is a practically significant sensitivity diagnostic. It does not measure latent creativity and the fixed-capacity curve still allows for changes in prior-paper age and topic composition.

Table 4 quantifies the contrast. Cumulative and rolling-three-year endpoints changed by −12.85% and −10.59%, whereas the fixed-capacity endpoint at 549 prior papers changed by +9.84%. The bootstrap 95% intervals for the 2020–2025 endpoint change was −14.55% to −11.09% for cumulative memory and −12.30% to −8.82% for rolling memory. These intervals measure the sampling variability of the focal papers in the observed sample and are different from the draw-to-draw variability of fixed capacity memory. The sign reversal is considered to be a sensitivity result. There are no slopes with confirmatory claims attached to them, as there are six dependent annual points.

Table 4. MiniLM distinctiveness summaries under cumulative, rolling-three-year, and fixed-capacity memory, 2020–2025.

Memory regime	2020	2025	Change	Slope/year
Cumulative previous years	0.3902	0.3401	−12.85%	−0.00830
Rolling three years	0.3902	0.3489	−10.59%	−0.00647
Fixed capacity (549; 250 draws)	0.3902	0.4286	+9.84%	+0.00891

In order to enhance computational reproducibility, the 5429 corpus record identifiers were matched to the 5429 rows of stored MiniLM embeddings. The reported cosine

k-nearest-neighbor procedure (k = 5) was applied to the archived embeddings to recalculate annual means and recover the endpoint changes for 2020–2025, yielding −12.85% and −10.59% for cumulative and rolling three-year annual means, respectively. For fixed-capacity memory (549 papers), reaggregation of the 250 scores stored for individual draws in each year yielded the published annual means and the +9.84% endpoint change.

4.4. Topic structure and drift

Of 5,429 papers, 4,158 were assigned to 39 clusters; 23.41% were outliers. The prevalence-weighted and maximum centroid movement for this fitted partition are reported in Figure 4.

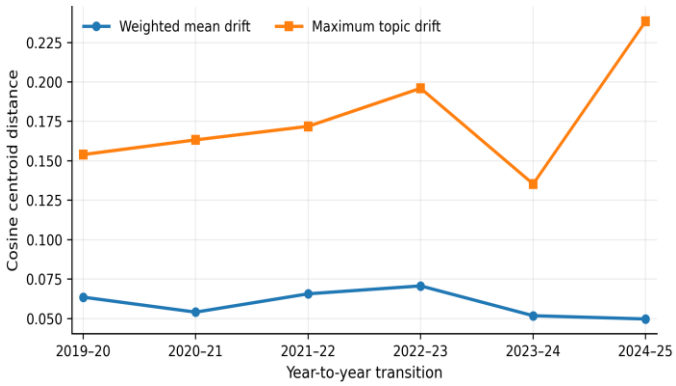


Figure 4. Exploratory topic-centroid drift across adjacent publication years in the fitted BERTopic partition. Weighted drift is prevalence-adjusted; maximum drift identifies the most mobile retained cluster.

The topic outputs are provided as examples of how the audit can be extended, and are not considered to be stable topic discovery.

Within this fitted partition, prevalence-weighted drift peaked at 0.0707 for 2022–2023 and was 0.0498 for 2024–2025; maximum drift reached 0.2384 in 2024–2025. These values are for one sample and one parameterization. They do not indicate a regime shift or stability of the topics in the population.

4.5. Retrieval-based provenance

The provenance layer retrieved 21 papers with high scores and 189 model-specific neighbor links in the model-specific embedding. The mean similarity for MiniLM, SciBERT, and SPECTER was 0.427, 0.885, and 0.766, respectively, to the three closest previous papers, while the mean year gap was 1.75, 1.73, and 1.86 years. Embedding geometries differ and similarity magnitudes are only interpretable within models.

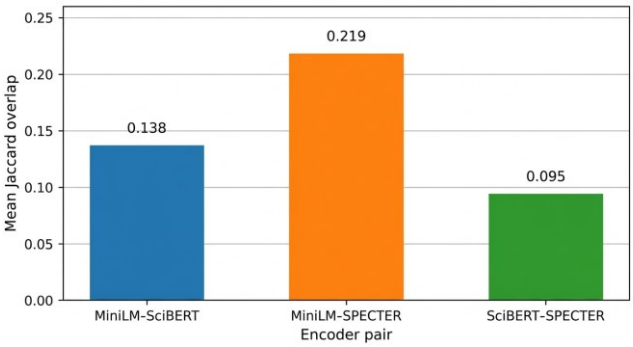


Figure 5. Mean pairwise Jaccard overlap of three-nearest-prior-paper sets for 21 high-score cases (three per year, 2019–2025), comparing MiniLM, SciBERT, and SPECTER.

The mean neighbor overlap for MiniLM-SPECTER was 0.219, followed by MiniLM-SciBERT (0.138), and SciBERT-SPECTER (0.095), as shown in Figure 5. The identity of the neighbor is thus highly variable even for the same set of selected focal papers.

Table 5. Retrieval-provenance summaries for 21 high-score cases (three per year, 2019–2025) under MiniLM, SciBERT, and SPECTER. Mean pairwise Jaccard is the average of the two pairwise overlaps involving each model.

Model	Mean prior similarity	Mean year gap	Mean pairwise Jaccard
MiniLM	0.427	1.75	0.179
SciBERT	0.885	1.73	0.117
SPECTER	0.766	1.86	0.157

Table 5 gives a summary of the provenance behavior for the selected cases. The average pairwise Jaccard overlap for MiniLM, SciBERT, and SPECTER was 0.179, 0.117, and 0.157, respectively, with the mean three-model intersection being 0.333 neighbors out of three.

4.6. Integrated result

There are three evidential layers in the integrated result. First, there are differences in cumulative-memory

trajectories between representations. Secondly, neighborhood and text-repair checks maintain the general ordering. Thirdly, the endpoint sign changes under fixed capacity, which indicates that the construction of the archive is a part of the estimand. Secondary diagnostics (topic drift and retrieval) are not required to set the archive-size result.

5. Discussion

The main result is a validity result for the longitudinal nearest-neighbor systems. The effect of the empirical memory regimes is sufficiently large to change the direction of the observed endpoint. A cumulative-memory line is not a pure time series of field behavior, but is a shared property of the focal documents and the growing comparison archive.

The analysis assumes that the cosine distance within the same representation is a measure of how distant the two texts are in terms of their semantic meaning, and that the comparison archive is part of the estimand and must be controlled or sensitivity-tested.

Cross-representation evidence does not eliminate but reveals model dependence. The MiniLM, SPECTER and temporal TF-IDF scores move downward over 2020–2025, and the score for SciBERT decreases, but not as dramatically or monotonously. The SciBERT result is only used as representation-sensitivity evidence. Its 50.76% truncation rate could affect the embedding geometry and reduce the observed variance, making it unsuitable to anchor the archive-size conclusion.

An AI monitoring system should therefore be able to report on the trajectory of each of the encoders or standardized ranks and should not be able to move the thresholds from one embedding space to another. The proposed audit protocol can be used to provide a report of archive scope, encoder choice, neighborhood size, and retrieval provenance that can be used to support AI-based literature monitoring systems, research dashboards, and semantic retrieval pipelines prior to interpreting temporal semantic-distance trends. The same audit principle can be applied, subject to domain-specific validation, to longitudinal semantic monitoring of biomedical literature, patent and industrial-technology archives, public-policy or regulatory document archives, as the monotonic archive-size mechanism is not AI-specific but rather based on k-nearest-neighbor candidate-set expansion.

The secondary modules are added for inspection but not for the sake of validating the main score. The single BERTopic fit does not have a stability ensemble, and only models' centroid movement at the sample level. Nearest-neighbor provenance reveals which previous documents contribute to individual scores, but low cross-encoder overlap indicates that this evidence is representation-conditional. These modules should be considered as audit diagnostics rather than causal explanations.

This sample is limited to 1,000 relevance-ranked records per year, is not a census, may be affected by year-dependent selection due to the fact that older papers have had more time to receive citations, which contributes to the OpenAlex relevance ranking, lacks full-text and citation context in title/abstract inputs, and includes no records before 2018. The entire previous archive from the 2018 cohort was used for the descriptive 2019 observations. All three MiniLM memory regimes used the same 549-paper reference archive (77 from 2018, 472 from 2019) at the 2020 baseline. The use of 549 historical papers does not keep publication ages or topical composition of papers constant across years as each fixed-capacity draw is drawn from the eligible archive for that year. The 250 draws represent the subsampling variation for each year's eligible archive; variations in archive composition are important when comparing results across years. The contrasting cumulative and fixed-capacity endpoint trends, therefore, suggest that it is not just the size of the archive that matters but also the construction of the archive itself. The topic analysis is based on one parameterization and the main period contains six annual observations. These constraints limit the inferences that can be made about the behavior of measurement in the observed sample.

Future work should involve more representative corpora, evaluate the stability of topics and construct validity, and replicate the audit over databases, disciplines and encoders.

More recent generative AI research includes retrieval-augmented generation (RAG)-enhanced large language model (LLM) retrieval [30] and generative semantic-communication frameworks [31]. The present audit may be expanded to include an optional LLM/RAG interface

in the future, allowing for queries of retrieval evidence without altering the archive construction and representation controls.

6. Conclusion

This study proposed a protocol for auditing the historical distinctiveness of AI research using transformer-based representations. It distinguishes between distance from earlier texts, the local compactness in the same year, the optional topic movement, and the retrieval provenance, which are not collapsed into a single claim of novelty. The mean k-nearest-neighbor distance for a fixed paper and representation cannot increase when a nested candidate archive is enlarged. The empirical sample also indicates opposite trends at the endpoint for cumulative and fixed-capacity memory. The result shows that archive construction changes the measured trajectory, not creativity itself. Without overstating the role, the audit can be extended via exploratory topic drift and retrieval provenance. The topic partition fitted is parameter-conditional and different encoders get different earlier neighbors. These findings support explicit uncertainty and provenance reporting in deployed semantic-monitoring systems.

Archive-based embedding systems should report and evaluate the scope, capacity, type of encoder, pooling, truncation, neighborhood size, corpus selection and local provenance. Temporal semantic-distance claims can only be interpreted in the context of their estimand. Further research is needed on a more representative archive, topic-stability analysis, benchmark validation and replication across databases or fields. The first stage is targeted within 6–12 months and the second stage of validation and replication is targeted within 12–18 months from project start.

The primary users are researchers and research-evaluation teams, developers of AI-based literature monitoring and semantic-retrieval systems, as well as research-management or policy users. They can use the audit to evaluate how sensitive observed semantic-distance patterns are to the design of the archives, and to better report on the choices of representation, memory, and provenance.

Competing Interest Statement

The authors declare no known competing financial interests or personal relationships that could have influenced the work reported in this paper.

Data Availability Statement

The bibliographic records used in this study were obtained from OpenAlex. The processed analytical data set, OpenAlex record identifiers and analysis code are available from the corresponding author upon request. The data retrieval and preprocessing procedures are described in Section 3.1.

References

[1] C. Lu et al., “Towards end-to-end automation of AI research,” *Nature*, vol. 651, no. 8107, pp. 914–919, Mar. 2026, doi: 10.1038/s41586-026-10265-5.

[2] Q. Hao, F. Xu, Y. Li, and J. Evans, “Artificial intelligence tools expand scientists’ impact but contract science’s focus,” *Nature*, vol. 649, no. 8099, pp. 1237–1243, Jan. 2026, doi: 10.1038/s41586-025-09922-y.

[3] M. Park, E. Leahey, and R. J. Funk, “Papers and patents are becoming less disruptive over time,” *Nature*, vol. 613, no. 7942, pp. 138–144, Jan. 2023, doi: 10.1038/s41586-022-05543-x.

[4] G. A. Toto, L. Grilli, L. Traetta, R. Villani, A. Petito, and G. Serviddio, “Convergence of disciplines: A systematic review of multidisciplinary development approaches in artificial intelligence,” *Front. Digit. Health*, vol. 7, Art. no. 1400338, Aug. 2025, doi: 10.3389/fdgth.2025.1400338.

[5] D. Abbonato, S. Bianchini, F. Gargiulo, and T. Venturini, “Interdisciplinary research in artificial intelligence: Lessons from COVID-19,” *Quant. Sci. Stud.*, vol. 5, no. 4, pp. 922–935, Oct. 2024, doi: 10.1162/qss_a_00329.

[6] D. Kobak, R. González-Márquez, E.-Á. Horvát, and J. Lause, “Delving into LLM-assisted writing in biomedical publications through excess vocabulary,” *Sci. Adv.*, vol. 11, no. 27, Art. no. eadt3813, Jul. 2025, doi: 10.1126/sciadv.adt3813.

[7] S. Bianchini, M. Müller, and P. Pelletier, “Drivers and barriers of AI adoption and use in scientific research,” *Technol. Forecast. Soc. Change*, vol. 220, Art. no. 124303, Nov. 2025, doi: 10.1016/j.techfore.2025.124303.

[8] N. Suzen, A. N. Gorban, J. Levesley, and E. M. Mirkes, “Predicting the impact of scientific articles through semantic analysis of abstracts,” *Scientometrics*, May 2026, doi: 10.1007/s11192-026-05647-5.

[9] N. Reimers and I. Gurevych, “Sentence-BERT: Sentence embeddings using Siamese BERT-networks,” in *Proc.*

2019 Conf. Empirical Methods in Natural Language Processing and 9th Int. Joint Conf. Natural Language Processing (EMNLP-IJCNLP), 2019, pp. 3982–3992, doi: 10.18653/v1/D19-1410.

[10] W. Wang, F. Wei, L. Dong, H. Bao, N. Yang, and M. Zhou, "MiniLM: Deep self-attention distillation for task-agnostic compression of pre-trained transformers," in *Advances in Neural Information Processing Systems*, vol. 33, 2020, pp. 5776–5788.

[11] I. Beltagy, K. Lo, and A. Cohan, "SciBERT: A pretrained language model for scientific text," in *Proc. 2019 Conf. Empirical Methods in Natural Language Processing and 9th Int. Joint Conf. Natural Language Processing (EMNLP-IJCNLP)*, 2019, pp. 3615–3620, doi: 10.18653/v1/D19-1371.

[12] A. Cohan, S. Feldman, I. Beltagy, D. Downey, and D. S. Weld, "SPECTER: Document-level representation learning using citation-informed transformers," in *Proc. 58th Annu. Meeting Assoc. Comput. Linguistics (ACL)*, 2020, pp. 2270–2282, doi: 10.18653/v1/2020.acl-main.207.

[13] E. Cunningham, B. Smyth, and D. Greene, "Facilitating interdisciplinary knowledge transfer with research paper recommender systems," *Quant. Sci. Stud.*, vol. 6, pp. 854–875, Sep. 2025, doi: 10.1162/qss.a.9.

[14] M. Shahinmoghadam and A. Motamedi, "Benchmarking pre-trained text embedding models in aligning built asset information," *Sci. Rep.*, vol. 15, no. 1, Art. no. 23866, Jul. 2025, doi: 10.1038/s41598-025-09052-5.

[15] Y. Zhao and C. Zhang, "A review on the novelty measurements of academic papers," *Scientometrics*, vol. 130, no. 2, pp. 727–753, Feb. 2025, doi: 10.1007/s11192-025-05234-0.

[16] M. Grootendorst, "BERTopic: Neural topic modeling with a class-based TF-IDF procedure," arXiv:2203.05794, 2022, doi: 10.48550/arXiv.2203.05794.

[17] L. McInnes, J. Healy, N. Saul, and L. Großberger, "UMAP: Uniform manifold approximation and projection," *J. Open Source Softw.*, vol. 3, no. 29, Art. no. 861, Sep. 2018, doi: 10.21105/joss.00861.

[18] R. J. G. B. Campello, D. Moulavi, and J. Sander, "Density-based clustering based on hierarchical density estimates," in *Advances in Knowledge Discovery and Data Mining*, ser. Lect. Notes Comput. Sci., vol. 7819, 2013, pp. 160–172, doi: 10.1007/978-3-642-37456-2_14.

[19] P. Baaden, M. Rennings, M. John, and S. Bröring, "On the emergence of interdisciplinary scientific fields: (How) does it relate to science convergence?," *Res. Policy*, vol. 53, no. 6, Art. no. 105026, Jul. 2024, doi: 10.1016/j.respol.2024.105026.

[20] K. Kim, D. F. Kogler, and S. Maliphol, "Identifying interdisciplinary emergence in the science of science: Combination of network analysis and BERTopic," *Humanit. Soc. Sci. Commun.*, vol. 11, Art. no. 603, May 2024, doi: 10.1057/s41599-024-03044-y.

[21] K. Beyer, J. Goldstein, R. Ramakrishnan, and U. Shaft, "When is 'nearest neighbor' meaningful?," in *Database Theory — ICDT '99*, ser. Lect. Notes Comput. Sci., vol. 1540, 1999, pp. 217–235, doi: 10.1007/3-540-49257-7_15.

[22] M. Radovanović, A. Nanopoulos, and M. Ivanović, "Hubs in space: Popular nearest neighbors in high-dimensional data," *J. Mach. Learn. Res.*, vol. 11, pp. 2487–2531, 2010.

[23] A. Kiourtis, S. Nifakos, A. Mavrogiorgou, and D. Kyriazis, "Aggregating the syntactic and semantic similarity of healthcare data towards their transformation to HL7 FHIR through ontology matching," *Int. J. Med. Inform.*, vol. 132, Art. no. 104002, Dec. 2019, doi: 10.1016/j.ijmedinf.2019.104002.

[24] G. Manias, A. Mavrogiorgou, A. Kiourtis, and D. Kyriazis, "SemAI: A novel approach for achieving enhanced semantic interoperability in public policies," in *Artificial Intelligence Applications and Innovations*, I. Maglogiannis, J. Macintyre, and L. Iliadis, Eds., IFIP Adv. Inf. Commun. Technol., vol. 627. Cham, Switzerland: Springer, 2021, pp. 687–699, doi: 10.1007/978-3-030-79150-6_54.

[25] J. Priem, H. Piwowar, and R. Orr, "OpenAlex: A fully-open index of scholarly works, authors, venues, institutions, and concepts," arXiv:2205.01833, 2022, doi: 10.48550/arXiv.2205.01833.

[26] J. H. Culbert et al., "Reference coverage analysis of OpenAlex compared to Web of Science and Scopus," *Scientometrics*, vol. 130, no. 4, pp. 2475–2492, Apr. 2025, doi: 10.1007/s11192-025-05293-3.

[27] M. Thelwall and X. Jiang, "Is OpenAlex suitable for research quality evaluation and which citation indicator is best?," *J. Assoc. Inf. Sci. Technol.*, vol. 76, no. 12, pp. 1660–1681, Dec. 2025, doi: 10.1002/asi.70020.

[28] K. Mavrogiorgos, A. Mavrogiorgou, A. Kiourtis, N. Zafeiropoulos, S. Kleftakis, and D. Kyriazis, "Automated rule-based data cleaning using NLP," in *2022 32nd Conference of Open Innovations Association (FRUCT)*, 2022, pp. 162–168, doi: 10.23919/FRUCT56874.2022.9953810.

[29] K. Voulgaris et al., "Data processing tools for graph data modelling big data analytics," in *2022 13th International Congress on Advanced Applied Informatics Winter (IIAI-AAI-Winter)*, 2022, pp. 208–212, doi: 10.1109/IIAI-AAI-Winter58034.2022.00048.

[30] D. Tomkou et al., "Bridging industrial expertise and XR with LLM-powered conversational agents," in *2025 21st International Conference on Distributed Computing in Smart Systems and the Internet of Things (DCOSS-IoT)*, 2025, pp. 1050–1056, doi: 10.1109/DCOSS-IoT65416.2025.00158.

[31] C. Liang and D. Li, "Generative AI-enabled semantic communication: State-of-the-art, applications, and the way ahead," *IEEE Commun. Surveys Tuts.*, vol. 28, pp. 3976–4015, Jan. 2026, doi: 10.1109/COMST.2025.3649707.