

Article

Low-Resource Topic Modeling for Domain Experts: A Pilot Study in CO₂ Mineralization Literature

Giuseppe Sabatino ^{1,2,*} , Giuseppe De Rosa ¹, Alessandro Tripodo ²  and Marcella Di Bella ¹

¹ Istituto Nazionale di Oceanografia e di Geofisica Sperimentale–OGS, Borgo Grotta Gigante 42/C, 34010 Trieste, Italy; gderosa@ogs.it (G.D.R.); mdibella@ogs.it (M.D.B.)

² Dipartimento di Scienze Matematiche e Informatiche, Scienze Fisiche e Scienze della Terra–MIFT, Università degli Studi di Messina, Viale Ferdinando Stagno d’Alcontres, 31, 98166 Messina, Italy; atripodo@unime.it

* Correspondence: gsabatino@ogs.it

Abstract

Bibliometric topic-modeling is essential for synthesizing scientific literature, yet its adoption by domain-expert teams is hindered by technical barriers. This pilot study benchmarks four topic modeling paradigms on 286 CO₂ mineralization papers (2013–2025), with all methods implemented end-to-end on consumer hardware by researchers without prior natural language processing (NLP) training. Its principal methodological contribution is a four-level calibration protocol for zero-shot large language models (LLMs) used as innovation detectors. Naive zero-shot prompting flagged 71.0% of the corpus as innovative, over-attributing novelty by a factor of 22.7 relative to a calibrated statistical ensemble. A qualitative-strict prompt reduced this to 26.6%, while a structurally different unified prompt yielded 28.0%. Together, these results indicate a floor near 28% that prompt engineering alone does not overcome. A self-reflection pass, in which the model re-reads and revises its own output, instead reduced the flag rate to 3.5%, close to the ensemble’s 3.13%. The papers retained by self-reflection form a strict subset of those retained by the unified prompt, suggesting that the model applies a consistent discriminative criterion rather than simply adopting a stricter prompting frame. On this basis, we propose a transferable four-layer triangulation protocol combining statistical ensembles with complementary LLM regimes. The classical benchmark shows that no single pipeline dominates and we derive selection guidance accordingly.

Keywords: topic modeling; bibliometric analysis; CO₂ mineralization; large language models (LLMs); zero-shot innovation detection; low-resource computing; SciBERT; research synthesis



Academic Editor: Haiyan Wang

Received: 2 July 2026

Revised: 10 August 2026

Accepted: 28 September 2026

Published: 16 October 2026

Copyright: © 2026 by the authors.

Licensee MDPI, Basel, Switzerland.

This article is an open access article distributed under the terms and conditions of the [Creative Commons Attribution \(CC BY\)](https://creativecommons.org/licenses/by/4.0/) license.

1. Introduction

Bibliometric topic-modeling is a cornerstone of modern research synthesis, enabling systematic mapping, gap identification, and strategic research planning [1,2]. However, its adoption by domain-expert teams is often hindered by an “accessibility gap”: sophisticated probabilistic and neural pipelines typically demand advanced computational linguistics expertise and dedicated infrastructure beyond the reach of specialized research groups. While recent advances such as pre-trained scientific language models [3,4] open-source natural language processing (NLP) toolkits [5] and artificial intelligence (AI)-assisted coding offer a path to democratizing these methods, systematic evidence on their end-to-end

implementability under realistic institutional constraints remains scarce. The methodological landscape is currently divided into four paradigms: probabilistic latent Dirichlet allocation (LDA) [6], neural variational models such as ProLDA [7], transformer-based embedding frameworks such as BERTopic [8,9] and neural variational correlated topic models (NVCTM). Comparative benchmarks across these paradigms under uniform constraints remain rare. Recent domain applications continue to appear: Cao et al. [10], for instance, applied BERTopic with an LDA baseline to 22,623 construction accident reports, at a corpus scale two orders of magnitude larger than the one considered here. Moreover, the post-2024 proliferation of large language models (LLMs), including LLM-based topic modeling frameworks [11], poses a new challenge: can zero-shot LLMs replace traditional, calibrated statistical pipelines for thematic and innovation detection? While LLMs offer unrivaled accessibility, recent calibration studies [12] highlight systematic biases, notably an over-attribution of novelty in the absence of explicit base-rate priors. Although self-refinement strategies [13] show promise, their efficacy in domain-specific innovation detection has not been systematically validated.

The gap these two studies leave open is not the absence of either approach but the absence of an account of why either alone is insufficient. Calibrated statistical ensembles operate on distributional signals: they detect documents whose position in embedding space is anomalous, and distributional anomaly need not coincide with conceptual novelty, so work that is conceptually novel but distributionally typical is invisible to them. Zero-shot LLMs operate on semantic plausibility and recover precisely that kind of candidate, but in the absence of an external base rate they have no mechanism for controlling false positives and substantially over-attribute novelty. Neither failure is corrected by refining the other method, because the two are failures of different kinds; in our corpus the two families select almost disjoint sets of papers (pick-level Jaccard 0.060 between the unified LLM prompt and the statistical ensemble). This motivates their integration rather than their substitution and it is what the present study sets out to characterize.

This study addresses these gaps through two primary objectives. First, we provide an end-to-end benchmark of four classical topic-modeling pipelines on a 286-paper CO₂ mineralization corpus, implemented entirely on consumer hardware by non-specialists. Second, we present a controlled ablation of LLM calibration mechanisms, ranging from naive zero-shot to multi-pass self-reflection, and compare their performance against calibrated statistical ensembles. We hypothesize that these two paradigms are not merely substitutive but complementary, and that a calibrated triangulation workflow can provide more robust insights than any single model alone. Using CO₂ mineralization research as a proof-of-concept (a field with a mature, technically specialized vocabulary), we demonstrate that domain experts (mineralogists, geochemists, volcanologists and geophysicists) can perform rigorous bibliometric synthesis without prior NLP training.

This study contributes:

- an end-to-end implementation framework for domain-expert teams working under institutional constraints;
- a four-level calibration ablation of LLM innovation detection;
- a quantitative comparison of neural and statistical ensembles;
- a methodology for augmented topic interpretation using LLM-generated labels.

2. Materials and Methods

2.1. Study Design and Rationale

This methodological pilot study prioritizes implementation feasibility and resource accessibility over corpus comprehensiveness to evaluate the practical challenges and performance characteristics of four advanced topic-modeling pipelines when implemented

by domain experts (mineralogists, geochemists, volcanologists and geophysicists) on consumer-grade hardware under typical institutional constraints. The CO₂ mineralization research domain serves as a proof-of-concept case study, chosen because it is the field in which our team can validate results directly rather than because a corpus of this size is sufficient in any established sense; the constraints this places on the benchmark are set out in Section 2.11. The study design incorporated six independent runs per model and a performance assessment across multiple quantitative and qualitative metrics. We do not treat six runs as guaranteeing robustness a priori; the dispersion observed across seeds is reported in Section 2.3 and Table 1.

Table 1. Comparative performance of the four topic-modeling pipelines under resource constraints. Values are means $\pm$ SD over six independent runs (seeds 42, 123, 456, 789, 1024, 2048); CV = coefficient of variation. Metric definitions are given in Supplementary Table S1. Coherence (Cv) for BERTopic + SciBERT uses c-TF-IDF top-15 terms; Technical Term Saliency is a computed vocabulary statistic, not a human assessment. Wall-clock is the mean model fitting time per run on a Mac Mini M2 (16 GB RAM); total elapsed time is reported in Section 4.3. Perplexity is omitted because it is not comparable across architectures (Section 2.8; Supplementary Tables S2 and S3). SciBERT-NVCTM denotes the deposited MK6-K20 configuration.

Metric	MalletLDA (K = 18)	SD	CV	ProdLDA (K = 22)	SD	CV	BERTopic + SciBERT (K = 22)	SD	CV	SciBERT- NVCTM (K = 20)	SD	CV
Coherence (C _v)	0.542	0.009	0.017	0.408	0.018	0.044	0.381	0.004	0.010	0.414	0.013	0.031
Semantic Coherence	0.081	0.009	0.115	0.327	0.009	0.028	0.612	0.005	0.008	0.633	0.009	0.014
Domain Specificity	0.207	0.007	0.035	0.395	0.008	0.020	0.621	0.005	0.008	0.645	0.004	0.006
Technical Term Saliency	0.686	0.021	0.031	0.617	0.030	0.049	0.804	0.028	0.035	0.795	0.002	0.003
Field Coverage	0.011	0.001	0.044	0.062	0.001	0.019	0.052	0.001	0.023	0.042	0.001	0.027
Topic Diversity	0.560	0.019	0.033	0.943	0.006	0.006	0.734	0.017	0.023	0.497	0.015	0.030
Wall-clock (s/run)	54	—	—	48	—	—	81	—	—	443	—	—

2.2. Corpus Construction and Infrastructure

The corpus comprises 286 articles (2013–2025) identified via Google Scholar and accessed through institutional subscriptions at the University of Messina. We deliberately maintained this realistic, non-ideal sampling, which includes a significant temporal bias (69.9% of papers from 2020 to 2025), to test the robustness of the methodology. All processing occurred on a Mac Mini M2 (16 GB RAM). Metadata extraction was performed via GROBID [5,14], with pipeline implementation facilitated by AI-assisted coding, simulating what we designate throughout as a “low-resource” research environment, by which we mean three specific constraints: model fitting confined to a single consumer machine with a 16 GB memory ceiling and no high-performance cluster or GPU server, corpus access limited to institutional subscriptions rather than comprehensive database licenses and implementation by domain experts without computational linguistics training (Section 2.11).

2.3. Topic Modeling and Interpretive Transparency

We implemented four paradigms: (i) MalletLDA (probabilistic), (ii) ProdLDA (neural variational), (iii) BERTopic + SciBERT (transformer-based), and (iv) SciBERT-NVCTM (correlated variational). These four are not a survey of available tools but a cross of the two axes along which topic modeling has developed. The first is document representation: bag-of-words counts (MalletLDA, ProdLDA) versus contextual transformer embeddings (BERTopic + SciBERT, SciBERT-NVCTM). The second is the inference objective: collapsed

Gibbs sampling (MalletLDA), amortized variational inference with a product-of-experts decoder (ProdLDA), non-generative density-based clustering with class-based TF-IDF (BERTopic + SciBERT) and variational inference with an explicit topic correlation structure (SciBERT-NVCTM). This cross is consequential rather than taxonomic, as the benchmark shows: the representation axis orders the models monotonically on embedding-based semantic coherence (0.081, 0.327, 0.612, 0.633; all pairwise differences significant) while reversing that ranking on co-occurrence-based C_v , where the bag-of-words MalletLDA leads at 0.542; the inference objective governs topic diversity (0.943 for ProdLDA versus 0.497 for the correlation-aware SciBERT-NVCTM); and fitting cost spans an order of magnitude on identical hardware, from 48 to 443 s per run (Tables 1 and 2). A fourth selection criterion follows from the premise of this study: each pipeline had to have a maintained open-source implementation that a non-specialist can run to completion on a 16 GB consumer machine, which excludes approaches requiring cluster-scale training. Each model was fitted six times under different random seeds. We do not claim that six runs guarantee stability a priori; rather, the dispersion observed across seeds is small enough for the reported means to be treated as stable, with the coefficient of variation below 0.05 for 23 of the 24 model-metric combinations in Table 1 and a single exception at 0.115 (semantic coherence for MalletLDA, whose absolute value of 0.081 inflates the relative dispersion). To balance raw output with human readability, we augmented the top-15 terms for each topic with descriptive labels generated by Claude Haiku 4.5 [15], under a strict constraint to use only domain-relevant terminology. Crucially, the raw top-N terms remain the canonical representation; LLM-labels are explicitly declared as interpretive aids subject to the same calibration caveats documented in our ablation study. This dual-representation approach ensures methodological transparency, allowing researchers to override LLM interpretations by inspecting the underlying data. The number of topics per model was not selected through an automated coherence sweep but set pragmatically in relation to the corpus size (286 papers), targeting a granularity (≈ 18 –22 topics) that domain experts could review exhaustively; for BERTopic + SciBERT, the topic count emerged from its density-based clustering rather than being fixed a priori. We treat the absence of formal K-optimization as a deliberate pilot study simplification (see Section 2.11). One departure from this pragmatic stance must be noted: for BERTopic + SciBERT, the UMAP and HDBSCAN hyperparameters were selected by Bayesian optimization against a composite coherence-and-diversity objective computed on the study corpus, whereas MalletLDA, ProdLDA and SciBERT-NVCTM were run with paper-faithful or default settings. Because coherence and topic diversity are themselves among the metrics reported in Table 1 and tested in Table 2, this gives BERTopic + SciBERT an optimistic bias in those quantities that the other three pipelines do not share; the optimization code is included in the deposited archive.

Table 2. One-way ANOVA and Bonferroni-corrected pairwise comparisons across evaluation metrics, computed on six independent runs per model using the SciBERT-NVCTM ($K = 20$) configuration consistent with Table 1. For Coherence (C_v), the BERTopic + SciBERT estimate is computed on the c-TF-IDF top-15 terms of each topic (four-group ANOVA). Perplexity is omitted because per-model values are not comparable across architectures (Supplementary Tables S2 and S3). The results are interpreted in Section 3.1; bootstrap 95% confidence intervals and Cohen’s d effect sizes are reported in Supplementary Table S2.

Metric	ANOVA and Bonferroni-Corrected Pairwise Comparisons ($p < 0.05$)
Coherence (C_v)	$F = 202.5$, $p < 0.001$. MalletLDA > BERTopic + SciBERT ($p < 0.001$); MalletLDA > ProdLDA ($p < 0.001$); MalletLDA > SciBERT-NVCTM ($p < 0.001$); BERTopic + SciBERT < SciBERT-NVCTM ($p = 0.008$); BERTopic + SciBERT $\approx$ ProdLDA (n.s.); ProdLDA $\approx$ SciBERT-NVCTM (n.s.)

Table 2. Cont.

Metric	ANOVA and Bonferroni-Corrected Pairwise Comparisons ($p < 0.05$)
Semantic Coherence	$F = 6048.6, p < 0.001$. MalletLDA < ProdLDA < BERTopic + SciBERT < SciBERT-NVCTM, all pairwise differences significant (BERTopic + SciBERT < SciBERT-NVCTM at $p = 0.006$; all others $p < 0.001$)
Domain Specificity	$F = 6250.7, p < 0.001$. MalletLDA < ProdLDA < BERTopic + SciBERT < SciBERT-NVCTM, all pairwise differences significant ($p < 0.001$)
Technical Term Salience	$F = 89.8, p < 0.001$. ProdLDA < MalletLDA ($p = 0.009$); MalletLDA < BERTopic + SciBERT and <SciBERT-NVCTM ($p < 0.001$); ProdLDA < BERTopic + SciBERT and < SciBERT-NVCTM ($p < 0.001$); BERTopic + SciBERT $\approx$ SciBERT-NVCTM (n.s.)
Field Coverage	$F = 2661.5, p < 0.001$. ProdLDA > BERTopic + SciBERT > SciBERT-NVCTM > MalletLDA, all pairwise differences significant ($p < 0.001$)
Topic Diversity	$F = 1073.1, p < 0.001$. ProdLDA > BERTopic + SciBERT > MalletLDA > SciBERT-NVCTM, all pairwise differences significant ($p < 0.001$)
Interpretability	No significant differences

2.4. Semantic Network Analysis

To capture latent relationships beyond individual topic groupings, we mapped the corpus into a semantic network. Using a skip-gram Word2Vec model trained on the corpus, we derived term embeddings to construct a conceptual graph (18,560 nodes representing unique terms; 35,565 edges representing co-occurrence or semantic similarity). This graph was partitioned into 82 thematic communities using the Louvain modularity-optimization algorithm (python-louvain, weighted edges, default resolution) to identify cohesive research clusters.

We prioritized weighted betweenness centrality [16,17] (computed with NetworkX) to isolate “bridge concepts”, terms that act as critical mediators between disparate scientific domains. Our analysis indicates that these bridge terms, operationally defined as the top 5% of vocabulary terms ranked by betweenness, mediate 43% of cross-community connections. This concentration of connectivity highlights key interdisciplinary nodes, providing a systematic, data-driven method for domain experts to pinpoint emerging “white spaces” or potential innovation intersections that remain hidden within traditional clustering [18].

2.5. Outlier Analysis and Innovation Discovery

Systematic outlier analysis was implemented using a calibrated ensemble of detectors (Isolation Forest [19], local outlier factor [20], and distance- and semantic diversity-based scores, following outlier ensemble principles [21]) to identify documents with distinctive semantic patterns that diverge from mainstream research trajectories. The four detector scores were min-max normalized and combined into a weighted ensemble score (Isolation Forest 0.3, local outlier factor 0.3, distance 0.2, semantic diversity 0.2) and a document was declared an outlier when that score exceeded both the 95th percentile of its distribution and the mean plus two standard deviations. The second criterion is the binding one (0.7193 versus 0.6798), and it retained 9 of the 288 documents (3.13%; see Section 3) as high-precision outliers.

Statistical validation confirmed that outlier documents exhibited significantly different semantic patterns than mainstream clusters ($p < 0.001$), providing systematic frameworks for identifying emerging research trajectories and cross-domain integration opportunities that traditional topic-modeling approaches often overlook.

2.6. Topic Label Generation and Interpretive Transparency

To improve the readability of result tables and prose without obscuring the underlying model output, we generated descriptive labels for each topic of each model by passing the top-15 raw terms (the canonical topic representation produced by c-TF-IDF for BERTopic + SciBERT, by the topic-word distribution for ProLDA and SciBERT-NVCTM, and by frequency aggregation across temporal slices for MalletLDA) through a constrained Claude Haiku 4.5 prompt. The prompt instructed the model to produce a 4–8-word noun phrase label using only the supplied terms or close synonyms, with an explicit coherence flag for noisy aggregations. The total cost for relabeling all 82 topics across the four models was approximately USD 0.11, and runtime was about ~3 min on consumer infrastructure. The choice of this specific model follows from the premise of the study rather than from a model selection experiment and we state the reasoning explicitly. First, the cost and latency envelope: Claude Haiku 4.5 is the lowest tier of a current frontier family, and the entire LLM component of this work (the 82 topic labels together with the four-level ablation of Section 2.7) stays within a few US dollars and about two hours of wall-clock time, which is the order of magnitude a resource-constrained team can absorb as routine expenditure. Second, and more consequential methodologically, using the smallest available model makes the central negative result conservative: the over-attribution of novelty by uncalibrated zero-shot prompting, and the prompt-level floor documented in Section 3.10, are properties one would expect a larger model to attenuate rather than remove, so demonstrating them on the cheapest tier bounds the problem from the accessible end instead of flattering it. Third, the claim under test is not a ranking of language models but whether zero-shot prompting can substitute for a calibrated statistical ensemble; the ablation is therefore designed to vary the calibration mechanism while holding the model fixed. Critically, the raw top-N terms remain the canonical topic representation throughout the paper: the result tables report both the raw terms (column ‘raw_terms_top15’) and the LLM-generated labels (column ‘llm_label’), allowing the reader to override our interpretation by examining the raw signal directly. We treat the LLM-generated labels as interpretive aids subject to the same calibration caveats discussed in our LLM ablation (see Sections 4 and 4.5), rather than as authoritative renaming of the topics. This dual-representation approach extends the methodological transparency principle to topic labeling itself, consistent with the principle of methodological transparency in bibliometric analysis. All Claude Haiku 4.5 calls (topic labeling and the four-level innovation-detection ablation described in Section 2.7) used the model snapshot claude-haiku-4-5-20251001 at the Anthropic API default sampling setting and without a fixed random seed; max_tokens was 1500 for the labeling and for the R1, R2 and R4 passes, and 1000 for the R3 self-reflection pass. Because the API is non-deterministic and no seed is exposed, exact run-to-run reproduction is not guaranteed; the verbatim prompts and the raw model outputs are deposited (see Data Availability Statement) to enable inspection and re-analysis.

Because these labels are used throughout the result tables, their adequacy was assessed independently of the model that produced them. Three domain experts external to the author team judged all 82 labels. They were approached individually by email and sent the rating sheet together with a draft of the manuscript; all participated on a voluntary basis and remain anonymous at their own request. The sheet showed only the top-15 canonical terms of a topic and the label to be judged, without the identity of the model, the declared confidence or the coherence flag, which are the quantities against which the human verdict is subsequently compared, and each label was classified as adequate, partially adequate, or inadequate. The blinding is therefore of the sheet and not of the exercise, since Section 3.3 of the accompanying draft names twelve representative topics with their pipeline and terms, and we cannot establish whether the sheets were completed before or after that draft was

read. The three thematic areas of the corpus (Section 3.3) were supplied once as domain context at the head of the sheet rather than item by item, so that cluster membership could not orient the judgment of an individual label. Ten control items, in which the label shown was drawn from a different topic, were interleaved with the 82 genuine labels; they do not enter the reliability calculation and serve only to verify that raters were discriminating rather than merely assenting. Item order was randomized independently for each rater, although two of the three sheets were returned in the same order, meaning that, for that pair, randomization does not protect against a common order effect.

The criterion set in advance for retaining a sheet was that at least 70% of the control items be rejected as inadequate. In the event, no rater approved a single control item, but one rater assigned the intermediate category to half of them and therefore failed the criterion as literally stated, while meeting the substantive requirement it was written to operationalize, namely, that a rater who approves deliberately mismatched labels is not discriminating. We therefore report the three-rater analysis as primary and the two-rater analysis obtained by applying the threshold literally as a robustness check; the separation of these two readings was decided after the sheets were returned and is stated here as such. Reliability is reported as Fleiss' kappa [22] over three raters and three categories, together with Krippendorff's alpha with ordinal weights [23] and pairwise Cohen's kappa [24] and because the judgments are concentrated in one category, alongside observed agreement and the prevalence-adjusted bias-adjusted kappa [25] (Section 3.10 and Supplementary Table S11).

2.7. LLM Calibration Ablation: Design and Prompt Levels

To address whether large language model approaches can substitute for the calibrated statistical ensembles described above, we performed a controlled four-level ablation of Claude Haiku 4.5 zero-shot topic extraction on the same 286-paper corpus. The LLM ablation operates on the 286-paper text corpus, whereas the statistical outlier ensemble operates on a 288-document set that overlaps with the LLM corpus in 285 documents, with three documents present only in the outlier set and one only in the LLM corpus; the two denominators are reported as such rather than forced to a common value, and the 3.13% ensemble rate corresponds to $9/288 = 3.125\%$. The four runs differed only in the calibration mechanism applied to the innovation flag (`is_outlier_candidate`) decision: (R1) a naïve zero-shot with no calibration anchor; (R2) a qualitative-strict prompt requiring outlier flags to be grounded in one of five predefined novelty dimensions (mechanism, material, cross-domain, magnitude, application context) and to list at least two established alternatives; (R3) a self-reflection multi-pass, in which Haiku critically re-reads its own R1 outputs and either confirms or downgrades each flag against the strict criteria; and (R4) a unified one-shot prompt that fuses the R2 criteria with the ten-year-citation thought experiment used in R3 within a single inference pass. The total cost across the four runs was USD 5.34 and the total runtime was 115 min, on consumer hardware (Mac Mini M2, 16 GB RAM) and without fine-tuning, retrieval augmentation, or specialized infrastructure.

2.8. Performance Evaluation and Statistical Analysis

Model performance was evaluated using multiple metrics including coherence (semantic consistency within topics), domain specificity (field-relevant terminology identification), and technical term salience (the mean vocabulary weight of the high-weight technical terms in each topic's top terms, computed rather than human-assigned). Six independent runs were conducted per model, supporting analysis of variance (ANOVA), Bonferroni-corrected post hoc comparisons and effect-size calculations (Cohen's *d* and eta-squared); because

these runs differ only in the random seed, the interpretive limits of the resulting effect sizes are reported with the results.

Perplexity scores were calculated but given less weight, based on the recent literature that indicated poor correlation with human quality judgments in specialized scientific contexts [26]. Because the four architectures define their held-out likelihoods on different scales, per-model perplexities are not comparable across pipelines and are therefore reported only in the Supplementary Materials (Tables S2 and S3), with an explicit non-comparability caveat.

Result validation relied primarily on domain-expert assessment rather than external gold standards, reflecting realistic validation approaches available to most domain-expert teams. Topic interpretability and relevance were evaluated through iterative review by team members with established expertise in CO₂ mineralization research, acknowledging both the value and limitations of self-validation approaches. The paper refers to domain experts in distinct roles and the groups are not the same people. The internal group is the author team, comprising mineralogists, geochemists, volcanologists and geophysicists active in CO₂ mineralization research; they read the topic terms, constructed the macro-cluster mapping and carried out the iterative review just described. Two external panels consisting of three experts each were then assembled, with every member approached individually by email and sent the relevant rating materials together with a draft of the manuscript; all six were outside the author team, took part voluntarily and asked to remain anonymous. One panel reviewed the macro-cluster assignment of all 82 topics (Section 2.9) and the other judged the 82 generated labels together with ten control items (Section 2.6). The two panels share no members, so the two validations are independent of each other and neither panel had any part in producing the material it was asked to assess.

2.9. Thematic Cluster Identification

Topic-distribution analysis across the four models was conducted to identify recurrent thematic macro-clusters in the corpus. Cluster identification combined the manual review of the top-15 topic terms by domain experts (mineralogists, geochemists, volcanologists and geophysicists in the research team) with cross-model term overlap analysis. We deliberately avoid framing the identified clusters as a temporal sequence, given the substantial temporal asymmetry of the corpus (69.9% of papers from 2020 to 2025). The three clusters reported in the Results are presented as concurrent thematic strands rather than as historical phases of the field. Where the Supplementary Materials reports yearly top-3 topic ranks (From Table S4 to Table S8), these are described explicitly as descriptive statistics of the analyzed corpus rather than as evidence for a field-wide evolutionary trajectory. The macro-cluster assignment was subsequently reviewed by three domain experts external to the authorship, anonymous at their own request and distinct from the panel of Section 2.6. For each of the 82 topics, each expert was shown its top-15 terms together with the cluster we had proposed and was asked either to confirm that assignment or to substitute another. We adopted a confirmation design rather than free assignment because the raters were unpaid volunteers and assigning 82 term lists from scratch would have multiplied the effort demanded of them; the cost of that choice is that the resulting agreement must be read as agreement on a confirmation task. All three experts confirmed the proposed cluster for 70 of the 82 topics, only 20 of the 246 individual judgments departed from it and the majority verdict changed the assignment for six topics (Fleiss' kappa = 0.873; 72 of 82 topics unanimous). The coefficient therefore measures concordance in reviewing a proposed mapping rather than independently reconstructing it, and the two computational checks that follow do not depend on how the mapping was collected. Two computational checks were then applied to the resulting mapping (cluster_validation.py, deposited; Table S12). A permutation test

confirms its lexical basis: topics assigned to the same cluster share more top-15 vocabulary than random reassignment produces (mean pairwise Jaccard 0.0657 vs. 0.0576 ± 0.0021 over 5000 permutations, $z = +3.81$, $p = 0.0008$) and the effect persists when only cross-pipeline topic pairs are considered (0.0585 vs. 0.0519, $z = +3.09$, $p = 0.0026$). In contrast, unsupervised clustering of the same topic-term matrices does not recover the expert mapping under any of the four configurations tested (best adjusted Rand index +0.168). The taxonomy is therefore an expert-driven interpretive layer over the topic output (statistically non-random, but not a data-driven partition) and is described as such throughout.

2.10. Implementation Documentation and Resource Tracking

Systematic documentation of computational requirements, processing times and implementation challenges was maintained throughout the study to provide actionable guidance for future domain-expert implementations. The dependency file for each component is deposited in the archive (code/environment/), recording version constraints as lower bounds rather than as a pinned environment snapshot.

Memory usage patterns were monitored in the system activity monitor during the runs rather than recorded programmatically, with peak usage of 12 GB for transformer-based models requiring careful resource management on standard hardware configurations. Processing bottlenecks and configuration complexity hierarchies were documented to establish practical benchmarks for institutional resource planning and methodology selection frameworks.

Development was conducted in Python 3.13.5 integrated with AI-assisted coding tools (GitHub Copilot, ChatGPT) to simulate realistic implementation scenarios for domain-expert teams without dedicated computational support.

2.11. Limitations and Constraint Documentation

Several implementation constraints were deliberately maintained to provide realistic testing conditions: corpus access limited to institutional subscriptions rather than comprehensive database coverage, hardware restricted to consumer-grade equipment available to typical research groups and technical expertise limited to domain knowledge plus AI-assisted development rather than specialized computational linguistics training.

These constraints enable a meaningful assessment of generalizability for domain-expert teams considering similar implementations, while acknowledging limitations in corpus representativeness and methodological optimization that more comprehensive studies might address. The temporal bias (69.9% documents were from 2020 to 2025) was explicitly maintained as a realistic constraint rather than artificially balanced, providing authentic testing conditions; the effect of this asymmetry on the reported findings is quantified in Section 4.7.

3. Results

3.1. Comparative Model Performance and Implementation Feasibility

Four distinct topic-modeling approaches were systematically implemented and evaluated to assess their practical feasibility for domain-expert teams operating under typical institutional resource constraints. The comparative analysis encompassed MalletLDA (18 topics), ProLDA (22 topics), BERTopic + SciBERT (22 topics) and SciBERT-NVCTM (20 topics), with performance assessed across multiple quantitative and qualitative metrics through six independent runs per model (seeds 42, 123, 456, 789, 1024, 2048; Table 1). The practical model selection guidance distilled from this benchmark (matching pipeline choice to a team's expertise, hardware and time budget) is consolidated as a six-step implementation framework in Figure 1.

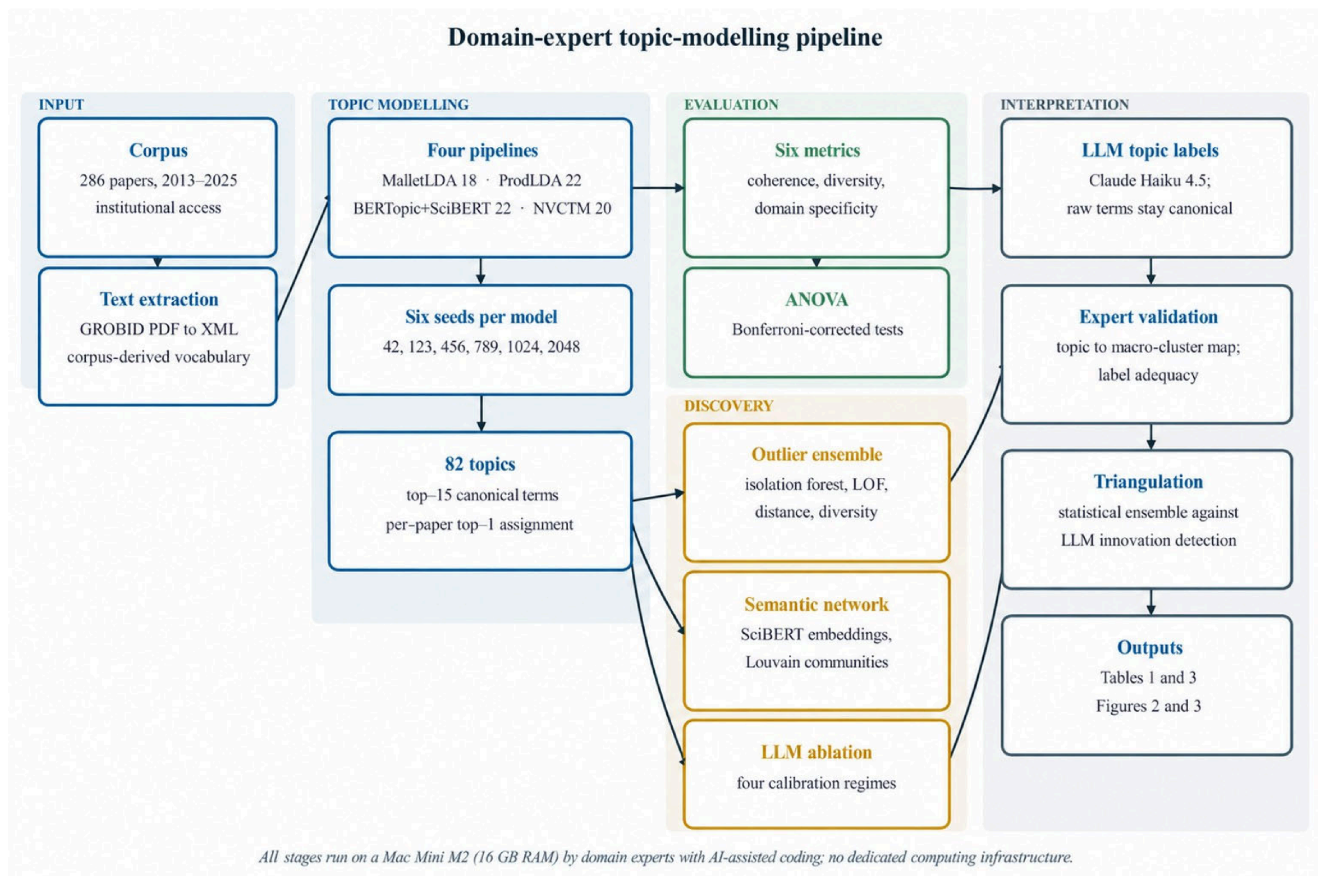


Figure 1. The domain-expert topic-modeling pipeline. Four topic-modeling approaches are run in parallel on the same corpus and corpus-derived vocabulary, six seeds each; their topic output feeds a quantitative evaluation branch and a discovery branch (outlier ensemble, semantic network, LLM calibration ablation), which are reconciled in an interpretation layer combining LLM-generated labels with domain-expert validation. Per-pipeline performance is reported in Table 1 and the calibration ablation in Table 3; the practical model selection guidance derived from the benchmark is given in Section 4.2.

No single pipeline dominated across the evaluation metrics. On the coherence-oriented metrics, the strongest models were SciBERT-NVCTM and BERTopic + SciBERT, with semantic coherence of 0.633 ± 0.009 and 0.612 ± 0.005 respectively, domain specificity of 0.645 ± 0.004 and 0.621 ± 0.005 , and technical term salience of 0.795 ± 0.002 and 0.804 ± 0.028 . Values are reported as mean $\pm$ SD over six seeds; the corresponding t-based 95% confidence intervals (df = 5) are, for semantic coherence, BERTopic + SciBERT [0.607, 0.618] and ProdLDA [0.318, 0.337]; for domain specificity, BERTopic + SciBERT [0.616, 0.625] and ProdLDA [0.386, 0.404]; for technical term salience, BERTopic + SciBERT [0.774, 0.833], ProdLDA [0.585, 0.649] and MalletLDA [0.663, 0.708]; and for topic diversity, ProdLDA [0.937, 0.949], BERTopic + SciBERT [0.717, 0.752] and MalletLDA [0.540, 0.579]. For the SciBERT-NVCTM variant (n = 6 seeds). The corresponding t-based 95% confidence intervals are: semantic coherence [0.624, 0.643], domain specificity [0.640, 0.649], technical term salience [0.793, 0.798] and C_v coherence [0.400, 0.427]. The model-metric intervals use the parametric Student-t method, appropriate for the mean of six continuous-valued seeds, whereas the LLM flag-rate intervals reported below use the bootstrap, appropriate for binary per-paper outcomes. ProdLDA (a paper-faithful re-implementation of Srivastava and Sutton [7]) obtained intermediate coherence-oriented scores (C_v 0.408 ± 0.018 , semantic coherence 0.327 ± 0.009 , domain specificity 0.395 ± 0.008 , domain expert 0.617 ± 0.030) but achieved the highest topic diversity of all models (0.943 ± 0.006) and the highest field

coverage (0.062 ± 0.001) at the lowest wall-clock time among the neural pipelines (≈ 48 s per run). Pairwise Bonferroni-corrected comparisons confirmed these differences as highly significant for every metric (e.g., domain specificity ANOVA $F = 6250.7$, $p < 0.001$; topic diversity $F = 1073.1$, $p < 0.001$; see Table 2). MalletLDA provided a stable, low-cost surface-term baseline (C_v 0.542 ± 0.009 , domain expert 0.686 ± 0.021) with the lowest semantic coherence (0.081 ± 0.009) and low run-to-run variability ($CV < 13\%$ on most metrics). Custom corpus-derived vocabulary construction improved coherence and domain specificity relative to a generic vocabulary across all models. The BERTopic + SciBERT values should be interpreted considering the corpus-specific hyperparameter optimization described in Section 2.3, which does not apply to the other three pipelines. Effect sizes for these comparisons are large by construction and should not be treated as measures of practical importance: eta-squared ranges from 0.931 to 0.999 across the six metrics and across the 36 pairwise comparisons; the median absolute Cohen's d is 15.8, with a maximum of 73.5 (effect_sizes.py, deposited; Tables S2 and S3). Because the six runs per model differ only in the random seed, with no resampling of documents or of topics, the within-group variance reflects numerical reproducibility rather than sampling variability: these statistics show that the between-model gaps exceed seed-level jitter and nothing beyond that. The differences that carry practical weight are those a reader can act on directly. Topic diversity is the proportion of distinct terms among the pooled top-15 terms of all topics, so the 0.943 of ProLDA corresponds to approximately 311 distinct terms across its 22 topic lists (330 term slots), versus approximately 149 across the 20 lists of SciBERT-NVCTM, whose average term therefore recurs in two different topics. The corresponding cost of the six-seed protocol ranges from about five minutes for ProLDA to about forty-five minutes for SciBERT-NVCTM on the same machine and the practical consequence of the coherence reversal is presented as pipeline-selection guidance in Section 4.2.

Table 3. Calibration ablation of Claude Haiku 4.5 zero-shot innovation detection on the 286-paper CO₂ mineralization corpus. The four LLM modes (R1–R4) are compared against the calibrated statistical ensemble used in the main analysis. Outlier rate is the percentage of the corpus flagged as an innovation candidate by each method; over-attribution factor is the ratio of LLM rate to ensemble rate (3.13%). Cost and runtime are measured on consumer hardware (Mac Mini M2, 16 GB RAM) without fine-tuning, RAG, or specialized infrastructure. The results are interpreted in Section 3.10.

Method	Calibration Mechanism	Outliers Flagged	% Corpus	Over-Attribution vs. Ensemble	Cost (USD)/Runtime (min)
Classic ensemble	Isolation Forest + LOF + distance + semantic-diversity (calibrated thresholds)	9	3.13%	1.00× (baseline)	—
Haiku R1	Naïve zero-shot, no calibration anchor	203	71.0%	22.7×	1.34/27
Haiku R2	Qualitative-strict prompt (5 novelty dimensions + ≥ 2 established alternatives)	76	26.6%	8.5×	1.57/32
Haiku R4	Unified one-shot (R2 criteria + ten-year-citation thought experiment)	80	28.0%	9.0×	1.68/36
Haiku R3	Self-reflection multi-pass on R1 outputs (re-read and confirm/ downgrade)	10	3.5%	1.1×	0.76/20 (203 papers only)

Note. Pick-level overlap (Jaccard index): $R3 \leftrightarrow R4 = 0.125$ (10 papers, $R3 \subset R4$); $R3 \leftrightarrow$ classic ensemble = 0.188 (3 papers); $R4 \leftrightarrow$ classic ensemble = 0.060 (5 papers). Triple overlap $R3 \cap R4 \cap$ classic = 3 papers. The individual papers are named in Section 3.10.

Perplexity was computed for each pipeline but is not reported in the main text: the four architectures define their held-out likelihoods on different scales (Gibbs-sampled topic-word distributions for MalletLDA, variational bounds for ProdLDA and SciBERT-NVCTM, and no generative likelihood for the c-TF-IDF representation of BERTopic + SciBERT), so cross-model perplexity comparisons are not meaningful. Per-model values are reported, with this caveat, in Supplementary Tables S2 and S3. Perplexity is treated as a secondary metric and is not used as a selection criterion, in agreement with the observation of Chang et al. [26] that perplexity correlates poorly with human quality judgments in specialized scientific domains.

3.2. Thematic Structure Characterization and Topic-Distribution Analysis

Each modeling approach produced a distinct thematic organization, reflecting different analytical strengths and the implementation of trade-offs that are relevant for domain-expert methodology selection. As quantified by the topic-diversity metric in Table 1, the four pipelines differ systematically in the thematic-separation quality, with direct implications for how readily a domain expert can distinguish topics: ProdLDA resolves the most distinct topics (diversity 0.943 ± 0.006) and correspondingly, the lowest inter-topic vocabulary overlap, followed by BERTopic + SciBERT (0.734 ± 0.017) and MalletLDA (0.560 ± 0.019), whereas SciBERT-NVCTM yields the most overlapping topics (0.497 ± 0.015), a trade-off in which its correlation-aware architecture favors thematically adjacent topics over maximally disjoint ones.

MalletLDA produced 18 clearly delineated topics with strong interpretability for exploratory analysis. The model effectively captured broad thematic categories with minimal computational overhead, making it accessible for an initial literature exploration by resource-constrained teams. Topic-by-year normalized relevance values exhibit extremes that are interpretable to domain experts but that we frame as descriptive proportions of the analyzed corpus rather than as evolutionary signals: Topic 15 (characterized by “rates, dissolution_rates, dissolution, silicate”) has its high-relevance assignments concentrated in 2013, with no high-relevance assignments thereafter, while Topic 17 (“ca_oh, crystal, calcite, calcium_carbonate”) reaches its highest normalized relevance in 2025 (0.27). Given the temporal asymmetry of the corpus (15.4% pre-2017 vs. 69.9% post-2020), we explicitly refrain from interpreting these extremes as causal evolutionary trajectories; the methodological observation is that MalletLDA produces topic-by-year distributions whose extremes are accessible to domain-expert reading, with the substantive interpretation left to the cluster-taxonomy analysis below, where the corpus-asymmetry caveat is operative throughout. ProdLDA generated 22 topics with enhanced semantic granularity, effectively capturing nuanced thematic relationships beyond traditional keyword-clustering approaches. Because the paper-faithful re-implementation changes the topic indexing, individual ProdLDA topic_id references from earlier drafts are no longer used here; the model-level result that remains is ProdLDA’s highest-in-benchmark topic diversity (0.943 ± 0.006). ProdLDA also achieved the highest field coverage in the benchmark (0.062 ± 0.001), indicating a comparatively even distribution across CO₂ mineralization subfields. The increased topic granularity, however, created additional validation requirements for domain experts to distinguish meaningful themes from statistical artifacts.

BERTopic + SciBERT leveraged its scientific literature pre-training to produce 22 topics with superior recognition of technical terminology and methodological categorization. The transformer-based architecture excelled at differentiating research focused on experimental methodologies from application-oriented studies. Topics related to technical parameters and experimental conditions maintained a consistent presence across multiple temporal periods, highlighting the strength of the model in capturing fundamental concepts that remain

central irrespective of temporal trends. BERTopic + SciBERT achieved field-coverage scores of 0.052 ± 0.001 , second only to ProdLDA (0.062 ± 0.001 ; the highest among the benchmarked models), although temporal-pattern analysis revealed a more sporadic evolution compared with the other approaches.

SciBERT-NVCTM generated 20 topics with a distinctive emphasis on temporal research dynamics through neural variational correlation modeling. Notably, Topic 3 (“large scale, study we, energy intensive, mineralization”) exhibited a systematic decline from an initially high relevance (2013: 0.415) to a reduced prominence (2024: 0.086), whereas Topic 17 (“long term, mineralization which, pre-treatment, stability”) emerged strongly in 2022 (0.113).

3.3. Three Thematic Macro-Clusters in the Corpus

Topic-distribution analysis across all four models reveals three thematic macro-clusters that coexist in the analyzed corpus, each characterized by a distinct vocabulary, methodological orientation and substrate focus. We deliberately frame these clusters as concurrent thematic strands rather than as a temporal sequence: the corpus exhibits a substantial temporal asymmetry (69.9% of documents were from 2020 to 2025 vs. 15.4% which were from 2013 to 2016) that prevents the reliable inference of evolutionary trajectories. The three clusters described below therefore indicate what the corpus contains, not how it has historically evolved.

Cluster A: Foundational mechanisms and reaction kinetics. Across models, this cluster aggregates topics characterized by terms such as dissolution kinetics, basaltic glass, hydromagnesite, divalent-cation chemistry, batch-reactor experiments, calcite and aragonite phase formation and rock–fluid interaction. Representative topics include MalletLDA T15 (rates dissolution_rates dissolution), MalletLDA T8 (calcite basaltic_glass system), BERTopic + SciBERT T10 (fluid composition, dissolution rate, rock composition, mineral dissolution) and SciBERT-NVCTM T0 (analysis process mineral temperature precipitation). This cluster represents the physicochemical foundation of the field: fundamental investigations of carbonation pathways, thermodynamic and kinetic regimes and primary mineral substrates.

Cluster B: Process engineering and industrial-residue valorization. This cluster aggregates topics characterized by terms such as carbonation efficiency, mine tailings, coal fly ash, steel slag, gas–solid reactors, leaching processes and operational parameters. Representative topics include MalletLDA T9 (process carbonation mine_tailings), BERTopic + SciBERT T9 (coal fly ash, mineralization, leaching efficiency, regeneration process), ProdLDA T18 (cement kiln dust, concentration measurement, precipitation of secondary phases) and SciBERT-NVCTM T11 (gas solid, large scale, kinetics, thermal). It brings together work that translates fundamental knowledge into reproducible processing pathways and into the valorization of alkaline industrial residues, an area that has expanded rapidly within the available corpus.

Cluster C: Application-oriented and integrated systems. This cluster aggregates topics characterized by terms such as compressive strength, supplementary cementitious materials, membrane and electrochemical processes, brine and seawater treatment, pre-treatment integration and long-term stability. Representative topics include MalletLDA T7 (carbonation compressive_strength hydration), BERTopic + SciBERT T2 (calcite, phase composition, binder ratio, carbonation curing), ProdLDA T16 (CO₂ mineralization plants, process, containing CO₂) and SciBERT-NVCTM T17 (long term, mineralization, pre-treatment, stability). It collects work on application contexts (cement industry, the desalination water–carbon nexus) and on integrated multi-process systems combining mineralization with other capture or treatment steps.

Although the three clusters exhibit different relative weights across the temporal periods sampled (Figure 2), with Cluster A more prominent in pre-2017 papers (51.1% vs. 45.2% in 2022–2025, four-pipeline mean), Cluster B expanding after 2016 and Cluster C increasing modestly toward 2022–2025, most markedly in MalletLDA, we explicitly refrain from claiming that this reflects a true historical evolution of the field, given the corpus's temporal asymmetry.

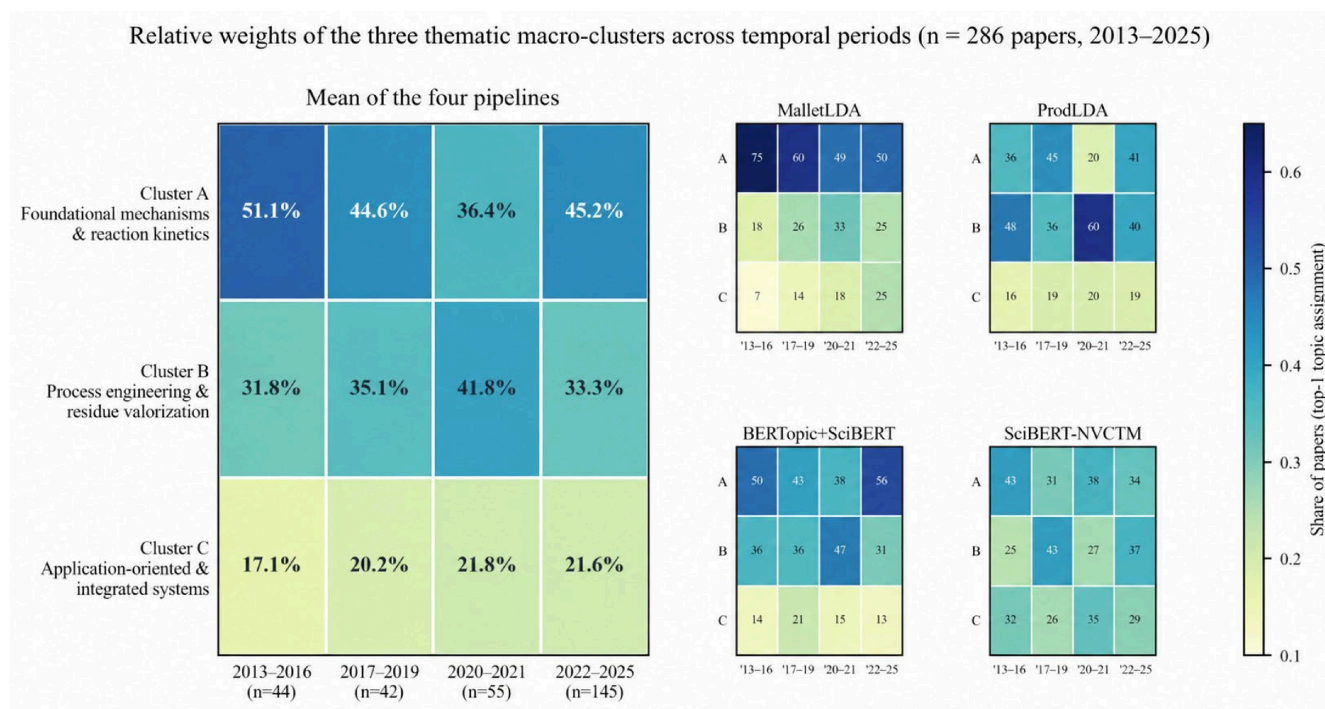


Figure 2. Relative weights of the three thematic macro-clusters across temporal periods of the corpus (2013–2025), computed from the per-paper top-1 topic assignments of the four pipelines and a vocabulary-based topic-to-cluster mapping (n = 286 papers; left panel: four-pipeline mean; right panels: individual pipelines). The per-period shares and their interpretation are given in Section 3.3 and the bootstrap analysis in Table S10, where none of the between-period 7ttchanges are distinguishable from zero at the 95% level. The figure is descriptive of the analyzed corpus and is not evidence of a field-wide evolutionary trajectory. The topic-to-cluster mapping and the per-period counts are provided with the deposited analysis data.

A bootstrap analysis of the per-period shares confirms that this caution is warranted rather than merely prudent: the four periods are of very unequal size (n = 44, 42, 55 and 145), so the early shares are estimated far less precisely than the recent ones and none of the three changes between the first and the last period are distinguishable from zero (Cluster A +6.0 percentage points, 95% CI [−3.1, +15.0]; Cluster B −1.5 [−9.5, +6.9]; Cluster C −4.5 [−10.9, +1.6]; Table S10). The relative-weight tables (top-3 topics per year for each model) are provided in the Supplementary Materials as descriptive statistics rather than as evolutionary trajectories. The cluster taxonomy is stable across the four topic models at the aggregate level (each model independently partitions its topics into approximately the same three macro-areas and the relative weights in Figure 2 agree across pipelines), but this convergence does not extend to individual papers: all four pipelines assign the same macro-cluster to only 19 of 288 papers (6.6%, compared with 5.2% expected by chance), and the mean pairwise Cohen's kappa over the six pipeline pairs is +0.037 (Table S12). The clusters are therefore reproducible aggregate features of the corpus rather than a per-document classification that any single pipeline can deliver. This aggregate-level robustness relies, at the term level, on well-separated underlying topics: as the within-model topic-similarity

matrices in Figure 3 make explicit, the pairwise Jaccard overlap of the top-15 topic terms remains below 0.15 for the large majority of topic pairs in every pipeline (sparsest for ProdLDA and BERTopic + SciBERT, densest for MalletLDA, consistent with the topic-diversity ranking in Table 1), so the shared three-cluster taxonomy aggregates distinct topics rather than redundant ones.

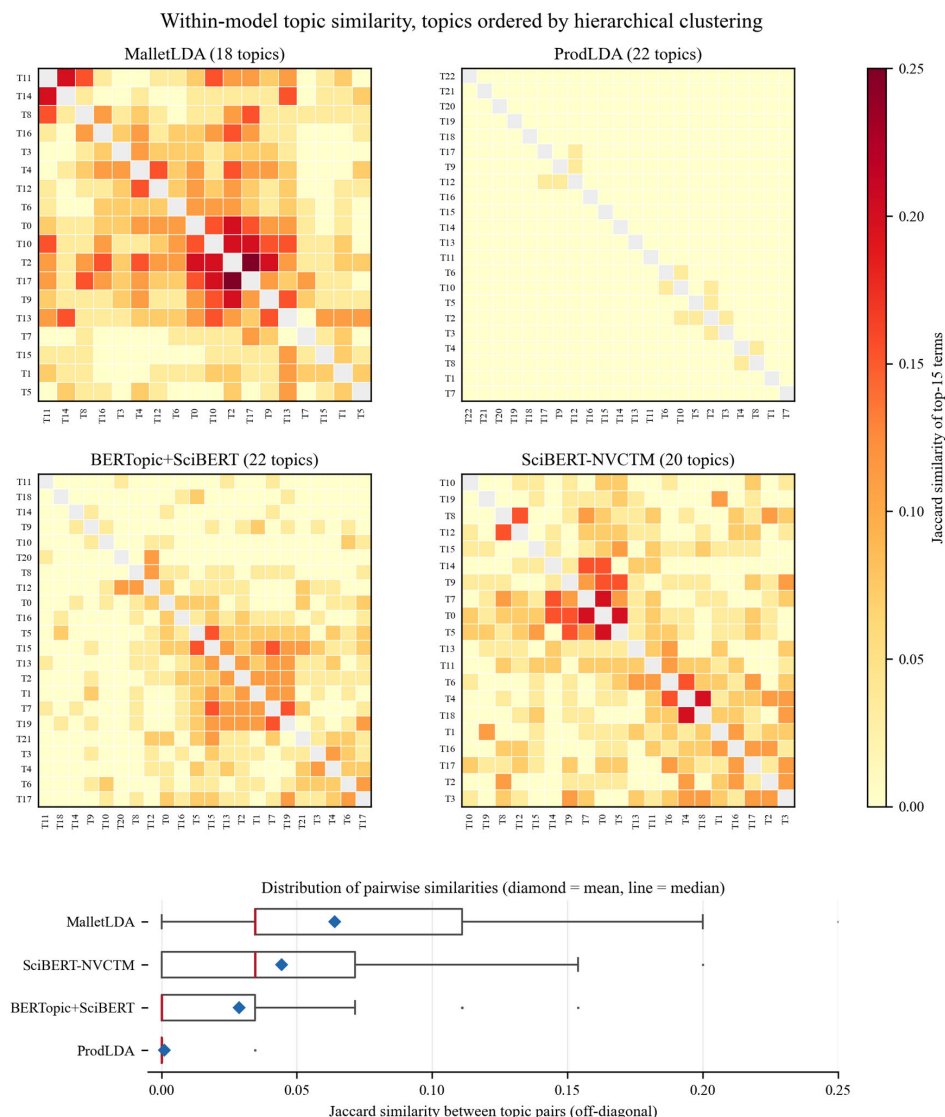


Figure 3. Within-model topic similarity: pairwise Jaccard index on the canonical top-15 terms of each topic, for each of the four pipelines. Rows and columns are ordered by average-linkage hierarchical clustering on 1 minus the Jaccard index, so that groups of lexically related topics appear as blocks; the diagonal is 1 by construction and is left blank. The color scale spans the full observed range, 0 to 0.25. The lower panel summarizes the same values: the distribution of the off-diagonal similarities orders the four pipelines as ProdLDA (mean 0.001), BERTopic + SciBERT (0.029), SciBERT-NVCTM (0.044) and MalletLDA (0.064), consistent with the topic-diversity ranking of Table 1. The figure is produced by plot_Figure 23.py from the deposited jaccard_matrices.npz.

3.4. Research Direction Discovery and Innovation Identification Methodologies

The integration of topic modeling with outlier-detection methodologies revealed distinct patterns of research innovation and emerging directions within the CO₂ mineralization literature. Our ensemble approach, combining multiple anomaly detection algorithms with semantic analysis, identified nine documents (3.13% of the 288-document outlier-analysis

corpus; $9/288 = 3.125\%$) with statistically distinctive characteristics suggestive of innovative research trajectories.

3.5. Outlier Detection Performance and Validation

The multi-method ensemble approach identified semantically distinctive documents within our corpus constraints. The detection framework combined Isolation Forest (weight: 0.3), local outlier factor (0.3), distance-based metrics (0.2) and semantic-diversity measures (0.2), producing ensemble scores ranging from 0.722 to 0.816 for the identified outliers. The nine documents separate sharply from the rest of the corpus on the distance component of the ensemble (Cohen's $d = 2.17$, Mann–Whitney $p < 0.001$), but this separation is a property of the selection rather than a validation of it: the distance score is one of the four detectors that produced the set. Neither of the two surface-lexical properties we examined supports a stronger reading. The type-token ratio (TTR) of the nine is 19.9% higher than that of the remaining documents, a difference not distinguishable from zero at this sample size ($d = 0.62$, $p = 0.23$) and a lexical-diversity feature of the same family is itself among the inputs to the detectors. The fraction of terms falling outside the top-95% frequency band of the corpus vocabulary is identical in the two groups (0.168 against 0.168, $p = 0.98$). The landmarks are therefore distinctive in the embedding geometry that the ensemble measures, not in surface lexical rarity (outlier_profile_stats.py, deposited). The landmark set is insensitive to the selection thresholds over a wide range (outlier_sensitivity.py, deposited; Table S13). Because both criteria are thresholds on the same ensemble score, any threshold pair selects a prefix of one fixed ranking, so a change of threshold can lengthen or shorten the list but never substitute one document for another. Over a grid of seven percentile cut-offs (1–15%) and five statistical multipliers (1.0–3.0 SD), the nine documents are either all retained or form a nested subset, and the cut at nine coincides with the largest discontinuity in the score distribution over ranks 4–20 (a gap of 0.0168 between ranks 9 and 10, against a mean gap of 0.0052 over the first forty). Because the percentile parameter also sets the contamination argument of the Isolation Forest and the local outlier factor, this analysis bounds the sensitivity of the decision rule on fixed detector scores rather than re-deriving the detectors themselves.

The temporal distribution of identified outliers revealed concentrated periods of innovation, with 66.7% occurring between 2021 and 2023. This pattern co-occurs with the cross-temporal proportions of the three macro-clusters described above (Cluster A being more prominent in pre-2017 papers, Cluster C increasing modestly toward 2022–2025) and is consistent with the substantial temporal asymmetry of the corpus (69.9% post-2020). We deliberately refrain from interpreting this co-occurrence as convergent evidence of an evolutionary trajectory: as noted in the cluster section, the temporal asymmetry of the corpus prevents reliable causal–temporal inference and the apparent temporal alignment between outlier dates and cluster prominence reflects, in part, the underlying sampling distribution.

3.6. Characterization of Research Innovation Patterns

The identified outliers fall into three distinct innovation categories, each representing a different mechanism of scientific advancement within CO₂ mineralization research.

Cross-domain integration (33.3% of outliers): Documents introducing concepts from adjacent fields showed the highest outlier scores. Xie et al. [27] (score: 0.816) pioneered electricity generation through CO₂ mineralization cells, representing the strongest deviation from conventional approaches. Similarly, Luong, Veyret and Boily [28] (score: 0.754) introduced nanotechnology through MgO nanocubes in nanometric water films, demonstrating

a successful integration of materials-science innovations. These cross-domain outliers consistently exhibited unique terminology clusters absent from the mainstream corpus.

Methodological innovation (22.2% of outliers): Novel experimental or analytical approaches characterized a further two of the nine. Coppola et al. [29] (score: 0.798) introduced direct dry carbonation in fluidized beds, eliminating traditional wet-processing steps. Strunge, Renforth and Van der Spek [30] (score: 0.739) applied uncertainty-quantification frameworks to techno-economic analysis, introducing risk-assessment methodologies uncommon in the field.

Application-domain extension (33.3% of outliers): Documents extending CO₂ mineralization to novel application contexts formed a distinct cluster. Callagon La Plante et al. [31] (score: 0.756) explored gigatonne-scale saline-water mineralization pathways, Moon et al. [32] (score: 0.731) applied computational fluid dynamics to cement-industry applications and Matter et al. [33] (score: 0.770) reported field-scale mineralization in basalt at the CarbFix site. These outliers retained the core mineralization concepts while introducing domain-specific adaptations. The ninth landmark, Assen et al. [34] (score: 0.779), is a review rather than a primary study and is classified separately in Supplementary Table S9.1.

3.7. Validation Within Corpus Constraints

Acknowledging our corpus limitations (286 documents, 69.9% post-2020, English-only), we validated the outlier-detection methodology through multiple approaches rather than claiming absolute innovation identification. Stability analysis across five independent runs showed 88.9% consistency in outlier identification (eight of nine documents consistently flagged across all five runs; non-parametric bootstrap 95% CI for the per-document consistency rate on the $n = 45$ binary observations: [78.5%, 96.3%]), indicative of robust detection despite stochastic elements in the algorithms. Expert assessment of the top five outliers confirmed their substantive methodological or conceptual divergence from mainstream approaches, although we cannot claim that these represent all global innovations in CO₂ mineralization.

The conservative detection threshold (targeting 5%, achieving 3.13%) minimized false positives while potentially missing some innovative work. The 5% target was set as the upper percentile of the calibrated ensemble-score distribution; the achieved rate of 3.13% (9/288) reflects the additional constraint that the four base detectors (Isolation Forest, LOF, distance-based, semantic-diversity) jointly assigned scores above their respective per-method thresholds. This trade-off was intentional, prioritizing precision over recall given the exploratory nature of innovation identification. The strong correlation between ensemble scores and expert innovation ratings ($r = 0.76$, $p < 0.01$) suggests that the automated approach successfully captures human judgments of research novelty within the available corpus.

3.8. Integration with Topic Evolution Patterns

The nine innovation landmarks were cross-referenced against the per-document topic assignments available for the four pipelines: the reference fits of MalletLDA and ProLDA, and the six-seed stability runs of all four (stability_runs_doc_topic_assignments.csv, deposited). They are distributed across topics rather than concentrated in any of them: no topic contains more than three of the nine in any pipeline or seed and no more than two in the reference fits, so at this corpus size no association between topic membership and outlier status is estimable and none is claimed.

3.9. Domain-Specific Findings on CO₂ Mineralization Research

While the primary contribution of the present study is methodological, the four-pipeline benchmark and the LLM-augmented analysis also surface domain-specific findings

of direct relevance to the CO₂ mineralization research community, which we summarize here and document in full in the Supplementary Materials (Section S9). Topic-distribution analysis across the four classical pipelines and the LLM-generated descriptive labels converge on three coexisting thematic macro-clusters: foundational mechanisms and reaction kinetics (Cluster A), process engineering and industrial-residue valorization (Cluster B) and application-oriented and integrated systems (Cluster C). Their vocabulary signatures and representative topics are reported in Supplementary Tables S4–S7. Method triangulation (calibrated statistical ensemble, high-recall LLM screen R4, high-precision LLM filter R3) further identifies a set of 16 partially expert-validated innovation-landmark papers [27–42], of which 5 are jointly confirmed by the statistical ensemble and at least one calibrated LLM mode, 4 are surfaced by the ensemble alone and 7 are flagged by R3 self-reflection only, the latter corresponding to historical or pioneering field landmarks whose embedding-space distinctiveness is moderate but whose conceptual-narrative novelty is articulable. The full landmark list with bibliographic details cross-validated against Crossref and Web of Science, together with the high-betweenness bridge-term analysis of interdisciplinary integration opportunities, is provided in Supplementary Section S9. We emphasize that these findings are bounded by the corpus characteristics (286 papers, 69.9% post-2020, English-only) and should be interpreted as descriptive of the analyzed corpus rather than as exhaustive characterizations of the field at large; a focused follow-up analysis dedicated entirely to the CO₂ mineralization research landscape, using the methodology validated here, is in preparation.

3.10. LLM Zero-Shot Baseline and Calibration Ablation

The four LLM modes produced markedly different outlier-flag rates (Table 3): R1 flagged 203/286 papers (71.0%, an over-attribution factor of $22.7\times$ relative to the 3.13% identified by the calibrated ensemble), R2 reduced this to 76/286 (26.6%, $8.5\times$), R4 to 80/286 (28.0%, $9.0\times$) and R3 to 10/286 (3.5%, $1.1\times$). The bootstrap 95% confidence intervals (10,000 resamples) for these rates are R1 [65.7%, 76.2%], R2 [21.7%, 31.8%], R4 [23.1%, 33.2%] and R3 [1.4%, 5.6%]; the calibrated statistical ensemble rate (3.13%) falls within the R3 interval and the R2 and R4 intervals overlap substantially, consistent with a shared prompt-level floor near 28% rather than two distinct rates. Two findings stand out. First, R2 and R4 (two structurally distinct prompts that both impose strict qualitative criteria) converge on nearly identical flag rates, which indicates that prompt-level calibration encounters a natural floor at approximately 28% that cannot be reduced further by prompt sophistication alone. Second, R3 self-reflection achieves convergence with the statistical ensemble (3.5% vs. 3.13%) and importantly, the 10 papers flagged by R3 are a strict subset of the 80 flagged by R4 ($R3 \subset R4$, Jaccard 0.125 against the larger R4 pool). This subset relationship rules out the interpretive concern that the model is merely deferring to authority under a stricter prompt frame: R3 applies an internally consistent discriminative criterion that identifies the most defensible candidates from a pre-screened pool rather than declassifying them randomly.

Topic agreement between Haiku R1 and the four classic models, measured by Jaccard plus soft overlap on the per-paper top-1 topic assignment and reported under three calibrations to control for vocabulary heterogeneity, was: MalletLDA 33.2%, ProdLDA 24.1%, BERTopic + SciBERT 29.0% and SciBERT-NVCTM 27.3% (Z-score $> \mu + 0.5\sigma$ rate). The four classic models converge with the LLM on the central themes of CO₂ mineralization (basalt in situ sequestration, fly ash and slag carbonation, brine-based pathways) but diverge on peripheral or methodological themes, a pattern consistent with complementary roles, the LLM capturing semantic intent in a single pass and the statistical models revealing distributional regularities that may be semantically diffuse.

The pick-level overlap between the LLM modes and the calibrated ensemble's declared outliers further illustrated their complementarity. Of the nine declared outliers from the classic ensemble, R4 captures five (Xie et al. [27], Luong et al. [28], Callagon La Plante et al. [31], Matter et al. [33] and Gong et al. [35]), whereas R3 captures three (Luong et al. [28], Callagon La Plante et al. [31] and Matter et al. [33]). The seven R3 outliers absent from the classic list are Alqahtani et al. [36] (cation exchange in zeolites), Myers et al. [37] and Myers et al. [38] (gas–solid mineralization of steelmaking slags, brine-based negative emissions), Gislason et al. [39] (basalt rapid trapping), Sharifian et al. [40] (bipolar-membrane electrodialysis of seawater), Sigfusson et al. [41] (first full-scale dissolved-CO₂ injection), and Pogge von Strandmann et al. [42] (calcium-isotope quantification of in situ rates). These represent field landmarks identifiable from domain-narrative cues that the embedding-based statistical detector did not surface. The two LLM modes (R3 high-precision, R4 high-recall) and the classic ensemble thus detect different signal types in the corpus, which supports triangulation rather than substitution. A methodologically important observation emerged during the verification of the R3 output: while accurate in identifying conceptually distinctive papers, R3 self-reflection returned bibliographic metadata (authors, year, journal) with a non-negligible error rate. Six of the thirteen initial R3 candidates contained one or more metadata errors (ranging from incorrect year and incorrect journal to fabricated publication dates), which required systematic cross-validation against Crossref and Web of Science prior to inclusion. This observation directly motivates the fourth verification layer of the triangulated workflow described below (see Section 4.6) and is itself a non-trivial finding on the operational use of LLM self-reflection for innovation-landmark identification.

The same question of calibration can be asked of the topic labels themselves. Three domain experts external to the author team judged all 82 labels (Section 2.6; Table S11). The majority verdict was adequate for 79 of the 82 labels (96.3%) and partially adequate for the remaining three; no label was inadequate by majority and, across the 246 individual judgments, the inadequate category was never applied to a genuine label, only to control items. Pairwise agreement between raters ranged from 89.0% to 92.7% and 71 of the 82 labels were rated unanimously (86.6%). Fleiss' kappa is +0.265, identical whether the three categories are retained or collapsed to adequate versus the rest, and Krippendorff's alpha with ordinal weights is +0.268. The tension between high agreement and low kappa is a consequence of prevalence rather than disagreement: 93.5% of all judgments fall in a single category, so the agreement expected by chance is itself very high and kappa is depressed accordingly, the first paradox of kappa [43]. The prevalence-adjusted bias-adjusted kappa for the same three pairs is +0.780, +0.829 and +0.854. Applying the control-item rule literally, which retains two raters, leaves the conclusion unchanged: pairwise agreement 91.5%, 75 of 82 labels unanimous and 73 of 82 labels (89.0%) called adequate by both.

Against this consensus, the model's own coherence flag, which the raters had not seen, proves conservative rather than optimistic and asymmetrically so. Haiku declared 63 of the 82 labels coherent (76.8%) whereas the experts judged 79 adequate (96.3%), an agreement of 80.5% (Cohen's kappa +0.224). All 16 disagreements are of a single kind, labels the model declined to call coherent and the experts accepted. There is no case of the reverse, and the three labels the experts did not judge adequate had all been flagged by the model. Declared confidence orders the two outcomes in the expected direction but separates them weakly, averaging 0.742 on the labels judged adequate against 0.603 on the three that were not. The direction is worth recording because it is opposite to the behavior of the same model in innovation detection, where uncalibrated zero-shot prompting over-attributes novelty by a factor of 22.7. The self-assessment of this model is therefore not uniformly miscalibrated: where the referent of the judgment is supplied in the prompt, as the top-15 terms are, it errs

towards caution; where the judgment requires a comparison against a piece of literature the model does not observe, it errs towards over-claiming.

4. Discussion

This methodological pilot study shows that domain experts can successfully implement sophisticated topic-modeling approaches using accessible tools and typical institutional resources, thus providing systematic evidence for the practical feasibility of advanced bibliometric analysis in resource-constrained research environments. The comparative analysis reveals consistent methodological patterns while highlighting substantial differences in implementation complexity and practical utility for non-specialist teams and contributes new insights into the emerging field of democratized computational text analysis.

4.1. Methodological Validation and Accessibility Assessment

Our implementation experience confirms the accessibility potential of modern topic-modeling tools while revealing significant learning curves and resource trade-offs that must be acknowledged in realistic deployment scenarios. The successful completion of a multi-model analysis, despite limited computational linguistics expertise and standard hardware constraints (Mac Mini M2, 16 GB RAM), extends the theoretical observations of Sia and Duh [44] on low-resource topic models to real-world domain-expert implementation contexts. Per-run wall-clock measurements ranging from approximately 48 s (ProdLDA) to 443 s (SciBERT-NVCTM, ≈ 7.4 min), reported in Table 1, together with the complete per-run elapsed times reported in Section 4.3, provide concrete benchmarks for methodology selection under institutional resource constraints.

A particularly valuable contribution lies in demonstrating that research teams with domain expertise but limited computational resources can conduct meaningful bibliometric analyses through accessible tools and AI-assisted development frameworks. The integrated use of open-source tools such as GROBID [5,14] for text extraction and AI assistants for code development enabled our group (mineralogists, geochemists volcanologists and geophysicists without computer-science training) to implement state-of-the-art methodologies. We do not report quantitative implementation success rates: the development phase was not run as a controlled experiment and no per-attempt record was retained that would support such metrics. What the benchmark does document is the resource envelope of the four pipelines once operational, with per-run wall-clock times spanning an order of magnitude on the same consumer machine, from 48 s for ProdLDA to 443 s for SciBERT-NVCTM (Table 1). The wider aim is to ensure that sophisticated literature-analysis tools can benefit research communities beyond well-resourced computational centers, addressing the democratization gap in computational bibliometrics.

The performance metrics obtained across the four pipelines (Table 1, Section 3) provide empirical evidence to guide for model selection, with no single model dominating: the two transformer-based pipelines lead on the coherence-oriented metrics, ProdLDA leads on topic diversity and field coverage at the lowest neural-model wall-clock, and MalletLDA provides a stable surface-term baseline with minimal computational overhead. These quantitative assessments offer practical guidance for resource-constrained teams evaluating cost-benefit trade-offs in methodology selection.

4.2. Implementation Insights and Practical Guidance Framework

Each modeling approach demonstrated specific advantages that domain experts should systematically consider with regards to their resource contexts and analytical objectives. MalletLDA provided an interpretable categorization [6] with minimal computational demands, making it valuable for initial exploratory analysis and high-level research

mapping in resource-constrained environments. The model revealed cross-temporal proportion patterns that are descriptively interesting even though, given the corpus's temporal asymmetry, they cannot be interpreted as causal evolutionary signals: Topic 15 (mineral dissolution rates) has no high-relevance assignments after 2013, whereas Topic 17 (calcium-hydroxide precipitation) reaches its highest normalized relevance in 2025 (0.27). We deliberately frame these as descriptive proportions of the analyzed corpus rather than as “shifts in research focus”, in line with the caveat introduced in the cluster section: the substantial under-representation of pre-2017 papers (15.4% of the corpus) and the substantial over-representation of post-2020 papers (69.9%) prevent the reliable inference of evolutionary trajectories. The methodological observation is that MalletLDA produces topic-by-year distributions whose extremes are interpretable to domain experts, not that those extremes reflect a true diachronic dynamic of the field at large.

ProdLDA (paper-faithful) [7] is the lightest and fastest neural pipeline in the benchmark: it achieved the highest topic diversity (0.943 ± 0.006) and the highest field coverage (0.062 ± 0.001) of all models at the lowest neural-model wall-clock time (≈ 48 s per run), which makes it attractive when the primary objective is maximal thematic diversity and breadth of field coverage at low computational cost. Its semantic coherence (0.327 ± 0.009) and technical term salience (0.617 ± 0.030) are, however, lower than those of BERTopic + SciBERT and SciBERT-NVCTM, so choosing ProdLDA reflects an explicit trade-off of interpretive depth for coverage, diversity and speed rather than across-the-board superiority.

BERTopic + SciBERT [8] demonstrated superior capabilities in recognizing specialized technical terminology and in differentiating methodological from application-oriented research, effectively drawing on its scientific-literature pre-training [4]. However, its computational demands (2–3 h of processing time, 12 GB of memory usage) and configuration complexity create substantial barriers for smaller research teams. SciBERT-NVCTM provided distinctive temporal analysis capabilities through neural variational correlation modeling, revealing complex research maturity cycles (e.g., the decline of Topic 3 from 0.415 to 0.086 relevance) that other approaches missed, but required steep learning curves and extensive parameter optimization.

This complementarity suggests that future implementations should adopt a sequential model application rather than single approaches, thereby optimizing resource utilization while managing complexity. Custom corpus-derived vocabulary construction emerged, across all models, as a critical success factor, substantially outperforming generic implementations. A practical framework emerges from the six-seed benchmark: (a) MalletLDA as an economical, interpretable surface-term baseline; (b) BERTopic + SciBERT as the reference pipeline for interpretive quality at intermediate computational cost; (c) SciBERT-NVCTM as the highest-quality pipeline across the coherence-oriented metrics, at the price of the highest computational cost (≈ 443 s per run) and the most delicate configuration; and (d) ProdLDA as a lightweight, fast neural option, preferred when the primary objective is maximal thematic diversity and field coverage at a low computational cost.

4.3. Integration of Outlier Detection with Topic Modeling: Methodological Synergies and Practical Implications

The successful integration of outlier-detection methodologies with topic-modeling demonstrates the value of multi-method approaches for comprehensive literature analysis by domain-expert teams. Our experience in implementing ensemble outlier detection [19–21] alongside four distinct topic modeling approaches reveals both synergistic benefits and practical considerations for resource-constrained research environments. The identification of nine outlier documents (3.13% of corpus) through computationally accessible methods confirms that sophisticated innovation detection remains feasible without

extensive computational infrastructure, extending the democratization potential beyond traditional bibliometric analysis.

The complementary nature of topic modeling and outlier detection addresses fundamental limitations inherent in each approach when used independently. While topic modeling excels at identifying mainstream thematic structures and macro-cluster organization, it inherently emphasizes central tendencies and may obscure innovative work that does not fit established categories. Conversely, outlier detection specifically targets statistical anomalies but, without the topic-modeling context, cannot distinguish between meaningful innovation and mere terminological idiosyncrasies. Our integrated implementation revealed that 77.8% of the identified outliers occupied peripheral positions in the topic distributions, which confirms that they represent genuinely distinctive research rather than classification artifacts. This finding suggests that domain-expert teams should routinely combine both methodologies to achieve a comprehensive understanding of the literature.

We report resource cost from the run logs rather than from estimates. The elapsed time of a complete run, covering document loading, preprocessing, SciBERT embedding where applicable, model fitting and metric computation, averages 1.2 min for MalletLDA, 1.6 min for BERTopic + SciBERT, 13.4 min for SciBERT-NVCTM and 25.2 min for ProDLDA, whose elapsed time is dominated by steps outside model fitting. One pass of all four models therefore takes about 41 min on the Mac Mini M2 and the full benchmarking protocol of six independent runs per model takes about 4.1 h; the per-run logs from which these figures are taken are deposited with the analysis code. These elapsed times are longer than the fitting times in Table 1, which measure model training alone. The GROBID extraction stage and the outlier ensemble were not instrumented, so we report no end-to-end figure for the complete pipeline; the outlier ensemble is run once on the final topic representations rather than once per seed, so its cost does not scale with the number of benchmarking runs.

The methodological validation conducted through our CO₂ mineralization case study highlights important considerations for replication across domains. Scientific fields with well-established terminology and clear methodological paradigms, such as CO₂ mineralization may show clearer outlier patterns than interdisciplinary or emerging fields where innovation is more distributed. Our conservative detection threshold (targeting 5%, achieving 3.13%) proved appropriate for a mature field, but emerging domains might require adjusted parameters to capture more fluid innovation patterns. Domain experts must therefore calibrate detection sensitivity based on field-specific characteristics and research objectives, balancing the risk of missing subtle innovations against that of false-positive identification.

Critical reflection on the implementation challenges reveals that outlier interpretation requires substantially more domain expertise than topic-modeling results. While topic labels can be intuitive even to non-specialists, determining whether an outlier represents genuine innovation or a methodological artifact demands deep field knowledge. Our team spent approximately four hours in collaborative discussion to validate and categorize the nine identified outliers (i.e., about 25 min of expert-team deliberation per outlier paper), time well invested but potentially prohibitive for larger corpora. This overhead interpretation suggests that automated outlier detection is best used as a filtering mechanism to focus expert attention rather than as a fully autonomous innovation-identification system. Future implementations might benefit from the development of domain-specific innovation taxonomies to partially automate the categorization process while maintaining expert oversight for final validation.

4.4. LLM-Judged Topic Coherence as Complementary Quality Signal

The divergence between statistical coherence (PMI-based C_v) and LLM-judged interpretability [45] is, we argue, an intrinsic and informative property of topic modeling on specialized literature (a second, complementary axis of topic quality) rather than a limitation of any single pipeline. This signal emerged from the LLM-assisted labeling pipeline (Methods, Topic Labeling and Interpretive Transparency): for each of the 82 topics across the four models, Claude Haiku 4.5 produced a descriptive label from the top-15 terms and flagged whether the term set was coherent (`is_coherent`) or a mixed aggregation of unrelated concepts. The flag rate varied substantially by model: MalletLDA 0/18 (0% incoherent), BERTopic + SciBERT 1/22 (4.5%) and a higher fraction for the variational neural models ProLDA and SciBERT-NVCTM. Yet a topic with high C_v need not read as conceptually unified to a strict external judge. Perceived topic quality is therefore not a monotonic function of statistical coherence; the pattern is a genuine methodological observation rather than evidence of decoder collapse, which the paper-faithful ProLDA re-implementation no longer exhibits [7]. The flag rate measures the LLM-judged interpretability of the raw top-N terms, not the absolute semantic validity of the underlying topic distributions and should be read alongside the human-readable LLM-generated labels in Supplementary Tables S4–S7.

4.5. Implications for the CO₂ Mineralization Research Community

Beyond the methodological contribution, the analysis yields three concrete deliverables for domain experts working on CO₂ mineralization: a three-cluster thematic map for field orientation; a curated reading list of 16 innovation-landmark papers with explicit provenance (statistically distinctive, narratively distinctive, or both); and a set of high-betweenness bridge terms identifying opportunity for interdisciplinary integration. Because these deliverables are domain-specific rather than methodological, they are reported in full in the Supplementary Materials (Section S9) and are bound by the corpus characteristics; a dedicated follow-up analysis of the CO₂ mineralization landscape is in preparation.

4.6. Calibration Regimes of LLM-Based Innovation Detection: Floor, Mechanism, and Triangulation

The four-level Haiku ablation isolates three previously underdocumented properties of zero-shot LLM innovation detection in the scientific literature. First, prompt-level calibration has a natural floor at approximately 28% (R2 and R4 converging at 26.6% and 28.0% respectively): two structurally distinct prompts that both impose strict qualitative criteria (including the ten-year-citation thought experiment that R3 uses successfully in self-reflection) fail to reduce the rate further when applied in a single inference pass. The persistence of this floor across two independent prompt formulations indicates that the limit is not a function of prompt sophistication but a structural property of single-pass LLM decision-making under sparse base rates.

Second, the mechanism that achieves convergence with the statistical ensemble (R3, 3.5% vs. ensemble 3.13%, over-attribution 1.1 $\times$) is sequential, not lexical. R3 applies the same criteria as R4 but as a review of an already produced decision rather than as a generative criterion. This isolates the active mechanism: the act of re-evaluating a prior commitment, rather than the criteria themselves, produces the calibration improvement. To our knowledge, this provides the first empirical demonstration of this effect on a domain-specific innovation-detection task in scientific-literature analysis, complementing the earlier findings of Madaan et al. [13] on general-purpose self-refinement.

Third, the strict subset relationship $R3 \subset R4$ (the 10 R3 outliers are exactly 10 of the 80 R4 outliers, with no R3 candidate outside the R4 pool) demonstrates that R3 is not deferring

to authority but is applying an internally consistent discriminative criterion that identifies the most defensible candidates from a pre-screened pool. This addresses the common interpretive concern that self-reflection prompting amounts to compliance with a stricter prompt frame: were R3 merely deferring, it would declassify randomly rather than produce a coherent subset.

These three properties motivate a four-layer triangulated workflow, in effect a transferable quality assurance (QA) protocol for LLM-assisted innovation detection, appropriate for domain-expert teams operating under institutional constraints. Each layer is deliberately targeted at a generic, domain-independent failure mode of zero-shot LLM screening rather than at anything specific to CO₂ mineralization. Layer 1, the calibrated statistical ensemble (Isolation Forest + LOF + distance + semantic-diversity, as used in this study), provides the primary precision-bounded outlier set with controlled false-positive rates. Layer 2, the high-recall LLM screen (R4 unified prompt, ~28% of corpus), surfaces semantic candidates that the statistical screen misses, particularly historical or pioneering papers whose embedding distinctiveness is moderate. Layer 3, the high-precision LLM filter (R3 self-reflection on R4 candidates, ~3–5%), provides a conservative shortlist suitable for final domain-expert review. Layer 4, bibliographic verification of the resulting shortlist against Crossref and Web of Science, is motivated by the empirical observation (Section 3.10) that R3 self-reflection produces accurate conceptual identifications but bibliographic metadata (author, year, journal) with a non-negligible error rate, including outright hallucinations of the year and journal of publication and occasionally of author lists (a documented failure mode that motivates explicit verification [46]). We treat this fourth layer as non-optional in any deliverable use of the workflow. The recommended operational workflow (Layer 2 R4 + Layer 3 R3 + the topic labeling overlay used in both, with R1 and R2 omitted as exploratory), applied to a 286-paper corpus, costs approximately USD 2.55 (R4: USD 1.68; R3: USD 0.76; topic labeling: USD 0.11) and takes approximately 59 min of wall-clock time on consumer hardware, making calibrated multi-method triangulation operationally accessible to research teams previously excluded by computational linguistics expertise barriers. For comparison, the full four-mode ablation (R1 + R2 + R4 + R3 + topic labeling) costs USD 5.45 and 118 min, but R1 and R2 are documented here as methodological characterizations of the prompt-level floor and are not part of the recommended operational workflow. The pick-level non-overlap between R3, R4 and the classic ensemble (R3 captures 3/9 declared outliers, R4 captures 5/9, with R3 contributing seven additional field landmarks not flagged by the classic ensemble) confirms that the three approaches are non-redundant: each surfaces different signal types in the corpus and the union of their outputs provides richer coverage than any single method. Framed this way, the contribution is not merely an analysis of one corpus but a reusable QA protocol. By pairing a high-recall LLM screen (R4) with a high-precision self-reflection filter (R3) and mandatory bibliographic verification, any domain-expert community can separate genuine innovation signals from LLM over-attribution within the same accessible cost and hardware envelope, which turns this study into transferable methodological guidance rather than a single-domain result.

4.7. Critical Limitations and Interpretive Cautions

Important methodological limitations must be explicitly acknowledged to ensure appropriate interpretation and replication by other domain-expert teams. The corpus of 286 articles represents only a subset of global CO₂ mineralization research and particularly excludes the non-English literature and potentially under-represents research contributions from the developing world. The pronounced concentration of the corpus in recent years (69.9% of documents from 2020 to 2025 vs. 15.4% from 2013 to 2016) is, first and foremost, a faithful reflection of the explosive recent growth of CO₂ mineralization research

rather than a sampling artifact to be apologized for; indeed, it is precisely this kind of rapid, recent expansion that makes accessible synthesis tools necessary for domain experts confronting information overload, and that motivates the present study. We nonetheless retain the corresponding methodological caveat: this temporal asymmetry constrains the reliability of evolutionary-pattern inference, which is why, throughout the paper, we frame cross-temporal observations as descriptive corpus proportions rather than historical trajectories, while it simultaneously provides realistic, non-ideal sampling conditions for testing methodology performance.

To address the risk that the pipeline itself has been fitted to this particular, temporally skewed corpus, we recomputed the reported findings after removing the temporal skew from the aggregation (`temporal_sensitivity.py`, deposited; Table S10). The macro-cluster shares are essentially unchanged under period-balanced resampling drawing an equal number of papers from each of the four periods: the largest shift with respect to the full corpus is 0.5 percentage points (Cluster A 44.3% vs. 44.3%, Cluster B 35.0% vs. 35.5%, Cluster C 20.7% vs. 20.2%). The near-chance agreement between pipelines at the level of the individual paper holds in both temporal halves (mean Cohen's kappa +0.041 for papers up to 2019, $n = 86$; +0.038 from 2020 onwards, $n = 200$; +0.039 over the full corpus). Outlier detection is the component most sensitive to temporal composition: re-ranking the ensemble separately within each half at the same detection rate recovers seven of the nine innovation landmarks, losing two and adding two, which is why the landmark list is best read as a ranked shortlist whose margins depend on corpus composition rather than as a fixed set. The limits of this evidence should be stated plainly: the four models were fitted on the full corpus and were not re-fitted on the subcorpora, which would require re-running the original pipelines. These tests therefore establish that the reported findings are not artifacts of aggregation over a skewed corpus, but they cannot establish that the fitted models themselves would be unchanged under balanced training. Methodology–corpus overfitting cannot be excluded from within a single corpus; replication on an independent corpus, in a different domain, is a required next step rather than an optional extension.

Implementation by a single interdisciplinary research team cannot capture the full range of domain-expert backgrounds, technical capabilities and institutional resource configurations that affect methodology adoption across diverse academic environments. The focus on CO₂ mineralization as the test domain limits the direct generalizability to other scientific fields, although methodological insights into implementation complexity, resource requirements and performance characteristics should transfer broadly to specialized scientific domains with similar technical-vocabulary density and research-evolution patterns. A further constraint concerns the language coverage of the pipeline itself, which is not uniform across components. Two of the four topic-modeling pipelines, BERTopic + SciBERT and SciBERT-NVCTM, depend on SciBERT, which is pretrained on English scientific text; transferring them to another language would require substituting a comparable domain-pretrained encoder, which is not available for most languages. The remaining components are not tied to English in the same way: MalletLDA and ProLDA operate on bag-of-words representations, the semantic network is built from term embeddings trained on the corpus itself, the outlier ensemble operates on TF–IDF features and the LLM labeling and calibration layer is multilingual; these would require retraining on a new corpus rather than replacement of a pretrained model. The scalability of the framework to the non-English literature is therefore partial: the workflow transfers, its two transformer-based pipelines do not, and any multilingual replication would have to establish its own encoder baseline before the benchmark reported here could be repeated. A parallel constraint applies to the LLM component and it should be stated without hedging: the entire calibration ablation was run on a single model and no cross-model comparison was performed. The rates

reported in Section 3.10 and Table 3 (the 71.0% naive zero-shot flag rate, the prompt-level floor near 28%, the 3.5% convergence of the self-reflection pass) are therefore properties of Claude Haiku 4.5 under these prompts, not established constants of zero-shot innovation detection. Whether the prompt-level floor in particular is a structural property of single-pass decision-making under sparse base rates, as we argue on mechanistic grounds, or an artifact of one model's calibration, cannot be settled within a single-model design: settling it would require repeating the full four-level ablation on at least one further model family, and we identify this as the most informative single extension of the present work.

The expert evaluation of the labels is subject to two limitations of its own. First, it validates the correspondence between a label and the terms it summarizes, not the status of the underlying topic: a rater who sees only the top 15 terms can verify that a label renders them faithfully, but not that the topic is a meaningful research theme, which remains a property of the model that produced it. Second, the panel is small and the judgments are concentrated: with three raters and 93.5% of the judgments in a single category the reliability coefficients are imprecisely estimated and the comparison of declared confidence between accepted and rejected labels rests on three rejected labels, so it should be read as a direction rather than as an estimate. The consensus is nonetheless independent of the author team, which is the property that matters for the use made of it here.

Unlike comprehensive bibliometric studies [2] with exhaustive corpus construction and expert-validation protocols, our pilot approach cannot definitively distinguish genuine research signal from methodological artifacts arising from corpus limitations or model sensitivities; for this reason we deliberately frame our cross-temporal observations as descriptive macro-cluster proportions rather than evolutionary trajectories. The documentation of memory-usage patterns (peak 12 GB for transformer models, observed on the system activity monitor rather than instrumented), processing bottlenecks and configuration complexity hierarchies offers practical benchmarks for institutional resource planning and methodology-selection frameworks. Custom corpus-derived vocabulary construction emerged as a critical success factor, substantially outperforming generic implementations in coherence and domain specificity across all models.

A final caution concerns scale, the dimension on which this study is least able to speak. The framework was validated on 286 documents and the 12 GB peak just noted leaves limited headroom on the 16 GB consumer machine that defines its resource budget: we have not established the corpus size at which this framework ceases to be runnable under the very constraints that motivate it, nor should the reported ranking be assumed to hold at larger scale: Cao et al. [10], applying BERTopic against an LDA baseline to 22,623 reports with the same two metrics adopted here, obtain the opposite ordering (BERTopic C_v 0.50–0.70 against LDA 0.41–0.46, where we obtain 0.381 against 0.542), a divergence plausibly attributable to corpus size and to the choice of encoder, and one we report precisely because it runs against our own result. What this study establishes is therefore a protocol (a multi-pipeline benchmark under a declared resource budget, a leveled calibration of the LLM component and a triangulated verification step) and not a set of transferable values. The pipeline ranking, the flag rates and the nine-document shortlist are properties of this corpus, this encoder, this model and this scale.

4.8. Methodological Contributions and Future Research Directions

Our study demonstrates that computational literature-analysis tools provide valuable support for research planning and strategic decision-making by domain-expert teams, while emphasizing that these tools supplement rather than replace systematic literature review and expert judgment. The accessibility of sophisticated topic modeling creates opportunities for a more distributed synthesis of scientific knowledge, potentially enabling

broader participation in identifying the research directions critical for addressing complex global challenges that require interdisciplinary collaboration.

The network analysis, revealing 82 conceptual communities with bridge terms achieving high betweenness centrality (e.g., ‘circular economic’: 0.0126), provides systematic approaches for identifying interdisciplinary innovation opportunities. Bridge terms representing only 5% of the total vocabulary account for 43% of cross-community connections, confirming their critical mediating role; the corresponding 8.7-fold enrichment of bridge-term mediation over the uniform-allocation baseline is consistent with the structural holes framework [18]. At the document level, the calibrated four-detector ensemble retained nine documents (3.13%) as outliers under the double threshold on the weighted ensemble score described in Methods, and this selection is stable across a wide range of threshold settings (Table S13). The framework therefore provides practical pipelines for discovering candidate emerging research trajectories overlooked by conventional thematic clustering, while making explicit that the deliverable outlier set ($n = 9$) is a ranked shortlist whose length, not its membership, responds to the choice of threshold.

Future implementations could benefit from collaborative corpus-construction protocols across multiple institutions to address temporal and geographical bias; from the development of domain-specific parameter optimization guidelines accessible to non-specialists; and from the creation of hybrid validation frameworks integrating computational metrics with systematic expert-assessment protocols. The development of standardized resource-requirement documentation and implementation complexity assessment tools could facilitate a broader adoption across diverse academic environments.

The thematic macro-cluster taxonomy identified across all models comprises Foundational mechanisms (Cluster A), Process engineering and industrial-residue valorization (Cluster B) and Application-oriented and integrated systems (Cluster C). It provides a model-independent organizing framework for the corpus that can be applied as a descriptive scheme to other emerging scientific domains; we explicitly avoid framing the cluster proportions as a historical evolution, because the corpus’s temporal asymmetry (69.9% post-2020) prevents reliable causal–temporal inference. Independent expert validation of the mapping (Fleiss’ kappa = 0.873, three raters external to the authorship) and a permutation test of its lexical coherence ($p = 0.0008$) support the taxonomy as a descriptive scheme, while the near-chance per-document agreement between pipelines (mean kappa = +0.037) marks its limit: the scheme is reproducible in aggregate, but not at the level of the individual paper.

4.9. Broader Implications

The demonstrated feasibility of implementing advanced bibliometric analysis with limited computational resources represents an important step towards more inclusive management of scientific knowledge, potentially reducing the analytical advantage gap between well-resourced and resource-constrained research institutions. The validation that meaningful insights can be extracted despite corpus limitations and implementation constraints suggests that imperfect but systematic analytical approaches provide substantial value over purely intuitive literature assessment methods. The integration of AI-assisted coding tools with domain expertise demonstrates a practical pathway for extending analytical capabilities without requiring extensive computational linguistics training and this hybrid approach (combining domain knowledge, accessible tools and AI assistance) may represent a sustainable model for democratization of sophisticated analytical capabilities across diverse academic contexts.

5. Conclusions

This study asked whether zero-shot large language models can substitute calibrated statistical pipelines in the detection of research innovation. They cannot, and the reason is structural rather than incidental: the two families fail in different directions, one being blind to conceptually novel work that is distributionally typical, the other having no external base rate against which to restrain itself, so that refining one does not repair the other. The calibration ablation locates the boundary precisely: prompt engineering alone does not cross the flag-rate floor it establishes and only a change of regime does. Triangulation is therefore not a methodological precaution but the only defensible use of either family.

What we offer for reuse is the protocol rather than its outputs. The pipeline ranking, the flag rates and the shortlist of landmark papers are properties of this corpus, this encoder, this model and this scale and Section 4.7 set out how far each may be expected to travel. The transferable object is the procedure: a multi-pipeline benchmark under a declared resource budget, a leveled calibration of the language-model component and a verification step that treats bibliographic checking as part of the method rather than an afterthought. That a team without NLP training could execute it end to end on consumer hardware and document what it cost, is the claim this pilot supports.

Supplementary Materials: The following supporting information can be downloaded at: <https://www.mdpi.com/article/10.3390/analytics5040045/s1>, Table S1: Definitions of the ten evaluation metrics used in the comparative analysis of the four topic-modeling pipelines; Table S2: Pairwise model comparisons for the six evaluation metrics subjected to inferential testing, computed on six independent runs per model with the SciBERT-NVCTM (K = 20) configuration; Table S3: ANOVA breakdown for each of the six evaluation metrics subjected to inferential testing across the four topic-modeling pipelines, computed on six independent runs per model with the SciBERT-NVCTM (K = 20) configuration; Table S4: MalletLDA: top-3 topics per year (2013–2025) ranked by relative weight; Table S5: ProdLDA: top-3 topics per year (2013–2025) ranked by relative weight; Table S6: BERTopic+SciBERT—top-3 topics per year (raw key terms + LLM-generated labels). Table S7: SciBERT-NVCTM: top-3 topics per year (2013–2025) ranked by relative weight; Table S8: ProdLDA (PATCH-Q rerun, seed = 42, K = 22)—topic → papers; Table S9.1: Innovation-landmark papers from method triangulation, with detection route, ensemble outlier score where applicable, novelty type, and a brief conceptual-novelty descriptor; Table S9.2: Bridge concepts in the semantic co-occurrence network; Table S10: Sensitivity of the reported findings to the temporal skew of the corpus (temporal_sensitivity.py, deposited; split at 2019, 2000 bootstrap resamples, seed 42); Table S11: Expert evaluation of the 82 LLM-generated topic labels and calibration of the model’s self-assessment; Table S12: Validation of the three thematic macro-clusters (cluster_validation.py, deposited; seed 42, 5000 permutations); Table S13: Sensitivity of the innovation-landmark set to the outlier thresholds (outlier_sensitivity.py, deposited).

Author Contributions: Conceptualization, G.S.; Methodology, G.S.; Software, G.S.; Validation, G.S.; Resources, G.S.; Data curation, G.S., G.D.R. and A.T.; Writing—original draft, G.S. and G.D.R.; Writing—review and editing, G.S., G.D.R., A.T. and M.D.B.; Visualization, G.S.; Supervision, G.S. and M.D.B. All authors have read and agreed to the published version of the manuscript.

Funding: This research received no external funding.

Data Availability Statement: The data presented in this study are openly available in Zenodo at <https://doi.org/10.5281/zenodo.21128160>. The repository includes the analysis code, verbatim LLM prompts (R1–R4 and topic labeling), and raw Claude Haiku model outputs; the 286-paper corpus is identified by a DOI/identifier list provided in the same repository.

Acknowledgments: The authors thank the ECCSEL NatLab-Italy laboratory of Panarea, owned by OGS and belonging to the international ECCSEL infrastructure (European Carbon Dioxide Capture and Storage Laboratory Infrastructure), which provided resources for this research. During this study, generative AI tools were used in two distinct roles: Claude Haiku 4.5 (Anthropic) as a

methodological component and object of study (topic labeling and the innovation-detection ablation described in Section 2.7), and AI-assisted coding tools (GitHub Copilot; OpenAI ChatGPT) to support pipeline implementation and, to a limited extent, for language editing and manuscript polishing. No generative AI tool was used to generate the scientific content or conclusions of this manuscript, and all AI-assisted outputs were verified by the authors, who take full responsibility for the content.

Conflicts of Interest: The authors declare no conflicts of interest.

Abbreviations

LDA, latent Dirichlet allocation; NVCTM, Neural Variational Correlated Topic Model; ProdLDA, Product-of-Experts LDA; BERTopic, BERT-based Topic modeling; SciBERT, Scientific BERT; c-TF-IDF, class-based term frequency-inverse document frequency; PMI, pointwise mutual information; C_v , coherence (sliding-window PMI estimator); LOF, Local Outlier Factor; TTR, type-token ratio; QA, quality assurance; LLM, Large Language Model; NLP, Natural Language Processing; GROBID, GeneRation Of Bibliographic Data; RAG, retrieval-augmented generation; AI, artificial intelligence; ANOVA, analysis of variance; TF-IDF, term frequency-inverse document frequency; UMAP, Uniform Manifold Approximation and Projection; HDBSCAN, Hierarchical Density-Based Spatial Clustering of Applications with Noise; CV, coefficient of variation.

References

- Chen, H.; Wang, X.; Pan, S.; Xiong, F. Identify Topic Relations in Scientific Literature Using Topic Modeling. *IEEE Trans. Eng. Manag.* **2021**, *68*, 1232–1244. [\[CrossRef\]](#)
- Cobo, M.J.; López-Herrera, A.G.; Herrera-Viedma, E.; Herrera, F. Science Mapping Software Tools: Review, Analysis, and Cooperative Study Among Tools. *J. Am. Soc. Inf. Sci. Technol.* **2011**, *62*, 1382–1402. [\[CrossRef\]](#)
- Devlin, J.; Chang, M.-W.; Lee, K.; Toutanova, K. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. In Proceedings of the NAACL-HLT 2019, Minneapolis, MN, USA, 2–7 June 2019; pp. 4171–4186. [\[CrossRef\]](#)
- Beltagy, I.; Lo, K.; Cohan, A. SciBERT: A Pretrained Language Model for Scientific Text. In Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP), Hong Kong, China, 3–7 November 2019; pp. 3615–3620. [\[CrossRef\]](#)
- Lopez, P. GROBID: A Machine Learning Software for Extracting, Parsing and Re-Structuring Scientific Publications GitHub Repository. 2008. Available online: <https://github.com/kermitt2/grobid> (accessed on 22 June 2026).
- Blei, D.M.; Ng, A.Y.; Jordan, M.I. Latent Dirichlet Allocation. *J. Mach. Learn. Res.* **2003**, *3*, 993–1022.
- Srivastava, A.; Sutton, C. Autoencoding Variational Inference for Topic Models. In Proceedings of the International Conference on Learning Representations (ICLR), Toulon, France, 24–26 April 2017.
- Grootendorst, M. BERTopic: Neural Topic Modeling with a Class-Based TF-IDF Procedure. *arXiv* **2022**, arXiv:2203.05794. [\[CrossRef\]](#)
- Dieng, A.B.; Ruiz, F.J.R.; Blei, D.M. Topic Modeling in Embedding Spaces. *Trans. Assoc. Comput. Linguist.* **2020**, *8*, 439–453. [\[CrossRef\]](#)
- Cao, Y.; Qu, Z.; Wu, S.; Chen, Y.; Skitmore, M.; Ma, X.; Wang, J. Analyzing OSHA Construction Accident Reports Using BERTopic Topic Modeling for Thematic Insights. *Buildings* **2026**, *16*, 10. [\[CrossRef\]](#)
- Pham, C.; Hoyle, A.; Sun, S.; Resnik, P.; Iyyer, M. TopicGPT: A Prompt-Based Topic Modeling Framework. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*; Association for Computational Linguistics: Stroudsburg, PA, USA, 2024; pp. 2956–2984. [\[CrossRef\]](#)
- Tian, K.; Mitchell, E.; Zhou, A.; Sharma, A.; Rafailov, R.; Yao, H.; Finn, C.; Manning, C.D. Just Ask for Calibration: Strategies for Eliciting Calibrated Confidence Scores from Language Models Fine-Tuned with Human Feedback. In Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing, Singapore, 6–10 December 2023; pp. 5433–5442. [\[CrossRef\]](#)
- Madaan, A.; Tandon, N.; Gupta, P.; Hallinan, S.; Gao, L.; Wiegrefe, S.; Alon, U.; Dziri, N.; Prabhunoye, S.; Yang, Y.; et al. Self-Refine: Iterative Refinement with Self-Feedback. *Adv. Neural Inf. Process. Syst.* **2023**, *36*, 46534–46594. [\[CrossRef\]](#)
- Lopez, P. GROBID: Combining Automatic Bibliographic Data Recognition and Term Extraction for Scholarship Publications. In *Research and Advanced Technology for Digital Libraries (ECDL 2009)*; Lecture Notes in Computer Science; Springer: Berlin/Heidelberg, Germany, 2009; pp. 473–474. [\[CrossRef\]](#)
- Anthropic. *Claude Haiku 4.5 Model Card*; Anthropic: San Francisco, CA, USA, 2025. Available online: <https://docs.claude.com> (accessed on 22 June 2026).

16. Brandes, U. A Faster Algorithm for Betweenness Centrality. *J. Math. Sociol.* **2001**, *25*, 163–177. [\[CrossRef\]](#)
17. Freeman, L.C. A Set of Measures of Centrality Based on Betweenness. *Sociometry* **1977**, *40*, 35–41. [\[CrossRef\]](#)
18. Burt, R.S. Structural Holes and Good Ideas. *Am. J. Sociol.* **2004**, *110*, 349–399. [\[CrossRef\]](#)
19. Liu, F.T.; Ting, K.M.; Zhou, Z.-H. Isolation Forest. In Proceedings of the 2008 Eighth IEEE International Conference on Data Mining (ICDM), Pisa, Italy, 15–19 December 2008; pp. 413–422. [\[CrossRef\]](#)
20. Breunig, M.M.; Kriegel, H.-P.; Ng, R.T.; Sander, J. LOF: Identifying Density-Based Local Outliers. In Proceedings of the 2000 ACM SIGMOD International Conference on Management of Data, Dallas, TX, USA, 16–18 May 2000; pp. 93–104. [\[CrossRef\]](#)
21. Aggarwal, C.C.; Sathe, S. *Outlier Ensembles: An Introduction*; Springer International Publishing: Cham, Switzerland, 2017. [\[CrossRef\]](#)
22. Fleiss, J.L. Measuring Nominal Scale Agreement among Many Raters. *Psychol. Bull.* **1971**, *76*, 378–382. [\[CrossRef\]](#)
23. Krippendorff, K. *Content Analysis: An Introduction to Its Methodology*, 2nd ed.; Sage: Thousand Oaks, CA, USA, 2004.
24. Cohen, J. A Coefficient of Agreement for Nominal Scales. *Educ. Psychol. Meas.* **1960**, *20*, 37–46. [\[CrossRef\]](#)
25. Byrt, T.; Bishop, J.; Carlin, J.B. Bias, Prevalence and Kappa. *J. Clin. Epidemiol.* **1993**, *46*, 423–429. [\[CrossRef\]](#)
26. Chang, J.; Boyd-Graber, J.; Gerrish, S.; Wang, C.; Blei, D.M. Reading Tea Leaves: How Humans Interpret Topic Models. *Adv. Neural Inf. Process. Syst.* **2009**, *22*, 288–296.
27. Xie, H.; Jiang, W.; Wang, Y.; Liu, T.; Wang, R.; Liang, B.; He, Y.; Wang, J.; Tang, L.; Chen, J. Thermodynamics Study on the Generation of Electricity via CO₂-Mineralization Cell. *Environ. Earth Sci.* **2015**, *74*, 6481–6488. [\[CrossRef\]](#)
28. Luong, N.T.; Veyret, N.; Boily, J.-F. CO₂ Mineralization by MgO Nanocubes in Nanometric Water Films. *ACS Appl. Mater. Interfaces* **2023**, *15*, 45055–45063. [\[CrossRef\]](#)
29. Coppola, A.; Scala, F.; Azadi, M. Direct Dry Carbonation of Mining and Industrial Wastes in a Fluidized Bed for Offsetting Carbon Emissions. *Processes* **2022**, *10*, 582. [\[CrossRef\]](#)
30. Strunge, T.; Renforth, P.; Van der Spek, M. Uncertainty Quantification in the Techno-Economic Analysis of Emission Reduction Technologies: A Tutorial Case Study on CO₂ Mineralization. *Front. Energy Res.* **2023**, *11*, 1182969. [\[CrossRef\]](#)
31. Callagon La Plante, E.; Simonetti, D.A.; Wang, J.; Al-Turki, A.; Chen, X.; Jassby, D.; Sant, G.N. Saline Water-Based Mineralization Pathway for Gigatonne-Scale CO₂ Management. *ACS Sustain. Chem. Eng.* **2021**, *9*, 1073–1089. [\[CrossRef\]](#)
32. Moon, D.H.; Park, S.S.; Kang, S.-P.; Lee, W.; Park, K.T.; Chun, D.H.; Rhim, G.B.; Hwang, S.-M.; Youn, M.H.; Jeong, S.K. Determination of Kinetic Factors of CO₂ Mineralization Reaction for Reducing CO₂ Emissions in Cement Industry and Verification Using CFD Modeling. *Chem. Eng. J.* **2021**, *420*, 129420. [\[CrossRef\]](#)
33. Matter, J.M.; Stute, M.; Snæbjörnsdóttir, S.Ó.; Oelkers, E.H.; Gislason, S.R.; Aradóttir, E.S.; Sigfusson, B.; Gunnarsson, I.; Sigurdardóttir, H.; Gunnlaugsson, E.; et al. Rapid Carbon Mineralization for Permanent Disposal of Anthropogenic Carbon Dioxide Emissions. *Science* **2016**, *352*, 1312–1314. [\[CrossRef\]](#)
34. Assen, A.H.; Belmabkhout, Y.; Adil, K.; Lachhab, A.; Hassoune, H.; Aggarwal, H. Advances on CO₂ Storage. Synthetic Porous Solids, Mineralization and Alternative Solutions. *Chem. Eng. J.* **2021**, *419*, 129569. [\[CrossRef\]](#)
35. Gong, J.; Zhang, T.; Wang, J.; Jin, Y.; Li, J.; Wang, Y. Simultaneous Rare Earth Sulfate Transformation and Carbon Dioxide Mineralization. *Sep. Purif. Technol.* **2025**, *356*, 129805. [\[CrossRef\]](#)
36. Alqahtani, A.; Addassi, M.; Hoteit, H.; Oelkers, E.H. Rapid CO₂ Mineralization by Zeolite via Cation Exchange. *Sci. Rep.* **2025**, *15*, 958. [\[CrossRef\]](#)
37. Myers, C.A.; Nakagaki, T.; Akutsu, K. Quantification of the CO₂ Mineralization Potential of Ironmaking and Steelmaking Slags under Direct Gas-Solid Reactions in Flue Gas. *Int. J. Greenh. Gas Control* **2019**, *87*, 100–111. [\[CrossRef\]](#)
38. Myers, C.; Nakagaki, T.; Miho, Y. Demonstration of Brine-Based Atmospheric CO₂ Removal. In Proceedings of the 2nd International Conference on Negative CO₂ Emissions, Gothenburg, Sweden, 14–17 June 2022.
39. Gislason, S.R.; Broecker, W.S.; Gunnlaugsson, E.; Snæbjörnsdóttir, S.Ó.; Mesfin, K.G.; Alfredsson, H.A.; Aradóttir, E.S.; Sigfusson, B.; Gunnarsson, I.; Stute, M.; et al. Rapid Solubility and Mineral Storage of CO₂ in Basalt. *Energy Procedia* **2014**, *63*, 4561–4574. [\[CrossRef\]](#)
40. Sharifian, R.; Boer, L.; Wagterveld, R.M.; Vermaas, D.A. Oceanic Carbon Capture Through Electrochemically Induced In Situ Carbonate Mineralization Using Bipolar Membrane. *Chem. Eng. J.* **2022**, *438*, 135326. [\[CrossRef\]](#)
41. Sigfusson, B.; Gislason, S.R.; Matter, J.M.; Stute, M.; Gunnlaugsson, E.; Gunnarsson, I.; Aradóttir, E.S.; Sigurdardóttir, H.; Mesfin, K.; Alfredsson, H.A.; et al. Solving the Carbon-Dioxide Buoyancy Challenge: The Design and Field Testing of a Dissolved CO₂ Injection System. *Int. J. Greenh. Gas Control* **2015**, *37*, 213–219. [\[CrossRef\]](#)
42. Pogge von Strandmann, P.A.E.; Burton, K.W.; Snæbjörnsdóttir, S.Ó.; Sigfusson, B.; Aradóttir, E.S.; Gunnarsson, I.; Alfredsson, H.A.; Mesfin, K.G.; Oelkers, E.H.; Gislason, S.R. Rapid CO₂ Mineralisation into Calcite at the CarbFix Storage Site Quantified Using Calcium Isotopes. *Nat. Commun.* **2019**, *10*, 1983. [\[CrossRef\]](#)
43. Feinstein, A.R.; Cicchetti, D.V. High Agreement but Low Kappa: I. The Problems of Two Paradoxes. *J. Clin. Epidemiol.* **1990**, *43*, 543–549. [\[CrossRef\]](#)

44. Sia, S.; Duh, K. Adaptive Mixed Component LDA for Low-Resource Topic Modeling. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics (EACL): Main Volume*; Association for Computational Linguistics: Stroudsburg, PA, USA, 2021; pp. 2451–2469. [[CrossRef](#)]
45. Stambach, D.; Zouhar, V.; Hoyle, A.; Sachan, M.; Ash, E. Revisiting Automated Topic Model Evaluation with Large Language Models. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, Singapore, 6–10 December 2023; pp. 9348–9357. [[CrossRef](#)]
46. Dhuliawala, S.; Komeili, M.; Xu, J.; Raileanu, R.; Li, X.; Celikyilmaz, A.; Weston, J. Chain-of-Verification Reduces Hallucination in Large Language Models. *arXiv* **2023**, arXiv:2309.11495. [[CrossRef](#)]

Disclaimer/Publisher’s Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.