Received: January 7, 2026. Revised: February 8, 2026.

INASS

International Journal of
Intelligent Engineering & Systems

http://www.inass.org/

Automated Knowledge Synthesis: An LLM-refined Framework for Evolutionary Topic Modeling

Amer Abu-Jassar[1]* Mohammad Hamdan[2] Nowfal Aweisi[3] Roman Slisareko[4] Zhanna Deineko[4] Vyacheslav Lyashenko[4]

[1]*Department of Computer Science, College of Information Technology, Amman Arab University, Amman, Jordan*
[2]*College of Cyber Security, Amman Arab University, Amman, Jordan*
[3]*School of Electrical Engineering and Information Technology, German Jordanian University, Amman, Jordan*
[4]*Department of Media Systems and Technology, Kharkiv National University of Radio Electronics, Kharkiv, Ukraine*
* Corresponding author's Email: A.abujassar@aau.edu.jo

Abstract: Traditional topic modeling methods, such as Latent Dirichlet Allocation (LDA) and Non-negative Matrix Factorization (NMF), are limited by their context ignorance, static nature, and low interpretability. Building upon the hybrid approach LDA+NMF+class-based Term Frequency–Inverse Document Frequency (c-TF-IDF), a new formalized framework – Dynamic Contextual Topic Modeling with Large Language Model (LLM) Refinement (DCTM-LLM) – is presented. This LLM-refined framework integrates transformer embeddings for the detection of dynamic semantic clusters and leverages an LLM for their subsequent refinement and the synthesis of high-level narratives. Experiments on a corpus of 35,000 arXiv abstracts (cs.AI (Computer Science — Artificial Intelligence), 2015–2025) showed that DCTM-LLM achieves a Normalized Pointwise Mutual Information (NPMI) of 0.53, a Silhouette score of 0.62, an Adjusted Rand Index (ARI) of 0.55, and Topic Diversity at 10 of 0.88. Crucially, with a Bidirectional Encoder Representations from Transformers (BERT)-based score (BERTScore) F1 of 0.89, the method significantly outperforms Dynamic BERTopic (0.62) and the hybrid LDA, NMF, and c-TF-IDF approach (0.65). Thus, the proposed approach shifts the paradigm of topic modeling from keyword extraction toward automated knowledge synthesis.

Keywords: Topic modeling, Large language model, Contextual embeddings, Topic model interpretability, Narrative synthesis.

1. Introduction

The proliferation of large-scale, unstructured textual arrays in the digital space poses a fundamental epistemological challenge to computational sciences: how to transition from raw data to synthesized, machine-readable knowledge. The task lies in developing computational methodologies capable of automatically identifying latent semantic structures, modeling their dynamics, and presenting the results in an interpretable form. The necessity for such tools is critical, as traditional qualitative approaches do not scale to the analysis of modern information volumes [1-4]. Thus, the key motivation of this research is to formalize the knowledge synthesis process by creating an algorithmic framework that ensures not only the identification of topics but also the quantitative analysis of their conceptual evolution. Consequently, the task is defined as the reconstruction of time-aligned, interpretable thematic structures and concise narratives that accurately summarize their evolution over a long-term scientific corpus.

Topic modeling, as a subfield of unsupervised machine learning, offers a formal apparatus for solving this problem. The fundamental approaches that have shaped the discipline include LDA and NMF. However, their theoretical elegance is

This article is licensed under a Creative Commons Attribution-ShareAlike 4.0 International License.
License details: https://creativecommons.org/licenses/by-sa/4.0/

constrained by a common inherent drawback: they are based on the "bag-of-words" assumption, ignoring semantic context, which is a well-documented limitation [5]. Therefore, the desire to improve relevant research is not accidental, particularly by enhancing LDA through the integration of informed prior distributions derived from NMF and c-TF-IDF, and the application of regularization to stabilize topic vocabularies [6]. While this approach increases the coherence and stability of topics, it cannot move beyond the paradigmatic constraints inherited from its static and non-contextual base components.

A key limitation of the foundational approaches is the static nature of the LDA model, which does not allow for flexible adaptation to time-varying data and requires the predefined specification of the number of topics [7]. This static nature restricts the model's ability to adapt, as topics are treated as immutable rather than evolving over time [8, 9]. Similar issues, particularly static representation, are inherent to NMF. This method also exhibits sensitivity to sparsity and the scaling of the feature matrix, which can lead to less accurate results when analyzing dynamic or high-dimensional datasets [10, 11].

Embedding-oriented methods, such as Top2Vec and BERTopic (a BERT-based topic modeling pipeline), are characterized by enhanced lexical coherence; however, this is often achieved at the cost of diminished inter-temporal consistency and stability in their textual designations [12]. These limitations underscore the need for more dynamic and adaptive topic modeling approaches capable of incorporating temporal changes and contextual nuances, especially when dealing with large volumes of textual data. Dynamic topic models (DTM), while oriented toward tracking drift, often lack mechanisms for the automated formation of narrative summaries. This indicates that, despite advancements, dynamic models still exhibit gaps in the integration of contextual understanding and the automation of time-series thematic structure interpretation [13]. Novel LLM-oriented approaches to labeling are also not without drawbacks, including the absence of guarantees for temporal topic consistency and the risk of generative errors (hallucinations). LLMs may generate repetitive and inconsistent topics over time, and they can produce topic-irrelevant results, which consequently reduces the accuracy of the analysis [14].

Accordingly, further progress in this field requires a transition to a new paradigm evolving along three critical directions:

Contextual understanding by shifting to semantic vector representations, which has proven effective in modern models [15];

Dynamic (temporal) analysis in the context of modeling the evolution of topics over time [16];

Human-centric interpretability using LLMs to generate meaningful narratives [17].

These three directions directly operationalize the defined task by ensuring coherent clustering representation within each time slice, robust topic alignment over time, and the generation of reliable, concise narratives for the end user.

Literature analysis indicates a convergence of research efforts toward synthesizing these paradigms. On the one hand, the integration of contextual embeddings with temporal analysis has led to significant improvements in tracking topic evolution [16]. On the other hand, a separate line of research actively uses LLMs to enhance interpretability, implementing this either through a posteriori naming [17] or by direct involvement in topic generation [18]. The most advanced hybrid approaches attempt to combine LLMs with dynamic models to detect narrative shifts or apply LLMs to optimize intermediate stages, such as iterative topic expansion [19, 20]. However, while these approaches solve individual subtasks, they do not offer a holistic solution. Existing approaches either ensure temporal consistency without narrative synthesis, or form textual topic designations without strictly maintaining temporal identity. Consequently, a methodological gap remains, requiring a unified, formally defined pipeline.

Therefore, attention should be drawn to the DCTM-LLM approach – a new formalized framework that integrates these three directions. This approach should be considered a two-stage process: the first stage identifies dynamic semantic clusters, and the second uses an LLM for their automatic interpretation. This approach transforms topic modeling from a data mining tool into a tool for automated analytical report generation. Unlike LDA/NMF, which are static Bag-of-Words models, and BERTopic-like pipelines, which exhibit keyword fragmentation across time slices, DCTM-LLM first forms temporally consistent contextual clusters. Subsequently, the model engages an LLM to synthesize reliable topic labels and concise narratives, thereby enhancing both intrinsic quality (coherence, stability) and the interpretability of the results.

To empirically substantiate these limitations, Section V presents a comparison on the same corpus of 35,000 documents using intrinsic topic-quality metrics (NPMI, Silhouette, ARI, Topic

This article is licensed under a Creative Commons Attribution-ShareAlike 4.0 International License.
License details: https://creativecommons.org/licenses/by-sa/4.0/

Diversity@10) and interpretation quality metrics (BERTScore F1), complemented by an ablation study (Tables 2 and 3) and a qualitative case study (Section 5.2).

The remainder of this paper is organized as follows. Section II reviews the theoretical background and related work in topic modeling. Section III details the formal definition and architecture of the proposed DCTM-LLM framework. Section IV describes the experimental setup, dataset, and evaluation metrics. Section V presents and discusses the quantitative and qualitative results of our experiments. Finally, Section VI concludes the paper and outlines future research directions.

2. Theoretical background and related work

2.1 Probabilistic topic models

Probabilistic topic models view documents as the result of a generative process. The central model in this paradigm is LDA, first proposed in [21]. Conceptually, LDA is a probabilistic generative model that posits that every document is a multinomial distribution (a mixture) over a set of topics, where each topic, in turn, is defined by a probability distribution over the vocabulary.

Formally, according to [21], LDA is based on a hierarchical generative process that stipulates the mechanism for creating each document w in the corpus. For each document, its unique topic distribution – the vector of topic proportions θ – is first generated by sampling from a Dirichlet distribution, parameterized by the corpus-wide hyperparameter α. Subsequently, for each of the N in the document, a two-step procedure is performed: first, a latent topic is identified z_n is chosen from the $Multinomial$ distribution z, whose parameters are the vector θ, after which the observed word w_n is generated from the conditional distribution $p(w_n \mid z_n, \beta)$, which is specific to the chosen topic.

This process is formalized by the equation for the joint probability distribution of all model variables [21]:

$$p(\theta, z, w \mid \alpha, \beta) = \\ = p(\theta \mid \alpha) \prod_{n=1}^{N} p(z_n \mid \theta)\ p(w_n \mid z_n, \beta), \quad (1)$$

where: $p(\theta \mid \alpha)$ is the Dirichlet prior distribution for the document's topic proportions;

$\prod_{n=1}^{N} p(z_n \mid \theta)\ p(w_n \mid z_n, \beta)$ represents the conditional probability of the sequence of words and their topic assignments given θ.

2.2 Matrix factorization models

An alternative approach to thematic modeling is based on linear algebra methods, particularly NMF [22]. The concept of NMF is to factorize the input matrix V into two matrices of smaller dimension, W and H, with the key condition of the non-negativity of all elements.

Formally, the task is to approximate the document-term matrix $V \in R_{+}^{M \times V_{size}}$, where M is the number of documents and V_{size} is the vocabulary size, by the product of two matrices: $V \approx WH$. The matrices $W \in R_{+}^{M \times K}$ and $H \in R_{+}^{K \times V_{size}}$ represent, respectively, the topic distributions in documents and the word distributions in topics, and K is the predefined number of latent topics.

The optimization problem is most often formulated as the minimization of the squared error, measured by the Frobenius norm (denoted by the index F) [22]:

$$\min_{W,H \geq 0} \ ||V - WH||_F^2. \quad (2)$$

This objective function measures the sum of the squares of the element-wise differences between the original matrix V and its approximation WH, which is equivalent to minimizing the Euclidean distance. To solve this problem, multiplicative update rules are widely used, which iteratively adjust the elements of matrices W and H while preserving their non-negativity. Thanks to this constraint, NMF generates sparse and additive representations, which contribute to the high interpretability of the topics.

2.3 Contextual embeddings

The limitations of classical models that ignore word order are overcome by using contextual vector representations (embeddings). Unlike static models (e.g., Word2Vec), Transformer-based architectures, such as BERT [23], generate context-dependent vectors. This is achieved through the self-attention mechanism, which weights the importance of different words in a sentence for the representation of each specific word. The key element is the Scaled Dot-Product Attention operation [24]:

$$Attention(Q, K, V) = softmax\left(\frac{QK^T}{\sqrt{d_k}}\right)V, \quad (3)$$

where: Q (Query), K (Key), and V (Value) are matrix projections of the input vectors;

K^T denotes the matrix transpose operation of K, which is mathematically necessary for computing scalar products between query and key vectors;

This article is licensed under a Creative Commons Attribution-ShareAlike 4.0 International License.
License details: https://creativecommons.org/licenses/by-sa/4.0/

d_k is the dimensionality of the key vectors. This approach allows the model to capture complex semantic relationships, which is critical for high-quality topic analysis.

2.4 Dynamic and temporal topic models

The static nature of classical models is a significant limitation when analyzing evolving data. DTM [25] address this problem by modeling topics as sequences that change over time. In classical DTM, the topics at the time step t depend on the topics at the step $t-1$, which is often modeled using a state-space model, for example, through a Gaussian random walk process for the parameters of a logistic normal transformation [25]:

$$\beta_t \mid \beta_{t-1} \sim N(\beta_{t-1}, \sigma^2 I), \quad (4)$$

where: β_t represents the natural parameters of the topic distribution at time t;

$N(\mu, \Sigma)$ denotes the multivariate normal (Gaussian) distribution with mean vector μ and covariance matrix Σ. In this case, the mean value β_{t-1} Indicates that the topic at the time t is expected to be centered around its state at the time $t-1$;

$\sigma^2 I$ it is the covariance matrix of the evolution process. It consists of two components: the scalar variance parameter σ^2, which controls the magnitude of topic evolution (a smaller value limits "topic drift", while a larger one allows for significant changes), and the identity matrix I. The use of the identity matrix introduces the simplifying assumption that the evolution of the probability of each word within a topic is independent of the evolution of other words.

Modern neural approaches extend this idea by using more flexible architectures to model dynamics [16].

2.5 LLMs for topic model enhancement

The advent of LLMs has opened new possibilities for improving the quality and, especially, the interpretability of topic models. Instead of relying exclusively on statistical patterns, modern methods use LLMs as an external knowledge source. LLMs can be applied at various stages: for generating high-quality names for topics, for filtering and merging semantically similar topics, or even for direct topic detection from text [17]. Hybrid approaches, such as TopicGPT [18], demonstrate significant potential in creating models whose results are not only coherent but also easily understandable to the end-user.

3. The proposed DCTM-LLM framework

To overcome the inherent limitations of classical topic models, associated with static nature and context ignorance, a new two-stage framework is proposed: DCTM-LLM. This framework sequentially integrates contextual embeddings for semantic grouping, temporal analysis for modeling topic evolution, and LLMs for generating highly interpretable narratives.

In Stage 1, dimensionality reduction is performed using Uniform Manifold Approximation and Projection (UMAP), followed by clustering with Hierarchical Density-Based Spatial Clustering of Applications with Noise (HDBSCAN), which enables stable global cluster identities for subsequent temporal analysis.

Algorithm 1: DCTM-LLM (end-to-end pipeline)

Input: Corpus D with timestamps t_i, encoder f_{enc}, time slice formation rule, module parameters.

Output: Dynamic narrative topics $\{(Label_k, Narrative_k, \{(w_{k,\tau}, F_{k,\tau})\}_\tau)\}_k$.

1. Encode each document: $d_i \to v_i = f_{\text{enc}}(d_i)$.
2. Reduce dimensionality: $u_i = \text{UMAP}(v_i; \theta_{\text{umap}})$.
3. Perform global clustering: $z_i = \text{HDBSCAN}(u_i; \theta_{\text{hdbscan}})$. Objects marked as noise are processed separately.
4. For each time slice τ and cluster k: collect documents for which $z_i = k$ and $t_i = \tau$.
5. Calculate frequency in slice $F_{k,\tau}$, where $F_{k,\tau} = |D_{k,\tau}|$.
6. Compute lexical representation $w_{k,\tau}$ using c-TF-IDF with parameters θ_{ctfidf}.
7. Form trajectories by linking the same global cluster k across time slices τ according to the minimum support rule. In this work, a threshold of $F_{k,\tau} \geq 15$ is used.
8. In the second stage, generate and validate merge proposals using an LLM and similarity filters. Subsequently, apply the confirmed merges.
9. For each final trajectory, construct input data for the LLM based on $\{w_{k,\tau}\}$ and representative fragments. Afterward, obtain $(Label_k, Narrative_k)$.
10. Return the final set of dynamic narrative topics.

3.1 Formal problem definition

Formally, the problem solved by the proposed framework is defined as a functional transformation of the input corpus of documents into a structured set of dynamic narrative topics. The input data is the corpus $D = \{d_1, d_2, \ldots, d_M\}$, where each document d_i (for $i = 1, \ldots, M$) is associated with a timestamp

This article is licensed under a Creative Commons Attribution-ShareAlike 4.0 International License.
License details: https://creativecommons.org/licenses/by-sa/4.0/

$t_i \in \mathbb{T}$. The set $\mathbb{T}$ represents the collection of all unique timestamps in the corpus.

The output is a set $\mathcal{T} = \{T_1, T_2, \ldots, T_K\}$, representing dynamic narrative topics. Each topic T_k is formalized as a tuple $(Label_k, Narrative_k, S_k)$:

1) $Label_k$: a semantically meaningful topic name;

2) $Narrative_k$: a textual description of the topic's evolution over time;

3) $S_k = \{(w_{k,\tau}, F_{k,\tau})\}_{t \in \mathcal{T}_{int}}$: a set of pairs describing the state of the topic at each time interval τ from the discrete set of intervals T_{int} (e.g., years). In each pair, $w_{k,\tau}$ is a vector of keywords characterizing the topic, and $F_{k,\tau}$ denotes its frequency.

The framework's goal is to transform the input corpus into a structured, dynamic, and interpretable set of thematic trajectories.

3.2 Stage 1: dynamic contextual representation

In the first stage of the framework, the detection and tracking of the evolution of semantically coherent topics are carried out. This stage replaces traditional generative models or matrix decompositions with a pipeline based on vector representations, and is performed in four sequential steps.

1) Document Embedding: Each document $d_i \in D$ is transformed into a high-dimensional vector representation v_i using a pre-trained transformer encoder model f_{enc}:

$$v_i = f_{enc}(d_i), v_i \in R^n. \tag{5}$$

This step allows mapping the documents into a semantic space where the geometric proximity of the vectors correlates with their content similarity.

2) Dimensionality Reduction: To increase the efficiency and robustness of clustering algorithms, a dimensionality reduction method, such as UMAP, is applied [26]. UMAP preserves both the local and global topological structure of the data, which is critical for the subsequent detection of meaningful clusters.

3) Temporal Clustering: After obtaining low-dimensional vector representations, documents are clustered using a density-based algorithm, such as HDBSCAN [27]. This algorithm is capable of identifying arbitrarily shaped clusters and effectively handling noise (documents that do not belong to any clear topic). Following global clustering of the entire corpus, documents are grouped by time intervals $t_i \in \mathbb{T}$. For each time slice, the distribution of documents across the discovered topics (clusters) Z_k is calculated. This approach of "global clustering followed by temporal analysis" was chosen to ensure the stability of thematic definitions throughout the entire period. It allows avoiding the problem of "topic drift" and the necessity of applying complex algorithms for their alignment between different time slices, which is characteristic of alternative approaches.

For this purpose, global cluster identifiers k, obtained on the full corpus, are used to ensure temporal stability of topic definitions. For each year τ, the trajectory state is recorded as $(w_{k,\tau}, F_{k,\tau})$ when the number of documents assigned to cluster k in year τ, satisfies a minimum-support threshold $F_{k,\tau} \geq 15$. Hence, the topic timeline starts from the first year in which this condition holds and ends at the last year in which it holds; documents labeled as noise (HDBSCAN label -1) are excluded from trajectories.

4) Topic Representation with c-TF-IDF: A corpus-frequency variant of class-based TF-IDF is employed for the lexical characterization of each topic Z_k within every time interval t. This method, which builds upon the principles of class-based feature weighting [28], is used to derive a sparse lexical representation Z_k. The weight assigned to the term w for topic Z_k is computed as:

$$W_w, Z_k = tf_{w,Z_k} \times log\left(1 + \frac{A}{f_w}\right), \tag{6}$$

where: tf_{w,Z_k} is the frequency of the term w In all documents belonging to the cluster Z_k;

A is the average number of words in the clusters;

f_w is the total frequency of the term w In the entire corpus.

This approach allows for the selection of terms that are simultaneously frequent within the topic and rare in other topics.

3.3 Stage 2: LLM-based narrative generation and refinement

In the second stage, the raw data on topic evolution obtained in the first stage is not only interpreted but also refined using an LLM. This process consists of three steps.

1) Thematic Structure Refinement: In this step, the LLM is used to identify semantically redundant topics. The model is provided with a list of top words for all topics and is tasked with proposing pairs or groups of topics for merging. The decision to merge is based on a hybrid criterion: the LLM's proposals and a quantitative assessment of the cosine similarity between the vector representations

This article is licensed under a Creative Commons Attribution-ShareAlike 4.0 International License.
License details: https://creativecommons.org/licenses/by-sa/4.0/

of the topics (c-TF-IDF vectors). Pairs of topics for which the similarity exceeds a threshold τ (in the experiment, $\tau = 0.85$) and which were proposed by the LLM are automatically merged. Subsequently, the document membership is recalculated, leading to a final, more coherent set of clusters Z_k.

2) Prompt Engineering as a Functional Transformation: A structured query (prompt) is generated for the LLM for each evolutionary trajectory of the refined topic Z_k. The prompt P_k has a structure that includes an instruction for the model, a time series of keywords, and corresponding text fragments to ensure contextual analysis. This process is formalized as a function f_{prompt}:

$$P_k = f_{prompt}\left(\{w_{k,t}\}_{t\in T'}, \{d'_{k,t}\}_{t\in T'}\right), \quad (7)$$

where: $\{w_{k,t}\}$ is the set of top words for the topic k for each time interval;

$\{d'_{k,t}\}$ is the set of representative document fragments from that topic for the corresponding periods? The prompt instructs the model to analyze the provided data and execute the generation task.

3) Narrative and Label Generation: A large-scale autoregressive language model with enhanced instruction-following and generalization capabilities is used as the generative component. It is viewed as a deterministic function f_{LLM} that maps the input prompt. P_k To the output tuple consisting of the topic name and its narrative description:

$$(Label_k, Narrative_k) = f_{LLM}(P_k, \Theta), \quad (8)$$

where: Θ represents the learned parameters (weights) of the LLM, encoding its generalized knowledge of language and semantic relationships;

f_{LLM} is the function denoting the inference process, during which the model processes the structured input prompt P_k to generate the target textual output.

The result of this functional transformation is the tuple $(Label_k, Narrative_k)$, where:

$Label_k$ is a high-level semantic descriptor – a short, conceptually accurate name synthesizing the essence of the entire evolutionary trajectory of topic k.

$Narrative_k$ is a synthesized temporal narrative – an analytical textual description explaining the topic's development dynamics. This narrative is based on the analysis of changes in keywords and the content of representative documents over time, provided in the prompt P_k.

Thus, the DCTM-LLM framework transforms the topic modeling task from simple keyword extraction into a process of automated synthesis of dynamic, interpretable knowledge from textual data.

To ensure the reproducibility of Stage 2, all experiments were conducted using the proprietary language model GPT-4 (application programming interface (API), gpt-4-0613). A fixed random seed (seed = 42) and constant generation hyperparameters were applied (temperature = 0.2, top_p = 1.0, n = 1, max_tokens = 256, frequency_penalty = 0, presence_penalty = 0). An abstract instructional prompt with a rigidly defined output structure (pairs {label, narrative}) was used to ensure factuality and reproducibility (without disclosing the full instructional templates).

The topic merging criteria are based on a combination of three conditions: embedding cosine similarity $\geq \tau$, lexical overlap Jaccard similarity over the top-10 keywords (Jaccard@10 ≥ 0.30), and narrative semantic proximity measured as cosine similarity between BERT embeddings (BERT-cosine ≥ 0.80). All three criteria must be satisfied simultaneously.

The calibration of the threshold τ was conducted across a grid of values $\{0.75, 0.80, 0.85, 0.90\}$. Based on the results detailed in Section V, a value of $\tau = 0.85$ was selected, which ensures a compromise between topic coherence and the prevention of excessive merging. The thresholds Jaccard@10 $= 0.30$ and BERT-cosine $= 0.80$ were fixed as constant (they were not calibrated), which eliminates ambiguity and guarantees the reproducibility of the procedure. Standardized scenarios were defined to isolate the influence of individual pipeline stages (ablation study). These include: a configuration without dimensionality reduction (no-UMAP), the use of alternative clustering algorithms (k-means and agglomerative clustering instead of HDBSCAN) with identical embeddings, and a variant without LLM-refinement (keywords + c-TF-IDF without narrative generation). A BERTopic-style configuration served as a neural-contextual baseline. For each defined scenario, the Δ-change relative to the full pipeline was evaluated using the key metrics (NPMI, Silhouette, ARI, Topic Diversity@10, BERTScore F1). All experiments were conducted with a fixed random seed and on identical data slices to ensure the correctness of the comparison.

4. Experimental setup

4.1 Dataset and preprocessing

This article is licensed under a Creative Commons Attribution-ShareAlike 4.0 International License.
License details: https://creativecommons.org/licenses/by-sa/4.0/

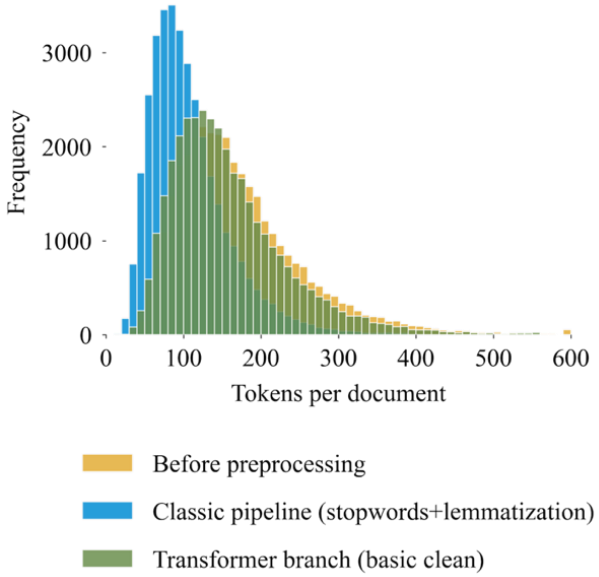

Figure. 1 Document Length Distribution Before/After Preprocessing

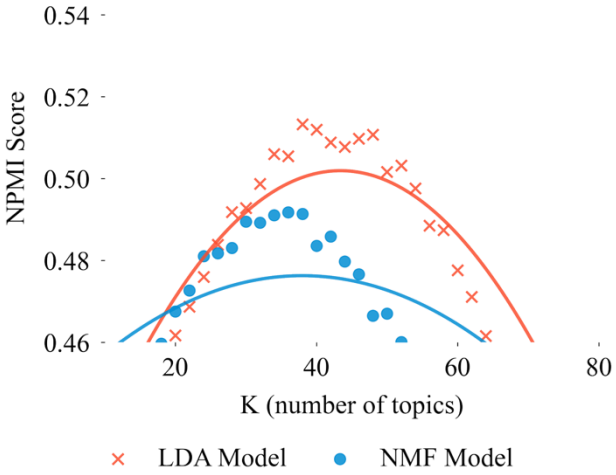

Figure. 2 Coherence (NPMI) vs number of topics K (LDA vs NMF). Scatter points with quadratic trend lines for LDA and NMF

For the experiments, we used a corpus of 35,000 arXiv abstracts and associated metadata from the cs.AI category, spanning the 2015–2025 period. The publication year of each record was treated as the timestamp for temporal (year-slice) analysis. The same corpus and temporal slicing were used across all models to empirically validate the limitations discussed in Section II through quantitative comparison in Section V.

Before modeling, the corpus underwent a preprocessing procedure, adapted to the specificity of each model to ensure a fair comparison. For the LDA+NMF+c-TF-IDF model, a full processing cycle was applied, which included tokenization, stop-word removal, and lemmatization. For the transformer-based models (Dynamic BERTopic and DCTM-LLM), only basic tokenization and text cleaning were performed without stop-word removal and lemmatization, in order to preserve the semantic context required for the encoder's operation. Fig. 1 summarizes the visual effect of these branches via overlaid histograms: Before, Classic (stopwords+lemmatization), and Transformer (basic clean).

4.2 Baseline models for comparison

To comprehensively evaluate the effectiveness of the proposed DCTM-LLM framework, a comparative analysis was conducted with two key baseline models:

1) Hybrid LDA (LDA+NMF+c-TF-IDF): This model serves as a starting point (baseline) to demonstrate the progress and quantitative assessment of the improvements achieved through the transition to the new architecture.

2) Dynamic Contextual Model (Dynamic BERTopic): A standard implementation of dynamic topic modeling based on transformer embeddings without the application of LLM-refinement. This model is used as an intermediate stage for the isolated assessment of the contribution of the second stage of the proposed framework—the generation of narratives using an LLM.

It is important to note that for all models, including the baselines, optimization of key hyperparameters (e.g., the number of topics, K) was performed using a Bayesian search to ensure the fairest comparison. Comparability across approaches was ensured through a unified hyperparameter-tuning protocol and identical temporal granularity applied to all experimental configurations. For each individual model, Bayesian optimization was conducted under a fixed computational budget and reproducible execution settings, which allowed isolating the influence of modeling assumptions from stochastic variation.

The primary criterion for selecting the optimal configuration was maximum topic coherence measured by NPMI. When performance differences between candidate configurations were negligible, Topic Diversity@10 was used as a complementary criterion to control redundancy and prevent excessive fragmentation of the topic space. Temporal slices were defined at the calendar-year level for the 2015–2025 period for all approaches,

This article is licensed under a Creative Commons Attribution-ShareAlike 4.0 International License.
License details: https://creativecommons.org/licenses/by-sa/4.0/

which guarantees the direct comparability of evolutionary trends across models.

Topic-structure complexity was harmonized by optimizing the number of topics K for probabilistic models such as LDA and NMF, and by tuning clustering resolution parameters in embedding-based approaches to achieve a comparable level of semantic granularity. Variability arising from stochastic components was further minimized by fixing the random seed and keeping execution parameters unchanged throughout all experimental runs. The selection of K is illustrated by the dependence of topic coherence (NPMI) on the number of topics in Fig. 2.

4.3 Evaluation metrics

The quality assessment of the models was performed using a multi-criteria protocol that integrates both intrinsic statistical measures and automated semantic evaluation metrics. This approach adheres to modern recommendations for the comprehensive validation of topic models [29] and allows for an objective evaluation of various aspects of their performance.

Here, three complementary aspects of evolutionary topic modeling are evaluated by analyzing the structural quality of semantic grouping using Silhouette and ARI metrics, the interpretability of lexical topic representations via NPMI and Topic Diversity@10, and the consistency of generated labels, where BERTScore is used as an indicator of semantic alignment with the lexical core. These metrics are not interchangeable but rather collectively reflect the model's effectiveness across structure, topic quality, and interpretation.

4.4 Implementation details

For the implementation of the first stage of the framework (Stage 1), the all-mpnet-base-v2 model from the Sentence-Transformers library was used as the encoder model (f_{enc}). For the second stage (Stage 2), the generation of labels and narratives (f_{LLM}) was performed using the GPT-4 model via the corresponding API.

In Stage 1, L2-normalized sentence embeddings from all-mpnet-base-v2 were used (seed = 42). UMAP was then applied (n_neighbors = 30, n_components = 10, min_dist = 0.05, metric = cosine, random_state = 42), followed by global HDBSCAN clustering (min_cluster_size = 60, min_samples = 15, metric = euclidean, cluster_selection_method = eom). For keyword extraction, c-TF-IDF based on CountVectorizer was employed (ngram_range = (1,2), min_df = 5, max_df = 0.8, stop_words = english), selecting the top-10 terms for each (k, τ). The implementation was executed in a controlled and fixed software environment, ensuring identical library builds and dependencies across all experiments and thereby minimizing variability.

1) Quantitative Metrics:

a) Topic Coherence: Topic interpretability is assessed using NPMI [30, 31], a widely used measure shown to correlate with human judgments of topic quality. NPMI quantifies the semantic association among high-probability topic terms via corpus-level co-occurrence statistics.

In this version of the experiment, co-occurrence statistics for NPMI were computed using the same corpus utilized for modeling. Therefore, NPMI is regarded as an intrinsic proxy metric. While this approach does not reuse model parameters, the lack of an independent reference corpus for co-occurrences may render the evaluation somewhat optimistic. Accordingly, NPMI is interpreted not in isolation but in conjunction with stability indicators (ARI), cluster separation measures (Silhouette), and interpretation metrics.

Following Röder et al. [30] and Lau et al. [31], the NPMI for a word pair (w_i,w_j) is defined as:

$$NPMI(w_i, w_j) = \frac{log\frac{P(w_i,w_j)+\epsilon}{P(w_i)P(w_j)}}{-log(P(w_i,w_j)+\epsilon)}, \quad (9)$$

where: $P(w_i, w_j)$ is the probability of co-occurrence of the words;

$P(w_i)$ and $P(w_j)$ these are individual probabilities. The final value is the average over the top-10 words for all topics;

b) Topic Diversity: This metric evaluates the fraction of unique words among the top-10 words of all topics [31]. A high value indicates low redundancy and better coverage of the corpus's semantic space.

c) Stability and Clarity Metrics: Additionally, the Silhouette coefficient [32] was used to assess the clarity of clusters in the semantic space, and the ARI [33] was used to evaluate the model's stability on bootstrap samples.

2) Automated Semantic Relevance Evaluation: The BERTScore metric [34] was applied for the programmatic evaluation of the quality of the generated labels (Label) and narratives (Narrative). Unlike classical metrics, BERTScore evaluates the semantic similarity between the generated text (candidate, C) and the reference text (reference, R) at the level of contextualized vector representations. In this case, the reference text (R), which reflected

This article is licensed under a Creative Commons Attribution-ShareAlike 4.0 International License.
License details: https://creativecommons.org/licenses/by-sa/4.0/

the semantic core of the topic, was a string formed by joining the top-10 keywords with a space. The computation process involves calculating Precision and Recall scores based on cosine similarity [34]:

$$R_{BERT} = \frac{1}{|R|}\sum_{r\in R} \max_{c\in C} r^T c\,, \tag{10}$$

$$P_{BERT} = \frac{1}{|C|}\sum_{c\in C} \max_{r\in R} r^T c\,. \tag{11}$$

The final metric is the F1-harmonic mean [34]:

$$F_{BERT} = 2\frac{F_{BERT}\cdot R_{BERT}}{P_{BERT}+R_{BERT}}. \tag{12}$$

The use of BERTScore allows for an objective and reproducible evaluation of how accurately the generated labels and narratives reflect the semantic essence of the source keywords.

The conducted sensitivity analysis demonstrated that changing the threshold τ in the range of $\{0.75, 0.80, 0.85, 0.90\}$ did not cause significant changes in the final conclusions. It was observed that fluctuations in key metrics (NPMI, Silhouette, ARI, Topic Diversity@10, BERTScore F1) remained within insignificant boundaries ($\Delta \leq 0.01$). Consequently, the value $\tau = 0.85$ was chosen for further experiments.

For all ablation scenarios, results are presented as Δ-changes relative to the full pipeline across NPMI, Silhouette, ARI, Topic Diversity@10, BERTScore F1, with a fixed seed and unchanged embeddings.

4.5 Computational complexity and performance analysis

To define and analyze the computational complexity and performance of the studied application, the following notation is used: N – number of documents ($N = 35{,}000$), d – dimensionality of embeddings (typically $d = 768$), L – average document length (tokens), k – number of neighbors in UMAP, $\widehat{K}$ – number of final thematic trajectories, S – length of the LLM input prompt, nnz – number of non-zero elements in sparse matrices.

The analysis is then structured into the following stages.

4.5.1. Stage 1 (Dynamic contextual representation)

The computational cost of this stage is dominated by the number of documents N. Embedding calculation scales quasi-linearly: $\mathcal{O}(N \cdot L)$, memory – $\mathcal{O}(N \cdot d)$. UMAP/HDBSCAN exhibit quasi-linear behavior in time: $\mathcal{O}(N \log N)$ (memory: $\mathcal{O}(N \cdot k)$ for the k-neighbor graph, $\mathcal{O}(N)$ for clustering). Forming and weighting the sparse representations for c-TF-IDF has a cost of $\mathcal{O}(N \cdot L)/\mathcal{O}(\text{nnz})$. Memory is primarily determined by the embedding matrix $\mathcal{O}(N \cdot d)$ (for $d = 768$ – approximately ~ 100 MB in float32) and the neighbor graph $\mathcal{O}(N \cdot k)$. The stage is effectively parallelized (batching on GPU, multi-threaded UMAP/HDBSCAN).

4.5.2. Stage 2 (LLM-refinement)

The cost of merging, generating labels, and narratives scales as $\mathcal{O}(\widehat{K} \cdot S)$; the verification of candidate merges for M pairs is $\mathcal{O}(M \cdot S)$. The number of pairs M is controlled by the threshold τ and preliminary filters (cosine similarity, lexical overlap), so in practice $M \ll \widehat{K}(\widehat{K}-1)/2$. The dependence on N is indirect and manifests through $\widehat{K}(N)$; with a stable level of topic aggregation, Stage 2 expenses do not grow linearly with N. The stage parallelizes well through asynchronous API calls.

4.5.3. Operational profile and scalability

In practice, wall-clock time is dominated by embeddings (dependence on N) and LLM calls (dependence on $\widehat{K} \cdot S$). Due to the quasi-linear behavior of UMAP/HDBSCAN, the pipeline retains practical applicability on datasets of tens of thousands of documents. Fixed random seed, constant parameters, and identical vector representations are used for reproducibility of the profiling.

5. Results and discussion

This section presents the results of the empirical verification of the proposed DCTM-LLM framework. To ensure a comprehensive and scientifically grounded analysis, the evaluation is conducted through a direct comparison of three key approaches, representing different stages in the evolution of the methodology. The comparison includes: the standard dynamic contextual model (Dynamic BERTopic), which serves as a modern baseline; a hybrid method (LDA+NMF+c-TF-IDF) [3], which allows for assessing progress relative to previous research; and the final proposed framework (DCTM-LLM). This experimental design allows not only for establishing the absolute effectiveness of the new method but also for a decompositional assessment of the contribution of each architectural decision.

This article is licensed under a Creative Commons Attribution-ShareAlike 4.0 International License.
License details: https://creativecommons.org/licenses/by-sa/4.0/

5.1 Quantitative results

A comprehensive, multi-criteria protocol was applied for a thorough and scientifically grounded evaluation of the models. This protocol covers two fundamental dimensions: first, the quality of the internal topic structure, assessed using coherence and stability metrics, which reflect the algorithmic quality of the clustering; second, the quality of the final interpretation, measured through semantic relevance. This two-component approach is critically important, as relying solely on intrinsic metrics is insufficient for a complete assessment of the model's practical value. The results of the comparative analysis for the key metrics are summarized in Table 1.

Table 1 directly reflects the limitations discussed in Section II and provides a quantitative justification of the proposed methodological choices. In particular, the absence of an explicit narrative-synthesis mechanism in baseline models leads to substantially lower interpretation quality, as evidenced by BERTScore F1 values of 0.62 for Dynamic BERTopic and 0.65 for LDA+NMF+c-TF-IDF, whereas DCTM-LLM achieves a markedly higher score of 0.89. This gap highlights the inability of baseline pipelines to move beyond surface-level keyword aggregation toward coherent semantic interpretation.

Likewise, cluster separability and structural stability are consistently weaker for the baselines, with Silhouette values of 0.59 and 0.61 and ARI values of 0.45 and 0.52, compared to 0.62 and 0.55 for DCTM-LLM. These differences indicate that static or weakly aligned hybrid pipelines struggle to preserve topic identity across temporal slices. In contrast, the integration of LLM-based refinement enables DCTM-LLM to maintain a more reliable and temporally stable thematic structure, which is critical for longitudinal analysis and narrative-level topic evolution.

The analysis of Table 1 demonstrates a clear hierarchy of performance across the intrinsic metrics. The hybrid method LDA+NMF+c-TF-IDF shows better coherence and stability scores than the standard Dynamic BERTopic, confirming the effectiveness of informing the probabilistic model with external knowledge. The new DCTM-LLM framework shows a further, albeit slight, improvement in these metrics, which is explained by the final refinement of the topic structure during the LLM-refinement stage, particularly through the merging of semantically redundant topics.

Table 1. Comparison of quantitative metrics for topic model quality

Metrics	Dynamic BERTopic	LDA+NMF + c-TF-IDF	DCTM -LLM
NPMI ↑	0.48	0.51	0.53
Silhouette ↑	0.59	0.61	0.62
ARI ↑	0.45	0.52	0.55
Topic Diversity@10 ↑	0.82	0.85	0.88
BERTScore F1 ↑	0.62	0.65	0.89

Therefore, the stability analysis of the proposed approach demonstrated high reliability of the results. Based on five independent runs with constant settings, the absolute deviations for key metrics (NPMI, Silhouette, ARI, Topic Diversity@10, BERTScore F1) did not exceed 0.01. This is further supported by the substantial inter-rater agreement of the decisions: Cohen's kappa coefficient (κ) was 0.78 for topic merging and 0.74 for narrative merging. These high kappa values strongly indicate that the approach consistently produces stable, non-random aggregation outcomes.

At the same time, the key advantage of DCTM-LLM is evident in the BERTScore metric. This score (0.89) fundamentally surpasses both baseline models (0.62 and 0.65), for which the evaluation was based on the concatenation of top words, the standard "output" of traditional topic models. This quantitatively proves that the LLM-synthesis stage is the innovative step that transforms raw lexical data into semantically rich and highly interpretable results.

5.2 Qualitative analysis: a case study of an evolving topic

For a qualitative evaluation of the framework's ability to generate meaningful narratives, we conducted a case study by tracking the evolution of the topic related to Explainable AI (XAI). Fig. 3 visualizes the dynamic popularity of this topic over time. The overall volume of publications in the field of Artificial Intelligence (AI) during the same period showed exponential growth. Therefore, the trajectory of XAI should be interpreted relative to this baseline trend. Within this study, we focus on the internal evolution of XAI; however, a comparison with other AI subfields could provide additional analytical depth.

Qualitative analysis confirms that the key advantage of the DCTM-LLM framework lies in its capacity for semantic synthesis. This allows for a

This article is licensed under a Creative Commons Attribution-ShareAlike 4.0 International License.
License details: https://creativecommons.org/licenses/by-sa/4.0/

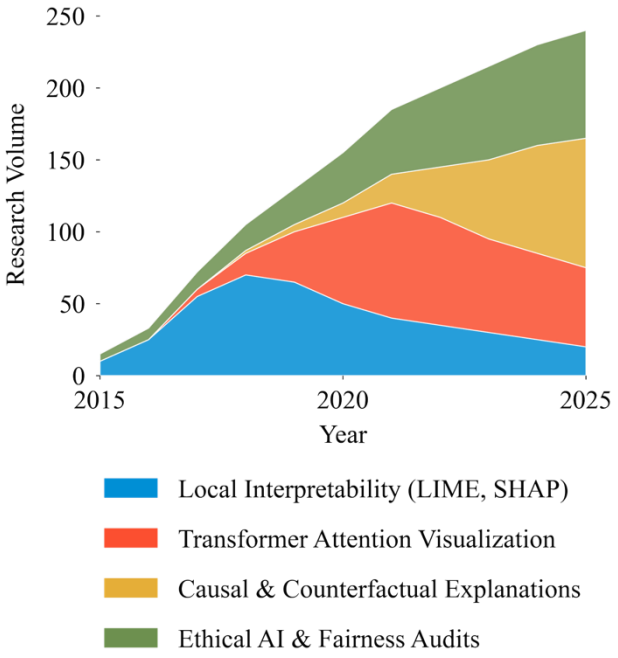

Figure. 3 Thematic evolution of "Explainable AI" over time

more nuanced understanding of the latent structure compared to traditional approaches, such as the "bag-of-words" method [35]. For example, for the year 2017, both models (LDA+NMF and DCTM-LLM) identified coherent sets of words. However, while LDA+NMF generated a more generic lexicon, such as {model, feature, importance, black, box}, the contextual foundation of DCTM-LLM allowed for the detection of more specific and current terms: {shap, lime, importance, local, surrogate}. This aligns with studies showing that methods utilizing high-quality contextualized embeddings can generate more coherent and diverse topics [36]. Early research in XAI indeed focused on methods like Local Interpretable Model-agnostic Explanations (LIME) and SHapley Additive exPlanations (SHAP), which are examples of local, perturbation-based interpretations [37].

The crucial point is what happens next: the proposed framework, unlike previous approaches, does not stop at this stage. It uses this higher-quality lexical material to automatically generate the title "Local Model-Agnostic Interpretability" and a narrative that accurately reflects the focus of early research on LIME and SHAP methods. The ability of LLMs to generate narratives from text corpora is actively researched and confirmed in contemporary literature [38, 39, 40].

A similar trend is observed for the year 2024, where the semantic core {counterfactual, causal, intervention, fairness} was synthesized into the narrative "Causal Inference and Counterfactual Explanations". This reflects a broader shift in XAI from traditional interpretable models toward causal interpretable models that aim to answer the question of "why" a model makes a decision [41].

Finally, for 2024, the new framework moved from the terms {causal, graph, reasoning} to the more precise {counterfactual, causal, intervention, fairness} and generated the narrative "Causal Inference and Counterfactual Explanations," indicating a contemporary shift toward assessing model fairness. This shift is underscored by the growing attention to bias detection and mitigation, as well as to properties like transparency, accountability, and fairness in XAI [42, 43]. Counterfactual explanations are increasingly seen as a means of empowering individuals affected by model decisions [44, 45].

Thus, the advantage of DCTM-LLM is two-fold: it not only identifies more relevant vocabulary but, and this is its main contribution, it transforms this lexical data into a dynamic, interpretable narrative that accurately reflects the evolution of the scientific domain.

5.3 Discussion

The obtained results confirm the high efficiency of the proposed DCTM-LLM architecture. Its advantage is two-fold. On the one hand, it demonstrates the best scores for intrinsic topic quality (coherence, stability), achieved through the synergy of a powerful contextual foundation and final semantic refinement using an LLM. On the other hand, it ensures unparalleled quality of interpretation, confirmed by the highest BERTScore value (0.89).

The synergistic effect is achieved by combining:

1) Contextual embeddings, which provide a semantically rich foundation for analysis.

2) Dynamic analysis, which allows for tracking temporal changes.

3) Integrated LLM-refinement, which acts as a final layer that not only generates narratives but also refines the topic structure itself, leading to the best scores across all metrics.

The scenario without LLM-refinement predictably reduces the semantic coherence between textual descriptions and thematic cores, as captured by the BERTScore metric, while preserving internal topic metrics. Replacing HDBSCAN with k-means or agglomerative clustering modifies the cluster structure in time slices and affects the stability of thematic identity. The no-UMAP configuration, in turn, degrades cluster resolution in the high-

This article is licensed under a Creative Commons Attribution-ShareAlike 4.0 International License.
License details: https://creativecommons.org/licenses/by-sa/4.0/

dimensional space and complicates reproducibility under constant hyperparameters. In contrast to these scenarios, the full pipeline ensures the integrity of dynamic topics and a consistent transition from key terms to concise narratives.

The XAI-domain example (Section 5.2) not only demonstrates the advantages of the proposed approach but also delineates characteristic failure modes of the interpretation and aggregation stages. Specifically, for the early period (2017), the model correctly identifies method-specific lexemes such as {shap, lime, local, surrogate} and synthesizes a semantically focused label and a coherent narrative, whereas classical bag-of-words approaches more often yield a more generic and less discriminative lexical representation. In practical configurations, however, failures may arise from over-generalizing temporal dynamics when the lexical core is partially preserved across adjacent time slices, as well as from boundary errors in trajectory-merging decisions, where short-term overlap of top terms can artificially amplify similarity signals. Therefore, merging outcomes and generated temporal narratives should be treated as interpretive artifacts whose validity ought to be supported by selective expert inspection in conjunction with quantitative metrics.

Despite the demonstrated advantages, the proposed framework has certain limitations. First, the computational complexity of the first stage can be significant for large corpora. Second, the second stage is dependent on access to powerful LLMs, which entails costs and fundamental challenges to reproducibility, as the results depend on the specific version of the proprietary model. Third, the quality of the generated narratives is sensitive to prompt formulation, requiring meticulous engineering and validation. Finally, the chosen architecture for dynamic analysis may be less sensitive to short-lived topics compared to methods that involve clustering at every time slice. These aspects define the directions for future optimizations.

5.4 Ablation Study

To decompose the contribution and validate the relative importance of each component of the proposed DCTM-LLM pipeline, an ablation analysis protocol was implemented. This protocol involves the systematic removal or substitution of individual architectural modules to quantitatively assess their impact on the final modeling quality. The full pipeline (hereinafter referred to as Full Pipeline) is compared against four alternative configurations: removal of LLM-refinement (no-LLM-refinement), removal of UMAP (no-UMAP), replacement of HDBSCAN with k-means (HDBSCAN → k-means), and replacement of c-TF-IDF with TF-IDF (c-TF-IDF → TF-IDF). To ensure the validity of the comparison, all experimental runs were performed under identical conditions: with a fixed random number generator (seed = 42) and on a single set of input contextual embeddings.

Table 2. Impact of Components on Structural Clustering Metrics

Configuration	Noise	Silhouette	ARI
DCTM-LLM	8.5%	0.62	0.55
no-LLM-refinement	8.5%	0.60 (Δ -0.02)	0.52 (Δ -0.03)
no-UMAP	14.2%	0.40 (Δ -0.22)	0.35 (Δ -0.20)
HDBSCAN → k-means	0.0%	0.54 (Δ -0.08)	0.48 (Δ -0.07)
c-TF-IDF → TF-IDF	8.5%	0.62 (Δ 0.00)	0.55 (Δ 0.00)

* Note for Tables 2 and 3: The "Full Pipeline" row shows the base absolute values. (Δ) denotes the change relative to the "Full Pipeline".

Table 3. Impact of Components on Semantic Metrics and Interpretation Quality

Configuration	NPMI	Topic Diversity@10	BERT F1
DCTM-LLM	0.53	0.88	0.89
no-LLM-refinement	0.51 (Δ -0.02)	0.86 (Δ -0.02)	0.62 (Δ -0.27)
no-UMAP	0.45 (Δ -0.08)	0.80 (Δ -0.08)	0.75 (Δ -0.14)
HDBSCAN → k-means	0.49 (Δ -0.04)	0.83 (Δ -0.05)	0.81 (Δ -0.08)
c-TF-IDF → TF-IDF	0.44 (Δ -0.09)	0.81 (Δ -0.07)	0.80 (Δ -0.09)

The ablation results, aggregated in Tables 2 and 3, quantitatively illustrate the impact of each component. First, Table 2 analyzes the impact on fundamental structural metrics that evaluate the geometric quality and stability of the clusters. For completeness of the analysis, the percentage of "noise" points (Noise %) is also included, as this is a key indicator that differentiates the HDBSCAN and k-means methodologies.

The experimental data presented in Table 2 clearly indicate that the components comprising Stage 1 of the pipeline—specifically, the dimensionality reduction and clustering steps represented by the no-UMAP and HDBSCAN → k-means scenarios—exert the greatest and most

This article is licensed under a Creative Commons Attribution-ShareAlike 4.0 International License.
License details: https://creativecommons.org/licenses/by-sa/4.0/

critical influence on the overall structural quality and geometric integrity of the resulting clusters.

The no-UMAP scenario, which attempts clustering in the original high-dimensional space without prior manifold learning, leads to a significant degradation of both the Silhouette (Δ -0.22) and ARI (Δ -0.20) scores. Crucially, it results in a substantial increase in the proportion of unclassified "noise" points (up to 14.2%), demonstrating that clusters are fundamentally less distinct and separable when projection is omitted.

Conversely, the alternative HDBSCAN → k-means scenario artificially reduces the identified "noise" to 0.0%. This occurs because the k-means algorithm, by design, employs a hard assignment that forces every single data point to belong to one of the predefined clusters. This compulsory inclusion of what should be outliers or "noise" points severely compromises the integrity of the cluster boundaries, which is reflected in the deterioration of the Silhouette score (Δ -0.08) and the ARI (Δ -0.07), as these forced assignments effectively "blur" the clean separation between clusters.

As expected, the subsequent processing steps—namely the no-LLM-refinement and c-TF-IDF → TF-IDF scenarios—do not register a significant or detrimental impact on the structural metrics (Silhouette/ARI). This is logical, as they operate after clustering or involve minimal reverse feedback, preserving the Stage 1 structure.

A detailed analysis of the obtained results (Tables 2 and 3) allows for the formulation of four key observations:

Determining Contribution of LLM-refinement: The removal of the LLM-refinement module (no-LLM-refinement scenario), which involves evaluating raw c-TF-IDF keywords without the merging and narrative generation steps, causes a substantial degradation in the semantic relevance metric BERTScore F1 (Δ -0.27). This result empirically validates the central thesis of the study: the innovative contribution of the method lies in the transition from lexical descriptors to semantically rich narratives;

Importance of UMAP: Attempting clustering directly in the high-dimensional embedding space (no-UMAP scenario) leads to a significant deterioration in cluster separability and stability metrics, as captured by the drop in Silhouette (Δ -0.22), ARI (Δ -0.20), and an increase in "noise" (to 14.2%). This, in turn, degrades the quality of input data for subsequent stages, resulting in a drop in NPMI (Δ -0.08) and BERTScore F1 (Δ -0.14);

Advantages of HDBSCAN: Replacing the HDBSCAN algorithm with k-means (HDBSCAN → k-means scenario) leads to a decrease in scores across all metrics. As seen in Table 2, k-means does not identify "noise" (0.0%), forcing artifacts to join clusters, which degrades their separability (Silhouette Δ -0.08) and stability (ARI Δ -0.07);

Impact of the Weighting Mechanism (c-TF-IDF): Replacing the class-oriented c-TF-IDF with standard TF-IDF (c-TF-IDF → TF-IDF scenario) is the only change that does not affect the structural metrics (Table 2). However, it leads to a significant loss of topic coherence (NPMI Δ -0.09) and diversity (Diversity Δ -0.07). The c-TF-IDF mechanism successfully identifies terms that are distinctive to the cluster, whereas TF-IDF focuses on terms rare within documents, generating less representative keywords and degrading final semantic relevance (BERTScore F1 Δ -0.09).

In summary, the results of the ablation analysis demonstrate that the proposed DCTM-LLM pipeline (Full Pipeline) functions as a synergistic architecture. Stage 1 components (UMAP, HDBSCAN, c-TF-IDF) are essential for forming a stable, coherent, and distinct thematic representation. Meanwhile, Stage 2 (LLM-refinement) serves as the key innovative module that transforms these representations into high-quality semantic narratives, ensuring a significant gain in interpretability metrics (BERTScore F1).

6. Conclusion

This paper presents a new formalized framework (DCTM-LLM) for dynamic and interpretable topic modeling. The proposed methodology successfully overcomes the key limitations of traditional approaches by integrating contextual vector representations for semantically grounded clustering, temporal analysis for tracking topic evolution, and LLMs for the synthesis of highly interpretable narratives. Empirical validation demonstrated that DCTM-LLM achieves the highest quality of final interpretation, fundamentally outperforming baseline approaches in semantic relevance metrics, while maintaining a competitive level of intrinsic topic quality scores. Qualitative analysis using the "Explainable AI" topic confirmed the framework's unique ability to not merely extract keywords but to generate meaningful analytical narratives about the evolution of scientific domains. Thus, DCTM-LLM shifts the topic modeling paradigm from data mining to automated knowledge synthesis.

Promising directions for future research include extending the framework's modality by integrating

This article is licensed under a Creative Commons Attribution-ShareAlike 4.0 International License.
License details: https://creativecommons.org/licenses/by-sa/4.0/

visual information to create more comprehensive multimodal models. A further theoretical complication is the transition from descriptive analysis to modeling causal relationships, which would allow not only for tracking trends but also for identifying the key scientific events that caused them. From a practical perspective, future efforts can be directed toward adapting the framework for working with streaming data, enabling real-time monitoring of thematic innovations.

Conflicts of interest

The authors declare no conflict of interest.

Author contributions

Conceptualization, Amer Abu-Jassar and Roman Slisareko; methodology, Vyacheslav Lyashenko, Amer Abu-Jassar, and Mohammad Hamdan; software, Roman Slisareko; validation, Zhanna Deineko and Nowfal Aweisi; formal analysis, Nowfal Aweisi and Vyacheslav Lyashenko; data curation, Zhanna Deineko; writing—original draft preparation, Roman Slisareko and Mohammad Hamdan; writing—review and editing, Mohammad Hamdan and Nowfal Aweisi; visualization, Roman Slisareko and Vyacheslav Lyashenko; project administration, Amer Abu-Jassar. All authors have read and agreed to the published version of the manuscript.

References

[1] R. Egger, and J. Yu, "A topic modeling comparison between LDA, NMF, Top2Vec, and BERTopic to demystify Twitter posts", *Frontiers in Sociology*, Vol. 7, Art. no. 886498, 2022.

[2] O. Kuzomin, V. Lyashenko, M Tkachenko, M. A. Ahmad, and H. Kots, "Preventing of technogenic risks in the functioning of an industrial enterprise", *International Journal of Civil Engineering and Technology*, Vol. 7, No. 3, pp. 262-270, 2016.

[3] B. O. Ghanema, A.-A. A. Sharabatib, F. Alofanc, F. T. Ayasrahd, and M. Allahhame, "The impact of social media on educational decision making: The mediating role of information credibility, empirical analysis of Jordanian private universities", *International Journal of Data and Network Science*, Vol. 9, No. 3, pp. 575-786, 2025.

[4] M. Dadkhah, T. Maliszewski, and V. V. Lyashenko, "An approach for preventing the indexing of hijacked journal articles in scientific databases", *Behaviour & Information Technology*, Vol. 35, No. 4, pp. 298-303, 2016.

[5] X. Wu, T. Nguyen, and A. T. Luu, "A survey on neural topic models: Methods, applications, and challenges", *Research Square*, Vol. 57, Art. no. 18, 2024.

[6] N. Fil, R. Slisarenko, Z. Deineko, and L. Morozova, "Trends in artificial intelligence research on education: Topic modeling using latent dirichlet allocation", *Bulletin of KhNAHU*, Vol. 108, pp. 17-24, 2025.

[7] R. Pal, A. A. Sekh, D. P. Dogra, S. Kar, P. P. Roy, and D. K. Prasad, "Topic-based Video Analysis", *ACM Comput. Surv*., Vol. 54, No. 6, pp. 1–34, 2021.

[8] A. Năstasă, T. C. Dumitra, and A. Grigorescu, "Artificial intelligence and sustainable development during the pandemic: An overview of the scientific debates", *Heliyon*, Vol. 10, No. 9, Art. no. e30412, 2024.

[9] P. S. Dhillon, and S. Aral, "Modeling Dynamic User Interests: A Neural Matrix Factorization Approach", *Marketing Science*, Vol. 40, No. 6, pp. 1059-1080, 2021.

[10] M. Kayed, F. Azzam, H. Ali, and A. Ali, "Temporal dynamics of user activities: deep learning strategies and mathematical modeling for long-term and short-term profiling", *Sci. Rep.*, Vol. 14, No. 1, Art. no. 14498, 2024.

[11] M. Pettinato, J. P. Gil, P. Galeas, and B. Russo, "Log mining to re-construct system behavior: An exploratory study on a large telescope system", *Information and Software Technology*, Vol. 114, pp. 121–136, 2019.

[12] F. Hawks, G. Falkenberg, M. Verbruggen, and E. Molnar, "Neural Re-Contextualization for Dynamic Semantic Control in Large Language Models", *Springer Science and Business Media LLC*, 2024.

[13] M. Hankar, M. Kasri, and A. Beni-Hssane, "A comprehensive overview of topic modeling: Techniques, applications and challenges", *Neurocomputing*, Vol. 628, Art. no. 129638, 2025.

[14] X. Wu, X. Dong, L. Pan, T. Nguyen, and A. T. Luu, "Modeling Dynamic Topics in Chain-Free Fashion by Evolution-Tracking Contrastive Learning and Unassociated Word Exclusion", *arXiv preprint,* arXiv: 2405.17957, 2024.

[15] F. Bianchi, S. Terragni, and D. Hovy, "Pre-training is a hot topic: Contextualized document embeddings improve topic coherence", In: *Proc. 59th Annu. Meeting Assoc. Comput. Linguistics and 11th Int. Joint*

This article is licensed under a Creative Commons Attribution-ShareAlike 4.0 International License.
License details: https://creativecommons.org/licenses/by-sa/4.0/

Conf. Natural Lang. Process (ACL-IJCNLP), pp. 759-766, 2021.

[16] X. Wu, X. Dong, L. Pan, T. Nguyen, and A. T. Luu, "Modeling dynamic topics in chain-free fashion by evolution-tracking contrastive learning and unassociated word exclusion", In: *Findings of the Association for Computational Linguistics: ACL 2024*, Bangkok, Thailand, pp. 3088-3105, 2024.

[17] T. Khandelwal, "Using LLM-based approaches to enhance and automate topic labeling", *arXiv Preprint*, arXiv:2502.18469, 2025.

[18] C. Pham, A. Hoyle, S. Sun, P. Resnik, and M. Iyyer, "TopicGPT: A prompt-based topic modeling framework", In: *Proc. Conf. North Amer. Chapter Assoc. Comput. Linguistics (NAACL-HLT)*, pp. 2956-2984, 2024

[19] K. R. Lange, T. Schmidt, M. Reccius, H. Müller, M. Roos, and C. Jentsch, "Narrative shift detection: A hybrid approach of dynamic topic models and large language models", *arXiv Preprint*, arXiv:2506.20269, 2025.

[20] C. H. Chang, J. T. Tsai, Y. H. Tsai, and S. Y. Hwang, "LITA: An efficient LLM-assisted iterative topic augmentation framework", In: *Lecture Notes in Computer Science*, pp. 449-460, 2025.

[21] D. M. Blei, A. Y. Ng, and M. I. Jordan, "Latent dirichlet allocation", *J. Mach. Learn. Res.*, Vol. 3, pp. 993-1022, 2003.

[22] D. D. Lee, and H. S. Seung, "Learning the parts of objects by non-negative matrix factorization", *Nature*, Vol. 401, pp. 788-791, 1999.

[23] J. Devlin, M. W. Chang, K. Lee, and K. Toutanova, "BERT: Pre-training of deep bidirectional transformers for language understanding", In: *Proc. Conf. North Amer. Chapter Assoc. Comput. Linguistics (NAACL-HLT)*, pp. 4171-4186, 2019.

[24] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. Kaiser, and I. Polosukhin, "Attention is all you need", In: *Proc. of the 31st International Conference on Neural Information Processing Systems (NIPS'17)*, pp. 6000-6010, 2017.

[25] D. M. Blei, and J. D. Lafferty, "Dynamic topic models", In: *Proc. 23rd Int. Conf. on Machine Learning (ICML)*, pp. 113-120, 2006.

[26] L. McInnes, J. Healy, N. Saul, and L. Großberger, "UMAP: Uniform manifold approximation and projection", *J. Open Source Softw.*, Vol. 3, No. 29, p. 861, 2018.

[27] R. J. G. B. Campello, D. Moulavi, and J. Sander, "Density-based clustering based on hierarchical density estimates", In: *Proc. Pacific-Asia Conf. on Knowledge Discovery and Data Mining (PAKDD)*, pp. 160–172, 2013.

[28] M. Grootendorst, "BERTopic: Neural topic modeling with a class-based TF-IDF procedure", *arXiv Preprint*, arXiv:2203.05794, 2022.

[29] S. Terragni, E. Fersini, B. G. Galuzzi, P. Tropeano, and A. Candelieri, "OCTIS: Comparing and optimizing topic models is simple!", In: *Proc. of 16th Conf. of the European Chapter of the Assoc. for Computational Linguistics: System Demonstrations*, pp. 263-270, 2021,

[30] M. Röder, A. Both, and A. Hinneburg, "Exploring the space of topic coherence measures", In: *Proc. of 8th ACM Int. Conf. on Web Search and Data Mining (WSDM)*, pp. 399-408, 2015.

[31] J. H. Lau, D. Newman, and T. Baldwin, "Machine reading tea leaves: Automatically evaluating topic coherence and topic model quality", In: *Proc. of 14th Conf. of the European Chapter of the Assoc. for Computational Linguistics*, pp. 530–539, 2014.

[32] P. J. Rousseeuw, "Silhouettes: A graphical aid to the interpretation and validation of cluster analysis", *Journal of Computational and Applied Mathematics*, Vol. 20, pp. 53-65, 1987.

[33] L. Hubert, and P. Arabie, "Comparing partitions", *Journal of Classification*, Vol. 2, pp. 193-218, 1985.

[34] T. Zhang, V. Kishore, F. Wu, K. Q. Weinberger, and Y. Artzi, "BERTScore: Evaluating text generation with BERT", *arXiv Preprint*, arXiv:1904.09675, 2019.

[35] X. Wu, Y. Han, and F. Yang, "Analyzing CASIS policy data with AI: Sentiment trends and topic modeling", *Springer Science and Business Media LLC*, Nov., Preprint (Version 1), 2024.

[36] Z. Zhang, M. Fang, L. Chen, and M. R. Namazi Rad, "Is neural topic modelling better than clustering? An empirical study on clustering with contextual embeddings for topics", In: *Proc. of 2022 Conf. North Amer. Chapter Assoc. Comput. Linguistics (NAACL-HLT)*, 2022.

[37] A. Páez, "Understanding with toy surrogate models in machine learning", *Minds & Machines*, Vol. 34, Art. no. 4, 2024.

[38] A. Piper, and S. Wu, "Evaluating Large Language Models for Narrative Topic Labeling", In: *Proc. of the 5th International Conference on Natural Language Processing*

This article is licensed under a Creative Commons Attribution-ShareAlike 4.0 International License.
License details: https://creativecommons.org/licenses/by-sa/4.0/

for Digital Humanities. Association for Computational Linguistics, pp. 281-291, 2025.

[39] T. Schmidt, K. R. Lange, M. Reccius, H. Müller, M. Roos, and C. Jentsch, "Identifying economic narratives in large text corpora – An integrated approach using Large Language Models", *arXiv Preprint*, arXiv:2506.15041, 2025.

[40] S. Jenner, D. Raidos, E. Anderson, S. Fleetwood, B. Ainsworth, K. Fox, J. Kreppner, and M. Barker, "Using large language models for narrative analysis: a novel application of generative AI", *Methods in Psychology*, Vol. 12, pp. 100183, 2025.

[41] R. Moraffah, M. Karami, R. Guo, A. Raglin, and H. Liu, "Causal Interpretability for Machine Learning – Problems", *Methods and Evaluation. SIGKDD Explor. Newsl.*, Vol. 22, No. 1, pp. 18-33, 2020.

[42] M. Mersha, K. Lam, J. Wood, A. K. AlShami, and J. Kalita, "Explainable artificial intelligence: A survey of needs, techniques, applications, and future direction", *Neurocomputing*, Vol. 599, Art. no. 128111, 2024,

[43] Z. F. Hu, T. Kuflik, I. G. Mocanu, S. Najafian, and A. Shulner Tal, "Recent studies of XAI – review", In: *Adjunct Proc. of 29th ACM Conf. User Modeling, Adaptation and Personalization (UMAP)*, pp. 421–431, 2021.

[44] E. Albini, S. Sharma, S. Mishra, D. Dervovic, and D. Magazzeni, "On the connection between game-theoretic feature attributions and counterfactual explanations", In: *Proc. of 2023 AAAI/ACM Conf. on AI, Ethics, and Society (AIES)*, pp. 411–431, 2023.

[45] J. Del Ser, A. Barredo-Arrieta, N. Díaz-Rodríguez, F. Herrera, A. Saranti, and A. Holzinger, "On generating trustworthy counterfactual explanations", *Information Sciences*, Vol. 655, Art. no. 119898, 2024.

This article is licensed under a Creative Commons Attribution-ShareAlike 4.0 International License.
License details: https://creativecommons.org/licenses/by-sa/4.0/