DVI-DTM: Dual-View Representation Learning for Interpretable Short Text Dynamic Topic Modeling

Di Liu[1]* **Zheng Fang**[2]* **Bin Wu**[1]†

[1]Beijing Key Laboratory of Intelligent Telecommunication Software and Multimedia, Beijing University of Posts and Telecommunications, China
[2]School of Artificial Intelligence, University of Chinese Academy of Sciences, China
liudi@bupt.edu.cn fangzheng20@mails.ucas.ac.cn wubin@bupt.edu.cn

Abstract

Dynamic topic modeling aims to capture topic evolution from temporal text corpora. However, existing methods face two major challenges when applied to short texts: semantic ambiguity and interpretation ambiguity. Semantic ambiguity arises from the sparsity of short texts and the neglect of temporal semantic shifts. Interpretation ambiguity refers to the latent topics that lack human-understandable descriptions. In this work, we propose a novel **D**ual-**V**iew representation learning-based **I**nterpretable short text **D**ynamic **T**opic **M**odel (**DVI-DTM**). To address semantic ambiguity, the Dual-View Representation Learning module is presented to learn robust document-topic distributions by aligning temporal-aware term view and sentence view representations of short texts. To tackle interpretation ambiguity, we introduce a GEA Topic Refiner that leverages LLM agents to generate topic descriptions and refine document-topic distributions through collaborative semantic reasoning. Furthermore, a Dual-Factor Ranking module is designed to capture the topic evolution through semantic relevance and temporal uniqueness. Comprehensive experiments demonstrate that DVI-DTM outperforms the state-of-the-art baselines in topic alignment and dynamic topic quality metrics while producing highly interpretable topic descriptions.

1 Introduction

Dynamic topic modeling focuses on the unsupervised discovery of latent topics in temporal text corpora, tracking their semantic evolution over time. Short texts such as news headlines, tweets, forum posts, and paper titles contain rich but highly fragmented information about events, trends, and public discourse. Dynamic topic modeling over short texts offers substantial practical and societal value in applications such as trend detection (Lu and Yang, 2012), social media monitoring (Inaam ul haq et al., 2023), and knowledge discovery in scientific (Zhou et al., 2018; Zeng et al., 2019; Liu et al., 2021; Li et al., 2023) and biomedical research (Sharma et al., 2022; Lee et al., 2025).

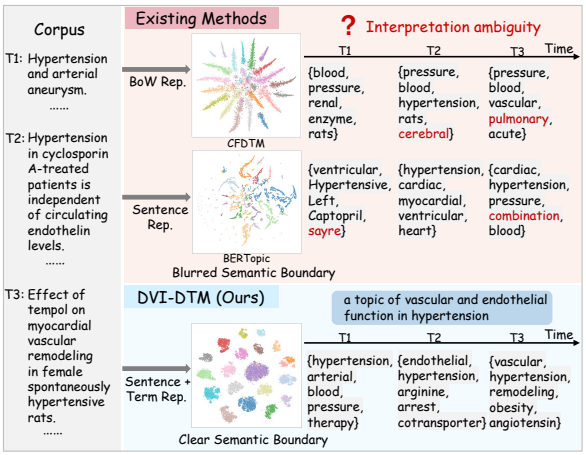

Figure 1: t-SNE visualizations of document-topic distributions and the top-5 terms on the Biomedical dataset. While existing methods lead to topic distributions with blurred semantic boundaries, irrelevant topic terms, and limited interpretability, our DVI-DTM produces well-separated topic distributions, coherent terms, and interpretable topic semantics. "*Rep.*": Representation.

Traditional dynamic topic modeling methods (Blei and Lafferty, 2006) are built primarily on probabilistic frameworks (Blei et al., 2003), introducing temporal dependencies to capture topic evolution. Recently, neural dynamic topic models have been popular, leveraging the VAE framework (Kingma and Welling, 2013) to reconstruct Bag-of-Words (BoW) representations (Dieng et al., 2019) and incorporating mechanisms such as self-attention (Miyamoto et al., 2023) and contrastive learning (Wu et al., 2024) to model document-topic (doc-topic) distributions. Clustering-based approaches (Grootendorst, 2022; Rahimi et al.,

*Equal contribution.
†Corresponding author.

Proceedings of the 64th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), pages 28576–28593

2024) model topic dynamics by grouping sentence-level embeddings over time, providing an alternative perspective for dynamic topic discovery.

Although these methods have achieved promising results on general documents, they still face certain limitations when applied to short texts. As illustrated in Figure 1, these limitations include: (1) **Semantic ambiguity**: Most methods (Blei and Lafferty, 2006; Dieng et al., 2019; Miyamoto et al., 2023; Wu et al., 2024) rely on BoW representations to infer topic evolution. However, short texts are inherently sparse and fragmented, offering limited word co-occurrence and contextual information. Short texts also contain numerous informal or filler words, and treating all tokens equally during sentence embedding extraction (Grootendorst, 2022; Rahimi et al., 2024) tends to dilute the representation of key terms, resulting in blurred semantic focus of topic embeddings in the clustering space. In addition, most neural methods adopt static term embeddings, which fail to capture semantic shifts of terms over time, leading to term semantic ambiguity. Consequently, the learned doc-topic distributions often exhibit blurred semantic boundaries. (2) **Interpretation ambiguity**: Existing methods typically output evolving sets of topic terms without explicit semantic summaries, requiring users to manually interpret and compare topics across time. While recent works (Doi et al., 2024; Nguyen et al., 2025; Vu et al., 2025) have attempted to address short text sparsity in static topic modeling, effective and interpretable dynamic topic modeling for short texts remains largely underexplored.

To address the above issues, we propose **DVI-DTM**, a novel dynamic topic modeling framework tailored for short texts that leverages dual-view representation information and the semantic reasoning capabilities of Large Language Models (LLMs) to achieve robust and interpretable dynamic topic discovery. To tackle the semantic ambiguity issue, we design a **Dual-View Representation learning (DVR)** module that simultaneously constructs short text embeddings and doc-topic distributions from both the sentence view and the term view. DVR introduces mutual information-based alignment and consistency constraints on both the text representations and the doc-topic distributions across the two views, encouraging the integration of fine-grained term semantics with global sentence-level context. This design effectively mitigates semantic ambiguity arising from the sparsity of short texts. Additionally, we introduce a temporal-aware term embedding extractor that captures dynamically evolving term embeddings over time. To address the issue of interpretation ambiguity, we design a **GEA (*Generator-Evaluator-Assigner*) Topic Refiner** that employs three cooperative LLM agents to generate interpretable topic descriptions and refine doc-topic distributions. Finally, a **Dual-Factor Ranking (DFR) module** is employed to extract topic evolutions by jointly considering topic semantic relevance and temporal uniqueness. To summarize, our main contributions are as follows:

- We identify two critical challenges in the short text dynamic topic modeling task, including semantic ambiguity and interpretation ambiguity, and propose a novel method, DVI-DTM, to address them effectively.
- We propose a DVR module that aligns fine-grained term semantics with global sentence contextual information, greatly mitigating semantic ambiguity due to short text sparsity.
- We design a GEA Topic Refiner to generate interpretable topic descriptions for the discovered dynamic topics and a DFR module to extract temporally coherent topic evolutions.
- Extensive experiments conducted on three widely used real-world short text datasets demonstrate the effectiveness and superior performance of our proposed DVI-DTM compared to the state-of-the-art methods.

2 Related Work

Dynamic Topic Modeling Dynamic topic modeling can be broadly categorized into Bag-of-Words (BoW) and clustering-based approaches. Within the BoW framework, the seminal DTM (Blei and Lafferty, 2006) extends LDA (Blei et al., 2003) to model temporal dependencies via state-space representations. Although subsequent studies (Hori et al., 2018; Li et al., 2019) optimize DTM using variational inference (Jähnichen et al., 2018) or Gibbs sampling (Acharya et al., 2018), these probabilistic methods suffer from scalability bottlenecks due to complex joint inference. To address this, neural approaches such as DETM (Dieng et al., 2019) employ VAEs (Kingma and Welling, 2013) to model topic evolution via embedding-based inner products. Building on this, the following methods (Zhang and Lauw, 2022a; Cvejoski et al., 2023) propose improvements. DSNTM (Miyamoto et al.,

2023) learns topic dependencies via self-attention, whereas CFDTM (Wu et al., 2024) recently introduced a contrastive learning strategy to model dynamic topics. Alternatively, clustering-based approaches (Grootendorst, 2022; Rahimi et al., 2024) derive dynamic topics directly from PLM embeddings, employing techniques like class-based TF-IDF (Grootendorst, 2022) or implicit distribution strategies (Rahimi et al., 2024). Despite these advancements, modeling evolution in short texts remains an open challenge.

Static Topic Modeling for Short Texts In static topic modeling, several prominent methods (Shi et al., 2018; Zhang and Lauw, 2022b; Doi et al., 2024; Vu et al., 2025; Nguyen et al., 2025; Ennajari et al., 2025) have been explicitly proposed for short texts. kNNTM (Lin et al., 2024) pioneered kNN-based aggregation to mitigate sparsity in short texts. Advancing this, GloCOM (Nguyen et al., 2025) utilizes PLM-driven global context clustering to enrich word co-occurrence patterns. To address the lack of explicit regularization in this framework, EnCOT (Vu et al., 2025) introduces an optimal transport-based enhanced global clustering method. The recent success of large language models (LLMs) in various text analysis tasks (Zhong et al., 2024; Zhou et al., 2025) has inspired new attempts to apply them to topic modeling (Stammbach et al., 2023; Pham et al., 2024; Doi et al., 2024; Liu et al., 2025; Yang et al., 2025). However, these methods overlook temporal dynamics, failing to address topic evolution under short-text sparsity.

3 Methodology

In this work, we propose DVI-DTM, a novel framework that enables robust and interpretable dynamic topic modeling for short texts. The overall architecture of DVI-DTM is illustrated in Figure 2. In this section, we first introduce the problem statement and notations. Subsequently, we detail our method.

3.1 Problem Statement and Notations

Given a timestamped corpus $\mathcal{D} = \{d_n, t_n\}_{n=1}^{N}$ consisting of N short texts, where each d_n denotes a sentence-level document and $t_n \in \{1, \ldots, T\}$ indicates its corresponding time slice. Short text dynamic topic modeling aims to mine K latent topics within the corpus and capture their temporal evolution. Unlike static topic modeling, this requires extracting representative terms from sequential time slices to characterize the semantic evolution of each topic. Another more challenging goal is to ensure the interpretability of discovered topics. Beyond merely outputting evolving topic term trajectories, the latent topics should be explained with human-understandable natural language descriptions, enabling users to grasp their semantic meaning intuitively.

In this paper, we define the term vocabulary as $\mathcal{W} = \{w_1, \ldots, w_W\}$, consisting of W unique terms. This vocabulary is constructed from the raw corpus by employing AutoPhrase(Shang et al., 2018) and removing stop words as well as low-frequency terms.

3.2 Dual-View Representation Learning

To address the challenge of semantic ambiguity prevalent in existing methods, we propose the dual-view representation learning module, as shown in Figure 2 (a). This module aligns term view and sentence view representations to enrich semantics.

Dual-View Representation Given a timestamped short text corpus $\mathcal{D}$, we encode it from two views and construct the doc-topic distributions. For the sentence view, we employ a trainable Pre-trained Language Model (PLM) to encode the entire document sentence, producing global sentence view representations $\mathbf{X} \in \mathbb{R}^{N \times C}$. For the term view, we design a temporal-aware term embedding extractor to capture semantic variations of terms across different time slices. Specifically, the vocabulary $\mathcal{W}$ is first fed into a frozen PLM to obtain static term embeddings, which are then replicated T times to form initial dynamic term embeddings $\mathcal{V}_{init} \in \mathbb{R}^{T \times W \times C}$. $\mathcal{V}_{init}$ is next passed into a Temporal Term Encoder (TTE) built upon Multi-Head self-Attention (MHA) (Vaswani et al., 2017) with learnable term positional encoding $\varphi \in \mathbb{R}^{W \times C}$ and time positional encoding $\psi \in \mathbb{R}^{T \times C}$ to encode temporal-aware term representations:

$$\mathcal{V} = MHA(\mathcal{V}_{init} + \varphi + \psi). \quad (1)$$

For each document d_n associated with a timestamp t_n, we retrieve the temporal embeddings for its constituent terms from $\mathcal{V}$. The document representation for the d_n is constructed as follows:

$$\mathbf{Y}_n = \text{Concat}(\mathcal{V}[t_n, idx_1], \ldots, \mathcal{V}[t_n, idx_l]), \quad (2)$$

where idx denotes the index of terms in $\mathcal{V}$, l denotes the term length of d_n. To handle varying lengths, we apply zero-padding to extend each $\mathbf{Y}_n$

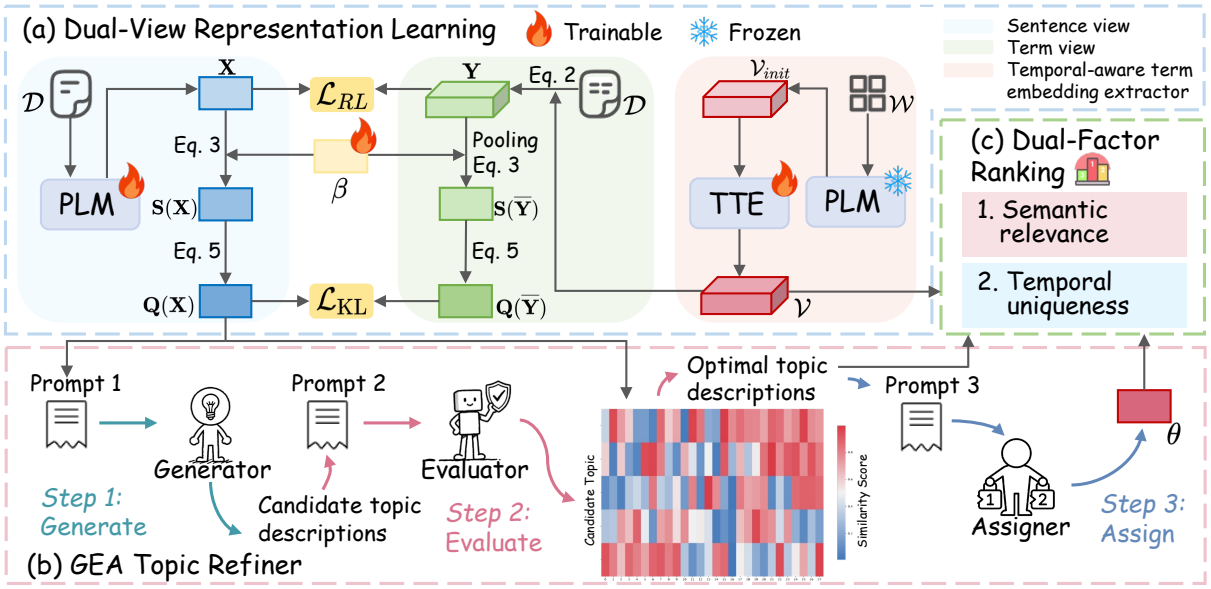

Figure 2: The framework of DVI-DTM. It consists of three components: (a) Dual-View Representation learning (DVR) module, (b) GEA Topic Refiner, and (c) Dual-Factor Ranking (DFR) module.

to the maximum length L, resulting in the batched tensor $\mathbf{Y} \in \mathbb{R}^{N\times L\times C}$. We perform masked mean pooling on $\mathbf{Y}$ to obtain the aggregated sequence representations, denoted as $\overline{\mathbf{Y}} \in \mathbb{R}^{N\times C}$.

To construct the doc-topic distributions, we encode the texts using a frozen PLM and then apply K-means clustering (MCQUEEN, 1967) to initialize the topic embeddings $\beta \in \mathbb{R}^{K\times C}$, which are added as trainable parameters within the model. To associate and align the documents with latent topics, we compute the cosine similarity scores $\mathbf{S} \in \mathbb{R}^{N\times K}$ between each view representation $v \in \{\mathbf{X}, \overline{\mathbf{Y}}\}$ and the topic embeddings β:

$$\mathbf{S}(v) = \text{softmax}\left(\cos(v, \beta)\right). \tag{3}$$

For each text d_n, we infer its topic assignment $\mathbf{K_n}$ based on $\mathbf{S}_n(v)$. To ensure temporal smoothness, the average similarity score of texts assigned to the same topic at time $t_n - 1$ is weighted and aggregated with the current similarity score of d_n at time t_n, yielding the final doc-topic distributions.

$$\mathbf{K}_n(v) = \arg\max_k(\mathbf{S}_{n,k}(v)), \tag{4}$$

$$\mathbf{Q}_n^{t_n}(v) = \lambda \mathbf{S}_n^{t_n}(v) + (1-\lambda)\frac{\sum_{m\in\mathcal{N}_{\mathbf{K}_n}^{t_n-1}} \mathbf{S}_m^{t_n-1}(v)}{|\mathcal{N}_{\mathbf{K}_n}^{t_n-1}|}, \tag{5}$$

where $\lambda \in [0, 1]$ is the weighting coefficient, and $\mathcal{N}_{\mathbf{K}_n}^{t_n-1}$ denotes the set of texts in the neighboring time slice $t_n - 1$ that are assigned to the topic $\mathbf{K}_n$.

Training Objectives Our DVR module is optimized with two objectives at the embedding representation level and the doc-topic distribution level. The first is the Representation Learning loss $\mathcal{L}_{\text{RL}}$. For each text, its sentence view representation and its corresponding term view representation form a positive pair, while representations from other texts are regarded as negative samples. This objective maximizes the average Mutual Information (MI) between positive pairs and minimizes MI between negative pairs, reinforcing cross-view semantic alignment and learning semantically enriched short text representations.

$$\mathcal{L}_{RL} = -\frac{1}{N}\left(\sum_{n=1}^{N} \mathcal{I}^{JSD}(\mathbf{X}_n; \mathbf{Y}_n)\right). \tag{6}$$

Following (Nowozin et al., 2016; Hjelm et al., 2019; Kamthawee et al., 2024), we use a Jensen-Shannon MI estimator to estimate a lower bound of MI:

$$\mathcal{I}^{JSD}(\mathbf{X}_n; \mathbf{Y}_n) := \mathbb{E}_{\mathbb{P}}[-sp(-\mathbf{X}_n \cdot \mathbf{Y}_n)] - \mathbb{E}_{\mathbb{P}\times\tilde{\mathbb{P}}}[sp(\mathbf{X}_n \cdot \tilde{\mathbf{Y}}_n)], \tag{7}$$

where $\tilde{\mathbf{Y}}_n$ is a negative input sampled from distribution $\tilde{\mathbb{P}} = \mathbb{P}$, and $sp(z) = log(1 + e^z)$ is the softplus function.

The second objective explicitly aligns the two views by minimizing the KL divergence between their doc-topic distributions, ensuring that both the

local term semantics and the global sentence context contribute coherently to the same latent topic representation:

$$\mathcal{L}_{\mathrm{KL}} = \mathrm{KL}(\mathbf{Q}(\mathbf{X}) \parallel \mathbf{Q}(\overline{\mathbf{Y}})). \tag{8}$$

The total training loss is calculated as:

$$\mathcal{L} = \mathcal{L}_{RL} + \mathcal{L}_{KL}. \tag{9}$$

3.3 GEA Topic Refiner

To overcome interpretation ambiguity, we propose a GEA (*Generator-Evaluator-Assigner*) Topic Refiner that consists of three cooperative LLM agents inspired by (Zhong et al., 2024), as shown in Figure 2 (b). While the DVR module learns robust and temporally coherent latent topic representations, these topics still lack explicit human-readable descriptions. This module bridges this gap by transforming latent topics into interpretable natural language descriptions and using them to refine the doc-topic distributions.

Generator The *Generator* produces candidate topic descriptions based on the doc-topic distributions $\mathbf{Q}(\mathbf{X})$ learned in the dual-view module. Specifically, we normalize $\mathbf{Q}(\mathbf{X})$ to obtain topic similarity scores and partition the corpus into three time periods to capture dynamic topic evolution. For each topic, we select the top A most relevant texts from each period, and randomly sample texts from the remaining corpus to increase lexical diversity and reduce sampling bias. The selected texts and their similarity scores are organized into a structured prompt to guide the *Generator* to produce E candidate descriptions for each topic.

Evaluator Although LLMs possess strong language generation capabilities, the *Generator* may occasionally produce hallucinated or irrelevant topic descriptions (Ji et al., 2023). To avoid the impact of such hallucinations, we design an explicit evaluation and verification mechanism to assess the most appropriate topic description. We randomly sample F texts from the corpus and prompt the *Evaluator* to judge whether each text falls within the semantic scope of a candidate description. The resulting binary decisions $\mathbf{U} \in \mathbb{R}^{F \times E}$ are compared with the corresponding doc-topic similarity scores, and the candidate description with the highest Pearson correlation (Benesty et al., 2009) is selected as the optimal description for each topic.

Assigner After obtaining the optimal topic descriptions, we employ them as semantic anchors to refine the doc-topic distribution. A structured prompt containing each text and all topic descriptions is provided to the *Assigner*, which assigns each text to its most semantically relevant topic and returns the corresponding index. This process yields a refined binary doc-topic distribution θ that is more interpretable and semantically faithful than the initial distribution, thereby further improving the reliability and clarity of topic assignments. Detailed prompts are provided in Appendix A.

3.4 Dual-Factor Ranking

To identify representative terms that characterize the evolution of each dynamic topic, we introduce a dual-factor ranking module that jointly considers semantic relevance and temporal uniqueness, inspired by (Balepur et al., 2023).

Semantic Relevance We define two semantic relevance scores to select topic terms that exhibit the strongest semantic association with the target topic. The first score is a topic description similarity score S_{des}, which measures the similarity between the temporal-aware term representations $\mathcal{V}$ and the optimal topic descriptions from GEA Topic Refiner:

$$S_{des} = softmax(\cos\left(\mathcal{V}^t, \mathbf{h}_k^t\right)), \tag{10}$$

where $\mathbf{h}_k^t$ denotes the average temporal-aware term embeddings of the terms that both appear in the k-th optimal topic description and term vocabulary $\mathcal{W}_t$ at time slice t.

The second score is a topic class similarity score S_{class}, which measures the similarity between the temporal-aware term representations and the average sentence view embedding $\mathbf{X}_k^t$ of texts assigned to the k-th topic at time t according to θ_k:

$$S_{class} = softmax(\cos\left(\mathcal{V}^t, \mathbf{X}_k^t\right)). \tag{11}$$

The semantic relevance factor ensures that the selected terms are semantically consistent with the topic description and representative of the documents associated with the topic, and it can be formulated as:

$$S_{SR} = \frac{S_{des} + S_{class}}{2}. \tag{12}$$

Temporal Uniqueness We introduce a temporal uniqueness score to identify topic terms that suddenly surge in prominence within a specific time

Table 1: Experimental results on different datasets. Higher means better for all metrics. The best values are marked in **bold** and the second values underlined. Symbol * indicates significant improvements over the baselines (except TopicGPT) through the T-test ($p \leq 0.05$). "-" denotes the metric is not applicable to the method.

Method	StackOverflow						NYT						Biomedical					
	Purity	NMI	NPMI	C_V	TD	TTC	Purity	NMI	NPMI	C_V	TD	TTC	Purity	NMI	NPMI	C_V	TD	TTC
BERTopic	0.377	0.364	0.055	0.372	0.886	0.273	0.509	0.304	0.108	0.420	0.939	0.265	0.265	0.257	0.045	0.394	0.946	0.115
ANTM	-	-	-0.021	0.499	0.831	0.363	-	-	0.037	0.501	0.722	0.411	-	-	-0.009	0.432	0.693	0.327
DTM	0.225	0.104	-0.135	0.459	0.846	0.385	0.351	0.101	-0.023	0.656	0.908	0.607	0.107	0.093	-0.029	0.518	0.694	0.454
DETM	0.267	0.210	0.245	0.561	0.727	0.444	0.347	0.119	0.121	0.547	0.932	0.484	0.176	0.124	0.070	0.434	0.732	0.332
DSNTM	0.253	0.126	0.011	0.595	0.702	0.541	0.345	0.103	0.069	0.651	0.529	0.590	0.102	0.092	-0.015	0.347	0.501	0.305
CFDTM	0.349	0.284	-0.069	0.418	0.982	0.364	0.463	0.269	0.066	0.579	0.902	0.521	0.297	0.246	0.012	0.501	0.766	0.438
TopicGPT (GPT-4o)	0.475	0.432	-	-	-	-	0.456	0.246	-	-	-	-	0.301	0.283	-	-	-	-
ECRTM+LLM-ITL	0.316	0.231	-	-	-	-	0.395	0.199	-	-	-	-	0.275	0.218	-	-	-	-
KNNTM	0.443	0.422	-	-	-	-	0.626	0.347	-	-	-	-	0.419	0.362	-	-	-	-
GloCOM-EnCOT	0.482	0.451	-	-	-	-	0.657	0.381	-	-	-	-	0.456	0.397	-	-	-	-
Ours (Qwen3-max)	0.538	0.474	0.280	0.618	**1.000**	0.580	0.687	0.409	0.145	0.672	0.990	0.624	0.504	0.407	0.081	0.668	0.989	0.635
Ours (Deepseek-R1)	0.534	0.470	0.282	0.617	0.996	0.578	0.697	0.418	0.145	0.673	0.995	0.625	0.507	0.413	0.084	0.670	0.991	0.637
Ours (GPT-4o)	0.531	0.468	0.278	0.612	0.995	0.571	0.685	0.408	0.139	0.671	0.989	0.622	0.498	0.406	0.079	0.667	0.984	0.632
Ours (GPT-5)	**0.545***	**0.482***	**0.290***	**0.620***	**1.000***	**0.581***	**0.704***	**0.424***	**0.149***	**0.680***	**0.996***	**0.632***	**0.511***	**0.418***	**0.088***	**0.678***	**0.994***	**0.644***

slice. Such abrupt changes in frequency often signal emerging events or shifts in attention relevant to evolving topics. The temporal uniqueness score is computed based on the ratio of the frequency $F(t, w)$ of the term w in the current time slice t to its frequency $F(t-1, w)$ in the previous time slice $t-1$, highlighting terms that are temporally distinctive and evolutionarily informative:

$$S_{TU} = softmax(log\frac{F(t,w)+1}{F(t-1,w)+1}). \quad (13)$$

Finally, our overall score can be formulated as:

$$S_{all} = \alpha \cdot S_{SR} + (1-\alpha) \cdot S_{TU}, \quad (14)$$

where α is a weighting parameter that balances the contribution of semantic relevance and temporal uniqueness. For each topic at time slice t, we select the highest-ranked terms and add them to the dynamic topic term list.

4 Experiments

4.1 Experimental Settings

Datasets The experiments are performed on three widely used real-world short text datasets: StackOverflow, NYT, and Biomedical. **StackOverflow** contains 19,796 question titles from 2008 to 2012. **NYT** consists of 24,952 news headlines from 2001 to 2024. **Biomedical** includes 20,463 paper titles from 1980 to 2013. Please refer to Appendix B for further details about the datasets.

Evaluation Metrics For topic alignment, we conduct text clustering to evaluate the quality of the doc-topic distributions, evaluated by **Purity** and Normalized Mutual Information (**NMI**) (Schütze et al., 2008), following (Wu et al., 2024; Vu et al., 2025). For dynamic topic quality, we consider four metrics: **NPMI** (Lau et al., 2014) and $\boldsymbol{C_V}$ (Röder et al., 2015) assess topic coherence based on pairwise co-occurrence and overall semantic consistency, respectively. Topic Diversity (**TD**) measures the uniqueness of topic words within a time slice, while Temporal Topic Coherence (**TTC**) (James et al., 2024) captures semantic stability across consecutive timestamps. We select the top 10 terms from each topic to calculate these metrics. Metrics are detailed in Appendix C.

Baselines We compare our model against representative baselines from four paradigms. For clustering-based dynamic topic models, we contain BERTopic (Grootendorst, 2022) and ANTM (Rahimi et al., 2024). For BoW-based dynamic topic models, we include: DTM (Blei and Lafferty, 2006), DETM (Dieng et al., 2019), DSNTM (Miyamoto et al., 2023) and CFDTM (Wu et al., 2024). For short-text topic modeling, we compare kNNTM (Lin et al., 2024) and GloCOM-EnCOT (Nguyen et al., 2025; Vu et al., 2025). For LLM-based topic modeling, we consider TopicGPT (Pham et al., 2024) and ECRTM+LLM-ITL (Wu et al., 2023; Yang et al., 2025). Since static methods lack temporal modeling capabilities, we report their results only for the topic alignment evaluation. Furthermore, neither ANTM nor TopicGPT supports specifying the number of topics, and ANTM cannot infer doc-topic distributions.

Table 2: Ablation Study. Symbol * indicates significant improvements of DVI-DTM through the T-test ($p \leq 0.05$).

#	DVR			GEA Topic Refiner			DFR			Purity	NMI	NPMI	C_V	TD	TTC
	$\mathcal{L}_{RL}$	$\mathcal{L}_{KL}$	TTE	Generator	Evaluator	Assigner	S_{des}	S_{class}	S_{TU}						
(a)	✗	✗	✗	✓	✓	✓	✓	✓	✓	0.349	0.310	0.212	0.463	0.937	0.404
(b)	✓	✗	✗	✓	✓	✓	✓	✓	✓	0.460	0.420	0.254	0.585	0.966	0.542
(c)	✓	✓	✗	✓	✓	✓	✓	✓	✓	0.503	0.445	0.269	0.606	0.981	0.565
(d)	✓	✓	✓	✗	✗	✗	✗	✓	✓	0.521	0.462	0.241	0.589	0.987	0.553
(e)	✓	✓	✓	✓	✗	✓	✓	✓	✓	0.534	0.471	0.284	0.612	0.993	0.571
(f)	✓	✓	✓	✓	✓	✓	✗	✓	✓	0.545	0.482	0.246	0.594	0.991	0.561
(g)	✓	✓	✓	✓	✓	✓	✓	✗	✓	0.545	0.482	0.244	0.588	0.961	0.545
(h)	✓	✓	✓	✓	✓	✓	✓	✓	✗	0.545	0.482	0.271	0.609	0.987	0.567
(i)	✓	✓	✓	✓	✓	✓	✓	✓	✓	**0.545***	**0.482***	**0.290***	**0.620***	**1.000***	**0.581***

Implementation Details The *all-mpnet-base-v2* SBERT model (Reimers and Gurevych, 2019) is adopted as the PLM within the DVR module. Following (Wu et al., 2024; Yang et al., 2025), we set the number of topics to 20 and report averages across five runs, excluding ANTM and TopicGPT. Detailed hyperparameter settings (such as λ and α), configurations of all LLM agents, and a cost analysis are provided in Appendix D.

4.2 Performance Comparison

We quantitatively compare our proposed DVI-DTM with baselines across three datasets, as reported in Table 1. Overall, our DVI-DTM achieves significant improvements over all baselines across all metrics. In terms of the topic alignment, static baselines driven by LLMs (TopicGPT and LLM-ITL) or specifically designed for short texts (kNN-TM and EnCOT) generally outperform traditional dynamic topic modeling baselines (BERTopic, ANTM, DTM, DETM, DSNTM, and CFDTM). Nevertheless, our method achieves average improvements of 5.5% in Purity and 3.2% in NMI, highlighting its stronger ability to capture short text semantics. Regarding dynamic topic quality, our approach consistently outperforms all dynamic topic modeling baselines by a significant margin. Specifically, DVI-DTM still achieves average gains of 3.0%, 7.0%, 4.1%, and 8.5% on NPMI, C_V, TD, and TTC, respectively. These results demonstrate that DVI-DTM not only produces more discriminative doc-topic distributions but also has the strong ability to capture temporally coherent and diverse topic evolution. We further evaluate four variants of DVI-DTM instantiated with different LLMs. The GPT-5-based version achieves the best overall performance, while the remaining variants consistently outperform all existing baselines.

4.3 Ablation Study

To evaluate the contribution of each module, we conduct extensive ablation studies on the StackOverflow dataset. As shown in Table 2, removing the whole DVR module results in substantial drops in both Purity and NMI (Table 2 (a)). Introducing the document embedding-level representation learning objective $\mathcal{L}_{RL}$ alone improves Purity by 11.1 points (Table 2 (b)), demonstrating its effectiveness in alleviating semantic ambiguity in short text representations. Further incorporating the doc-topic distribution alignment objective $\mathcal{L}_{KL}$ (Table 2 (c)) and the TTE module (Table 2 (i)) leads to additional Purity gains of 4.3 and 4.2 points, respectively, validating the roles of cross-view topic consistency and temporal-aware term modeling in enhancing doc-topic distribution quality. We also examine the impact of the GEA Topic Refiner. Removing this module causes a noticeable performance degradation (Table 2 (d)), and disabling only the *Evaluator* (Table 2 (e)) also degrades the results, indicating that the *Evaluator* is crucial for mitigating hallucinated topic descriptions and refining doc-topic distributions. Finally, we analyze the DFR module by removing each scoring factor (Table 2 (f-h)). Removing either S_{des} or S_{class} results in a decline in topic coherence, confirming that both interpretable topic descriptions and document class semantics are essential for reliable topic term extraction. This also highlights the synergistic interaction between the dual-view representations and the topic descriptions. Excluding S_{TU} also degrades performance, demonstrating its importance in capturing emerging and time-specific topic terms.

We further analyze the **number of topics** and **hyperparameters** of DVI-DTM in Appendix E.

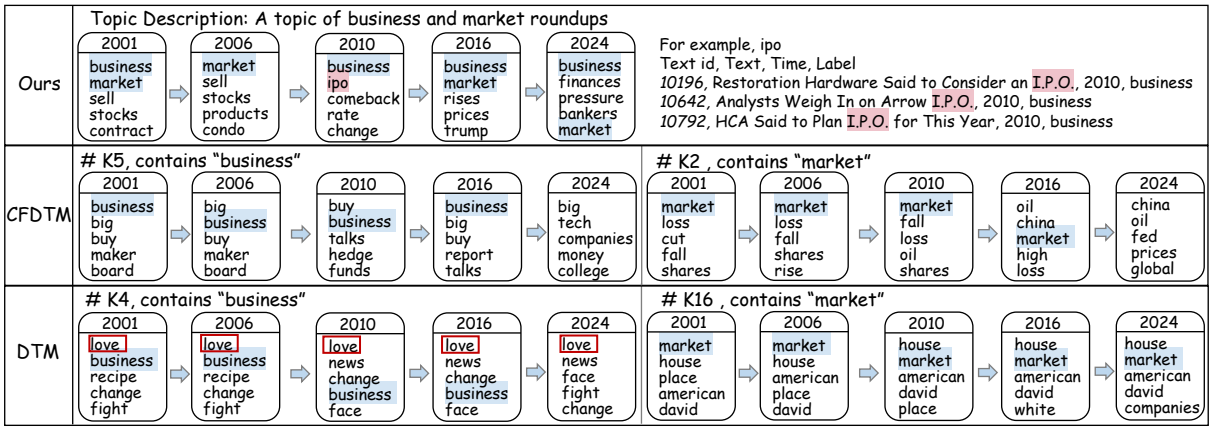

Figure 3: Evolution of Discovered Topics. Top 5 representative terms of the *business and market roundups* topic across different time slices in the NYT dataset, as extracted by Our DVI-DTM, CFDTM, and DTM.

Table 3: Alignment of discovered interpretable topic descriptions with the true labels on the NYT dataset.

Label	#Sample	Model-Generated Topic Descriptions
business	6,364	a topic of business and market roundups a topic of bank regulation and bailouts a topic of mergers and acquisitions a topic of corporate restructuring and strategic pivots a topic of white-collar crime and corporate misconduct
arts	2,501	a topic of visual arts and museum exhibitions
dining	2,098	a topic of dining guides and restaurant features a topic of food features
real estate	1,482	a topic of real estate spotlights
science	2,262	a topic of wildlife and animal behavior a topic of infectious disease outbreaks and vaccination
society	2,031	a topic of personal love stories tied to weddings a topic of digital culture and online platforms
sports	4,176	a topic of MLB roster moves and trade decisions a topic of New York City pro basketball a topic of sports opinion columns
styles	2,061	a topic of fashion industry culture and commentary a topic of fashion and style culture
travel	1,977	a topic of travel destination guides and itineraries a topic of contemplative travel and place

4.4 Interpretability and Evolution of Topics

Table 3 presents the 20 latent topics discovered by our model on the NYT dataset. The generated topic descriptions are highly consistent with the corresponding ground-truth labels, demonstrating that DVI-DTM is able to produce semantically meaningful and human-interpretable topic representations. Notably, categories with a larger number of documents are automatically decomposed into multiple fine-grained topics. This observation indicates that DVI-DTM can adaptively adjust topic granularity based on data distribution.

Figure 3 presents the dynamic evolution of topics in the NYT dataset. Since our method can generate a human-understandable semantic description,

Figure 4: Visualization of term evolution in embedding space using t-SNE (Maaten and Hinton, 2008).

we can easily identify that the topic is related to business and market roundups. In contrast, for baselines, users must search numerous topics to find those containing terms such as "business" or "market". Critically, attributed to the DVR module and the temporal uniqueness score, our method successfully captures the mutation of "IPO" (Initial Public Offering) in 2010, which is missed by baselines. Additionally, our extracted terms exhibit high topic coherence at each time slice, whereas DTM includes irrelevant terms like "love".

4.5 Analysis of Term Evolution

To assess whether our temporal-aware term embeddings capture semantic shifts over time, we visualize the term embeddings of "business" and "IPO" in 2009 and 2010 using t-SNE. As shown in Figure 4, these two terms are far apart in 2009 but become notably closer in 2010, consistent with corpus statistics showing a fivefold increase of "IPO" in "business" labeled documents. This demonstrates that our term embeddings effectively capture temporal semantic evolution, which is essential for accurate dynamic topic modeling.

Table 4: Human evaluation results.

Method	Evolution	Semantic	Temporal
DTM	3.200	0.867	0.533
DETM	2.800	0.867	0.667
DSNTM	2.933	0.800	0.600
CFDTM	3.730	0.800	0.733
Ours	**4.070**	**0.933**	**0.800**

4.6 Human Evaluation

Given that automated metrics do not always align with human judgment (Chang et al., 2009; Hoyle et al., 2021), we conduct a human evaluation to assess the quality of dynamic topics. For each model, we randomly sample 15 dynamic topic term trajectories on the StackOverflow dataset for comparison. We design two evaluation tasks: Topic Evolution Rating and Dynamic Intrusion Detection. The former assesses the topic **evolution** coherence using a 5-point Likert scale. The latter comprises two evaluation dimensions: **semantic** outlier detection and **temporal** slice completion. More details about the evaluation protocol are provided in Appendix F. We recruit three graduate students to perform evaluations in a blind setup. The evaluation process yielded substantial inter-annotator agreement, evidenced by a Kendall's tau of 0.816 for the rating task and a Fleiss' kappa of 0.722 for the detection tasks. As shown in Table 4, our model consistently outperforms all baselines, demonstrating superior temporal coherence and semantic interpretability in short text dynamic topic modeling. This is consistent with the observations from our main results.

5 Conclusion

In this paper, we propose a Dual-View representation learning-based Interpretable Dynamic Topic Model (DVI-DTM) to resolve the semantic and interpretation ambiguities in short-text dynamic topic modeling. To address semantic ambiguity, we introduce a dual-view representation learning framework that captures dynamic term embeddings and mitigates the impact of semantic sparsity by aligning term view and sentence view representations. To overcome interpretation ambiguity, we design a GEA topic refiner that generates human-readable topic descriptions, enabling more transparent topic understanding. Furthermore, a dual-factor ranking module is incorporated to extract temporally coherent topic evolutions. Extensive experiments demonstrate the effectiveness of our DVI-DTM.

Limitations

While our DVI-DTM achieves strong performance and interpretability in short text dynamic topic modeling, it relies on the capability of LLMs in the GEA Topic Refiner. When smaller open-source LLMs (e.g., Llama-3-8B) are used as agents in the refiner, they may fail to generate discriminative topic descriptions and reduce their effectiveness in refining doc-topic distributions due to limited reasoning and language generation capacity compared to large-scale models (e.g., GPT-5). Even without the GEA topic refiner, DVI-DTM remains competitive compared with existing baselines by leveraging the DVR module to produce robust doc-topic distributions, as shown in Table 2 (d). Therefore, we believe this limitation is acceptable in practice. An important direction for future research is to explore knowledge distillation and parameter-efficient fine-tuning (PEFT) techniques to equip lightweight models with stronger semantic reasoning capabilities, thereby balancing effectiveness with deployment efficiency.

Ethical Considerations

We adhere to the ACL Code of Ethics and all relevant license terms. The dynamic topic terms shown in Figure 1 and Figure 3 are generated by models trained on the Biomedical and NYT datasets, respectively. As these datasets are collected from real-world sources, the generated outputs may reflect existing biases or dominant perspectives present in the dataset. The authors do not intend to endorse or promote any biased or sensitive content that may be implicitly reflected in the model outputs. We make every effort to objectively analyze and present the discovered topics. Any unintended biases reflected in the model outputs are considered part of the limitations of automatic dynamic topic modeling.

The human evaluation in this study was conducted with voluntary participants who provided informed consent. No personally identifiable information was collected, and the study posed no foreseeable risk to participants.

Acknowledgments

This work is supported by the National Natural Science Foundations of China under Grant (62372060).

References

Ayan Acharya, Joydeep Ghosh, and Mingyuan Zhou. 2018. A dual markov chain topic model for dynamic environments. In *Proceedings of the 24th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, pages 1099–1108.

Nishant Balepur, Shivam Agarwal, Karthik Venkat Ramanan, Susik Yoon, Diyi Yang, and Jiawei Han. 2023. Dynamite: Discovering explosive topic evolutions with user guidance. In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 194–217.

Jacob Benesty, Jingdong Chen, Yiteng Huang, and Israel Cohen. 2009. Pearson correlation coefficient. In *Noise reduction in speech processing*, pages 1–4. Springer.

David M Blei and John D Lafferty. 2006. Dynamic topic models. In *Proceedings of the 23rd international conference on Machine learning*, pages 113–120.

David M Blei, Andrew Y Ng, and Michael I Jordan. 2003. Latent dirichlet allocation. *Journal of machine Learning research*, 3(Jan):993–1022.

Jonathan Chang, Sean Gerrish, Chong Wang, Jordan Boyd-Graber, and David Blei. 2009. Reading tea leaves: How humans interpret topic models. *Advances in neural information processing systems*, 22.

Kostadin Cvejoski, Ramsés J Sánchez, and César Ojeda. 2023. Neural dynamic focused topic model. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 37, pages 12719–12727.

Adji B Dieng, Francisco JR Ruiz, and David M Blei. 2019. The dynamic embedded topic model. *arXiv preprint arXiv:1907.05545*.

Tomoki Doi, Masaru Isonuma, and Hitomi Yanaka. 2024. Topic modeling for short texts with large language models. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 4: Student Research Workshop)*, pages 21–33.

Hafsa Ennajari, Nizar Bouguila, and Jamal Bentahar. 2025. Correlated topic modeling for short texts in spherical embedding spaces. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 47(6):4567–4578.

Maarten Grootendorst. 2022. Bertopic: Neural topic modeling with a class-based tf-idf procedure. *arXiv preprint arXiv:2203.05794*.

R Devon Hjelm, Alex Fedorov, Samuel Lavoie-Marchildon, Karan Grewal, Phil Bachman, Adam Trischler, and Yoshua Bengio. 2019. Learning deep representations by mutual information estimation and maximization. In *International Conference on Learning Representations*.

Koichi Hori, Naoya Takeishi, Takehisa Yairi, and Rem Hida. 2018. Dynamic and static topic model for analyzing time-series document collections. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*. Association for Computational Linguistics (ACL).

Alexander Hoyle, Pranav Goel, Andrew Hian-Cheong, Denis Peskov, Jordan Boyd-Graber, and Philip Resnik. 2021. Is automated topic model evaluation broken? the incoherence of coherence. *Advances in neural information processing systems*, 34:2018–2033.

Muhammad Inaam ul haq, Qianmu Li, Jun Hou, and Adnan Iftekhar. 2023. Detecting the research structure and topic trends of social media using static and dynamic probabilistic topic models. *Aslib Journal of Information Management*, 75(2):215–245.

Patrick Jähnichen, Florian Wenzel, Marius Kloft, and Stephan Mandt. 2018. Scalable generalized dynamic topic models. In *International Conference on Artificial Intelligence and Statistics*, pages 1427–1435. PMLR.

Charu Karakkaparambil James, Mayank Nagda, Nooshin Haji Ghassemi, Marius Kloft, and Sophie Fellenz. 2024. Evaluating dynamic topic models. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 160–176.

Ziwei Ji, Nayeon Lee, Rita Frieske, Tiezheng Yu, Dan Su, Yan Xu, Etsuko Ishii, Ye Jin Bang, Andrea Madotto, and Pascale Fung. 2023. Survey of hallucination in natural language generation. *ACM computing surveys*, 55(12):1–38.

Krissanee Kamthawee, Can Udomcharoenchaikit, and Sarana Nutanong. 2024. Mist: mutual information maximization for short text clustering. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 11309–11324.

Diederik P Kingma. 2014. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*.

Diederik P Kingma and Max Welling. 2013. Auto-encoding variational bayes. *arXiv preprint arXiv:1312.6114*.

Jey Han Lau, David Newman, and Timothy Baldwin. 2014. Machine reading tea leaves: Automatically evaluating topic coherence and topic model quality. In *Proceedings of the 14th Conference of the European Chapter of the Association for Computational Linguistics*, pages 530–539.

Seonae Lee, Taehun Lee, Ren Liu, Soohyung Joo, and Dexin Shi. 2025. Research trends and challenges in diagnostic classification models: Insights from dynamic topic modeling. *Measurement: Interdisciplinary Research and Perspectives*, pages 1–32.

Ang Li, Yawen Li, Yingxia Shao, and Bingyan Liu. 2023. Multi-view scholar clustering with dynamic interest tracking. *IEEE Transactions on Knowledge and Data Engineering*, 35(9):9671–9684.

Qian Li, Liangyun Liu, Ming Xu, Bin Wu, and Yunpeng Xiao. 2019. Gdtm: A gaussian dynamic topic model for forwarding prediction under complex mechanisms. *IEEE Transactions on Computational Social Systems*, 6(2):338–349.

Yang Lin, Xinyu Ma, Xin Gao, Ruiqing Li, Yasha Wang, and Xu Chu. 2024. Combating label sparsity in short text topic modeling via nearest neighbor augmentation. In *Findings of the Association for Computational Linguistics ACL 2024*, pages 13762–13774.

Jianghan Liu, Ziyu Shang, Wenjun Ke, Peng Wang, Zhizhao Luo, Jiajun Liu, Guozheng Li, and Yining Li. 2025. Llm-guided semantic-aware clustering for topic modeling. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 18420–18435.

Lu Liu, Nima Dehmamy, Jillian Chown, C Lee Giles, and Dashun Wang. 2021. Understanding the onset of hot streaks across artistic, cultural, and scientific careers. *Nature communications*, 12(1):5392.

Rong Lu and Qing Yang. 2012. Trend analysis of news topics on twitter. *International Journal of Machine Learning and Computing*, 2(3):327.

Laurens van der Maaten and Geoffrey Hinton. 2008. Visualizing data using t-sne. *Journal of machine learning research*, 9(Nov):2579–2605.

J MCQUEEN. 1967. Some methods for classification and analysis of multivariate observations. In *Proc. Fifth Berkeley Symposium on Mathematical Statistics and Probability, 1967*, pages 281–297.

Nozomu Miyamoto, Masaru Isonuma, Sho Takase, Junichiro Mori, and Ichiro Sakata. 2023. Dynamic structured neural topic model with self-attention mechanism. In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 5916–5930.

Quang Duc Nguyen, Tung Nguyen, Duc Anh Nguyen, Linh Ngo Van, Sang Dinh, and Thien Huu Nguyen. 2025. Glocom: A short text neural topic model via global clustering context. In *Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 1109–1124.

Sebastian Nowozin, Botond Cseke, and Ryota Tomioka. 2016. f-gan: Training generative neural samplers using variational divergence minimization. *Advances in neural information processing systems*, 29.

Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, and 1 others. 2019. Pytorch: An imperative style, high-performance deep learning library. *Advances in neural information processing systems*, 32.

Chau Pham, Alexander Hoyle, Simeng Sun, Philip Resnik, and Mohit Iyyer. 2024. 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)*, pages 2956–2984.

Hamed Rahimi, Hubert Naacke, Camelia Constantin, and Bernd Amann. 2024. Antm: Aligned neural topic models for exploring evolving topics. *Transactions on Large-Scale Data-and Knowledge-Centered Systems LVI: Special Issue on Data Management-Principles, Technologies, and Applications*, 14790:76–97.

Nils Reimers and Iryna Gurevych. 2019. Sentence-bert: Sentence embeddings using siamese bert-networks. 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)*, pages 3982–3992.

Michael Röder, Andreas Both, and Alexander Hinneburg. 2015. Exploring the space of topic coherence measures. In *Proceedings of the eighth ACM international conference on Web search and data mining*, pages 399–408.

Hinrich Schütze, Christopher D Manning, and Prabhakar Raghavan. 2008. *Introduction to information retrieval*, volume 39. Cambridge University Press Cambridge.

Jingbo Shang, Jialu Liu, Meng Jiang, Xiang Ren, Clare R Voss, and Jiawei Han. 2018. Automated phrase mining from massive text corpora. *IEEE Transactions on Knowledge and Data Engineering*, 30(10):1825–1837.

Anubhav Sharma, Seba Susan, Anmol Bansal, and Arjun Choudhry. 2022. Dynamic topic modeling of covid-19 vaccine-related tweets. In *Proceedings of the 2022 5th International Conference on Data Storage and Data Engineering*, pages 79–84.

Tian Shi, Kyeongpil Kang, Jaegul Choo, and Chandan K Reddy. 2018. Short-text topic modeling via non-negative matrix factorization enriched with local word-context correlations. In *Proceedings of the 2018 world wide web conference*, pages 1105–1114.

Dominik Stammbach, Vilém Zouhar, Alexander Hoyle, Mrinmaya Sachan, and Elliott Ash. 2023. Revisiting automated topic model evaluation with large language models. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 9348–9357.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz

Kaiser, and Illia Polosukhin. 2017. Attention is all you need. *Advances in neural information processing systems*, 30.

Tu Vu, Manh Do, Tung Nguyen, Linh Ngo Van, Sang Dinh, and Thien Huu Nguyen. 2025. Topic modeling for short texts via optimal transport-based clustering. In *Findings of the Association for Computational Linguistics: ACL 2025*, pages 7666–7680.

Xiaobao Wu, Xinshuai Dong, Thong Thanh Nguyen, and Anh Tuan Luu. 2023. Effective neural topic modeling with embedding clustering regularization. In *International Conference on Machine Learning*, pages 37335–37357. PMLR.

Xiaobao Wu, Xinshuai Dong, Liangming Pan, Thong Nguyen, and Luu Anh Tuan. 2024. 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*, pages 3088–3105.

Xiaohao Yang, He Zhao, Weijie Xu, Yuanyuan Qi, Jueqing Lu, Dinh Phung, and Lan Du. 2025. Neural topic modeling with large language models in the loop. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1377–1401.

An Zeng, Zhesi Shen, Jianlin Zhou, Ying Fan, Zengru Di, Yougui Wang, H Eugene Stanley, and Shlomo Havlin. 2019. Increasing trend of scientists to switch between topics. *Nature communications*, 10(1):3439.

Delvin Ce Zhang and Hady Lauw. 2022a. Dynamic topic models for temporal document networks. In *International Conference on Machine Learning*, pages 26281–26292. PMLR.

Delvin Ce Zhang and Hady Lauw. 2022b. Meta-complementing the semantics of short texts in neural topic models. *Advances in Neural Information Processing Systems*, 35:29498–29511.

Ruiqi Zhong, Heng Wang, Dan Klein, and Jacob Steinhardt. 2024. Explaining datasets in words: Statistical models with natural language parameters. *Advances in Neural Information Processing Systems*, 37:79350–79380.

Xiaokang Zhou, Wei Liang, I Kevin, Kai Wang, Runhe Huang, and Qun Jin. 2018. Academic influence aware and multidimensional network analysis for research collaboration navigation based on scholarly big data. *IEEE Transactions on Emerging Topics in Computing*, 9(1):246–257.

Yuxuan Zhou, Margret Keuper, and Mario Fritz. 2025. Balancing diversity and risk in LLM sampling: How to select your method and parameter for open-ended text generation. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 26352–26365.

A Prompts

In this section, we provide the prompt designs used in the GEA Topic Refiner. As shown in Figure 5, the *Generator* is prompted to produce candidate natural language descriptions for each latent topic based on representative short texts and corresponding scores sampled across different time periods. The *Evaluator* is designed to mitigate hallucination and select the most semantically aligned topic descriptions. As shown in Figure 6, the prompt consists of a binary question asking whether the text sampled from the corpus falls within the semantic scope of the given topic description. These binary judgments are later aggregated and compared with document-topic similarity scores to select the optimal topic description. The *Assigner* is instructed to assign the text to the single most relevant topic, which produces a refined and more interpretable document-topic distribution, as shown in Figure 7. These prompt designs enable effective collaboration among the *Generator*, *Evaluator*, and *Assigner*, ensuring that the refined topics are both semantically accurate and human-interpretable.

B Dataset Details

To ensure a comprehensive evaluation, we conducted experiments on three short text datasets: StackOverflow, NYT, and Biomedical. A brief introduction to these datasets is given below:

StackOverflow: A dataset derived from the public Kaggle challenge data[1], where we randomly sample question titles between 2008 and 2012 as short texts and use the accompanying technical tags provided on the Q&A website as reference labels.

NYT: A dataset contains news headlines collected from *The New York Times* online archive[2] spanning the years 2001 to 2024.

Biomedical: A dataset consists of paper titles randomly selected from the challenge data published on the BioASQ official website[3], covering publications from 1980 to 2013.

The statistics of the processed datasets are summarized in Table 5. For each dataset, time slices are partitioned by year. In the data preprocessing phase, we employ AutoPhrase (Shang et al., 2018) to extract terms from the temporal text, then remove

[1] https://www.kaggle.com/competitions/predict-closed-questions-on-stack-overflow/data?select=train.zip

[2] https://www.kaggle.com/datasets/aryansingh0909/nyt-articles-21m-2000-present

[3] http://participants-area.bioasq.org/

Please suggest topic descriptions about the text samples that are more likely to achieve higher scores. Here is a corpus of text samples each associated with a score. The text samples are sorted from the lowest to the highest score.

Please suggest {num_candidate_descriptions} topic descriptions to me, one in a line, starting with "-" and surrounded by quotes "". For example:

- "a topic of technology innovation."

Please generate the response based on the given datapoints as much as possible. Do not output anything else.

{*Text 1: content. (Score)*

Text 2: content. (Score)

Text 3: content. (Score)

...

...

...}

Response:

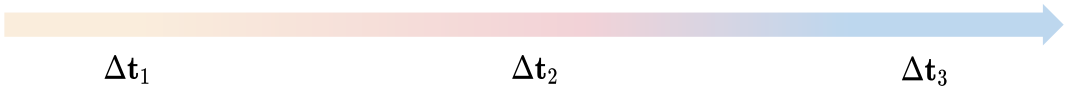

Figure 5: Prompt for the *Generator*. The content enclosed in curly braces {} represents placeholders for dynamic data instances. As depicted in the bottom timeline, we partition the corpus into three chronological periods: early, middle, and late phases, denoted as $\Delta t_p, p \in \{1, 2, 3\}$. The input text samples (highlighted in different colors) are sampled from these respective phases and sorted by their associated scores.

Confirm whether the Text satisfies a Topic. Respond with Yes or No. When uncertain, output No.

Example 1:

TOPIC: "a topic of a natural scene."

TEXT: "I love the way the sun sets in the evening."

Response: Yes.

Example 2:

TOPIC: "a topic of historical dates."

TEXT: "A member of the Democratic Party, he was the first African-American president of the United States."

Response: No. The TEXT does not mention date.

Now complete the following example,

TOPIC: "{topic_description}"

TEXT: "{text}"

Response:

Figure 6: Prompt for the *Evaluator*. The placeholders `{topic_description}` and `{text}` are instantiated with candidate topic descriptions and random text samples, respectively.

In this task, you will need to select the most appropriate topic for a given text and return the corresponding index, surrounded with []. Do not provide any explanations. For example,
TOPICS:
0. a topic of sports
1. a topic of biomedical
2. a topic of science
TEXT: Researchers developed a new nanoparticle for targeted drug delivery.
Response:
[1]
Now classify this text by selecting the most relevant topic index. Return only the index in brackets, without explanations.
TOPICS:
{topics_with_index}
TEXT: {text}
Response:

Figure 7: Prompt for the *Assigner*. The placeholders `{topics_with_index}` and `{text}` are instantiated with indexed candidate topic descriptions and the target text samples to be assigned, respectively.

Table 5: Statistics of the datasets used in our experiments. #Docs: Number of documents; Avg Len: Average document length (number of tokens); #Term: Term size; #Labels: Number of ground-truth categories; #Time: Number of time slices.

Dataset	#Docs	Avg Len	#Labels	#Term	#Time
StackOverflow	19,796	4.974	9	3,246	5
NYT	24,952	4.512	20	4,601	24
Biomedical	20,463	9.832	20	5,803	34

stopwords and filter out terms with a frequency of less than 3.

C Details of Evaluation Metrics

In this section, we provide more detailed explanations and mathematical definitions of evaluation metrics.

C.1 Purity

Purity measures the extent to which a cluster contains only data from a single class. Since the datasets contain ground-truth labels, we evaluate the quality of the temporal doc-topic distribution by treating the model as a document clustering method. We assign each document to the topic with the highest probability. For a set of predicted clusters $\mathcal{C} = \{c_1, \ldots, c_K\}$ and ground-truth classes $G = \{g_1, \ldots, g_M\}$, Purity is defined as:

$$\text{Purity} = \frac{1}{N} \sum_{k=1}^{K} \max_{j} |c_k \cap g_j|, \tag{15}$$

where N is the total number of documents. High purity implies that each discovered topic maps cleanly to a real-world category.

C.2 NMI

Normalized Mutual Information (NMI) (Schütze et al., 2008) is an information-theoretic measure that quantifies the mutual dependence between the predicted clusters and the ground-truth labels, and is normalized to allow comparison. It is defined as:

$$\text{NMI} = \frac{2 \times I(\mathcal{C}; G)}{H(\mathcal{C}) + H(G)}, \tag{16}$$

where $I(\mathcal{C}; G)$ is the mutual information between clusters and labels, and $H(\cdot)$ represents the entropy. NMI balances the trade-off between the number of clusters and clustering quality, with values in $[0, 1]$.

C.3 NPMI

Normalized Pointwise Mutual Information (NPMI) (Lau et al., 2014) measures topic coherence by

Table 6: The cost of running the GEA Topic Refiner on three datasets.

LLM (*Generator*, *Evaluator/Assigner*)	StackOverflow	NYT	Biomedical
Qwen3-max, Qwen-max	$2.74	$2.91	$4.08
DeepSeek-R1, DeepSeek-V3	$1.99	$2.08	$2.78
GPT-4o, GPT-3.5-turbo	$4.87	$5.33	$8.24
GPT-5, GPT-5-nano	$0.96	$1.13	$2.01

focusing on pairwise co-occurrence of topic words. The NPMI between words is defined as:

$$\text{NPMI}(w_i, w_j) = \frac{\log \frac{P(w_i,w_j)}{P(w_i)P(w_j)}}{-\log P(w_i, w_j)}, \tag{17}$$

where $P(w_i)$ is the probability of word w_i, and $P(w_i, w_j)$ is the joint probability of words w_i and w_j co-occurring in a sliding window. NPMI ranges from -1 to 1, with higher values indicating better coherence. We calculate the average NPMI across all time slices.

C.4 C_V

The C_V (Röder et al., 2015) evaluates topic coherence from the perspective of the overall semantic consistency of the topic. It employs a sliding window and computes the cosine similarity between the normalized pointwise mutual information vectors of the top words. Given a k-th topic z at time slice t represented by its top m words $\{w_1, \ldots, w_m\}$. e_i is a context vector for each word $w_i, i \in [1, m]$ and is defined as:

$$e_i = [\text{NPMI}(w_i, w_1), \ldots, \text{NPMI}(w_i, w_m)]^\top. \tag{18}$$

The C_V score is the average cosine similarity between each word's vector e_i and the aggregate vector of all top words $\bar{e} = \sum_{j=1}^{m} e_j$:

$$C_V(z) = \frac{1}{m} \sum_{i=1}^{m} \cos(e_i, \bar{e}), \tag{19}$$

where $\cos(\cdot, \cdot)$ denotes the cosine similarity. We calculate the average C_V across all time slices. A higher C_V indicates that the top words share a consistent semantic context.

C.5 TD

Topic Diversity (TD) assesses the uniqueness of discovered topics. We compute the percentage of words that appear only once and exist in the vocabulary of the current time slice among the top m words of the K topics at the time slice t, to measure the TD following (Wu et al., 2024).

$$\text{TD} = \frac{N_o^{(t)}}{mK}, \tag{20}$$

where $N_o^{(t)}$ represents the number of words that appear only once and also exist in the vocabulary set $W^{(t)}$. We calculate the average TD across all time slices.

C.6 TTC

Temporal Topic Coherence (TTC) (James et al., 2024) captures the semantic stability of a topic over time by considering word pairs between two consecutive timestamps. Given a k-th topic z at time slice t represented by its top m words, TTC is defined as follows:

$$\text{TTC}(z) = \frac{1}{m^2} \sum_{i=1}^{m} \sum_{j=1}^{m} C_v(w_i^{(k,t)}, w_j^{(k,t+1)}), \tag{21}$$

where $w_i^{(k,t)}$ is the i-th word of topic k at time slice t. We calculate the average TTC between each pair of time slices. Higher TTC means the same topic remains stable and coherent across time slices. Conversely, a lower TTC indicates that the topic words within the same topic exhibit incoherence over time.

D Implementation Details

We implement our method using the PyTorch framework (Paszke et al., 2019) and train it for 1000 iterations on a single NVIDIA A100 40 GB GPU with the Adam (Kingma, 2014) optimizer. The weighting coefficient λ in Eq. 5 and α in Eq. 14 are set to 0.7. We set the number of topics to 20 and report averages over five runs in the main experiment, except for ANTM and TopicGPT, which do not support specifying the number of topics, and TopicGPT is evaluated on a single run. For TopicGPT, we use GPT-4o for topic generation and GPT-3.5-turbo for topic assignment.

Table 7: Ablation study results on the StackOverflow dataset with different numbers of topics K. Symbol * indicates significant improvements over the baselines through the T-test ($p \leq 0.05$). "-" denotes the metric is not applicable to the method.

	K=10						K=30						K=50					
Method	Purity	NMI	NPMI	C_V	TD	TTC	Purity	NMI	NPMI	C_V	TD	TTC	Purity	NMI	NPMI	C_V	TD	TTC
BERTopic	0.241	0.247	-0.082	0.391	0.886	0.282	0.301	0.270	0.193	0.362	0.903	0.245	0.367	0.303	0.452	0.349	0.888	0.213
DTM	0.122	0.062	-0.119	0.375	0.776	0.309	0.144	0.096	-0.108	0.500	0.880	0.430	0.226	0.110	-0.131	0.542	0.929	0.479
DETM	0.155	0.101	-0.067	0.445	0.872	0.358	0.203	0.159	0.224	0.551	0.694	0.457	0.283	0.250	0.481	0.588	0.558	0.493
DSNTM	0.136	0.076	-0.114	0.526	0.918	0.429	0.155	0.102	0.241	0.498	0.642	0.408	0.220	0.102	0.339	0.462	0.482	0.385
CFDTM	0.208	0.143	-0.042	0.367	0.991	0.330	0.246	0.159	-0.048	0.419	0.976	0.362	0.235	0.103	0.469	0.474	0.720	0.425
ECRTM+LLM-ITL	0.216	0.157	-	-	-	-	0.231	0.144	-	-	-	-	0.227	0.105	-	-	-	-
KNNTM	0.303	0.322	-	-	-	-	0.460	0.379	-	-	-	-	0.447	0.343	-	-	-	-
GloCOM-EnCOT	0.359	0.367	-	-	-	-	0.502	0.416	-	-	-	-	0.501	0.419	-	-	-	-
Ours	**0.427***	**0.422***	**0.038***	**0.602***	**1.000***	**0.524***	**0.533***	**0.452***	**0.250***	**0.612***	**1.000***	**0.568***	**0.537***	**0.441***	**0.492***	**0.617***	**1.000***	**0.580***

For the agents of the GEA Topic Refiner, we utilize four different LLM families in our experiments. For each family, we employ the flagship version for the *generator*, while using the lightweight version for the *evaluator* and *assigner* to balance performance and computational costs. Specifically, the model pairings are: (1) **GPT-5** for *generator* paired with **GPT-5-nano** for *evaluator* and *assigner*, (2) **GPT-4o** paired with **GPT-3.5-turbo**, (3) **DeepSeek-R1** paired with **DeepSeek-V3**, and (4) **Qwen3-max** paired with **Qwen-max**. Table 6 lists the specific API costs for these models across the three datasets. Notably, the combination of GPT-5 and GPT-5 nano is the most cost-efficient option in our experiments. It achieves strong performance while incurring the lowest overall inference cost. Therefore, we adopt this configuration as the default setting for the GEA Topic Refiner in our experiments. We set $A = 30$, $E = 5$ in the *generator* and $F = 256$ in the *evaluator*.

E Ablation Study

E.1 Number of Topics

We conduct an ablation study to analyze the sensitivity of DVI-DTM to the number of topics K. As shown in Table 7, we examine how topic granularity influences the quality of dynamic topic modeling. When $K = 10$, the topics are coarse, causing multiple semantically distinct concepts to merge into a single topic, which results in lower topic coherence. As K increases to 30, our model achieves a better balance between topic granularity and semantic coherence, leading to relatively improved topic alignment performance. Further increasing K to 50 introduces redundant topics, slightly degrading performance. The results show that DVI-DTM maintains stable performance across different topic numbers, indicating strong robustness to the choice of K. Regardless of the value of K, our method consistently achieves the best performance across all metrics compared with the existing state-of-the-art baselines, demonstrating its ability to adapt to different topic granularities without sacrificing topic alignment or temporal coherence.

E.2 Hyperparameter Analysis

Another ablation study is conducted to evaluate the impact of the weighting hyperparameters in Eq. 5 and Eq. 14, as shown in Figure 8. The parameter λ controls the strength of the temporal smoothing operation in the DVR module. Setting λ too large weakens temporal continuity, leading to unstable topic assignments across time, while small values over-smooth the distributions and obscure emerging topics. The model achieves the best topic alignment performance at $\lambda = 0.7$. The parameter α adjusts the trade-off between semantic relevance and temporal uniqueness in the DFR module. Optimal performance on dynamic topic quality metrics is achieved when $\alpha = 0.7$, demonstrating that jointly considering semantic relevance and temporal uniqueness is essential for accurately capturing meaningful topic evolution.

E.3 Case Studies on Hallucination Mitigation

To demonstrate the efficacy of GEA Topic Refiner in mitigating hallucinations, we present representative cases from three datasets in Table 8. As shown in the first row of Table 8, the LLM completely ignores the required prompt format and generates a meta-commentary sentence for a text instead of a

Table 8: Qualitative validation of topic description evaluation.

Dataset	Output topic description	Rejected hallucinated topic description
StackOverflow	A topic of system-level design over syntax fixes	Text 308: Discussing JPA ORM suitability
NYT	A topic of personal love stories tied to weddings	A topic of recipe-driven food coverage
Biomedical	A topic of experimental toxicology with protective agents	A topic of chemical modulation of platelet pathways

Table 9: Comparison of runtime on the StackOverflow dataset. * indicates KNNTM requires an additional 65 hours for pre-computation of optimal transport distance before running.

Method	Type	Category	Runtime (s)
BERTopic	Dynamic	Non-LLM	217
ANTM	Dynamic	Non-LLM	1,075
DTM	Dynamic	Non-LLM	3,563
DETM	Dynamic	Non-LLM	937
DSNTM	Dynamic	Non-LLM	1,687
CFDTM	Dynamic	Non-LLM	2,093
KNNTM	Static	Non-LLM	100*
GloCOM-EnCOT	Static	Non-LLM	712
TopicGPT	Static	LLM-based	8,628
ECRTM+LLM-ITL	Static	LLM-based	11,113
Ours (without GEA)	Dynamic	Non-LLM	795
Ours (with GEA)	Dynamic	LLM-based	8,420

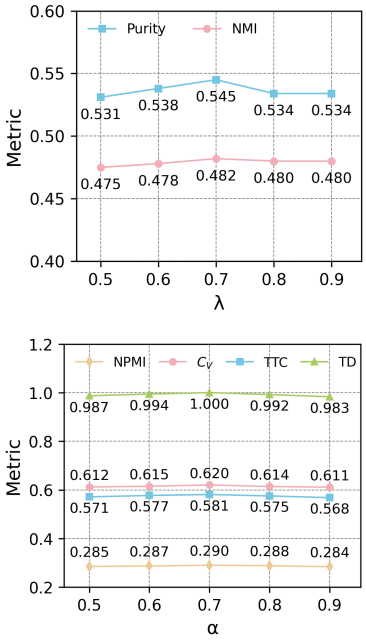

Figure 8: Ablation study results of hyperparameter λ and α on the StackOverflow dataset.

topic description for a latent topic. Furthermore, as shown in the second and third rows of Table 8, the LLM is easily misled by isolated or out-of-context words in sparse short texts, producing hallucinated topic descriptions with severe domain drift. The *Evaluator* can effectively reject descriptions that are malformed and inconsistent with the context, ensuring that only descriptions with high thematic alignment and grounding are preserved.

E.4 Runtime Analysis

Although topic modeling is typically an offline task, a runtime analysis is valuable for understanding practical trade-offs. We conducted a systematic evaluation of the execution time of DVI-DTM against all baselines under the same hardware and software settings. As shown in Table 9, the *Assigner* in the GEA Topic Refiner introduces additional computational overhead due to frequent LLM inference at the document level. Even with this component included, our method remains more efficient than other LLM-based baselines, while simultaneously achieving the best topic modeling performance and providing interpretable topic descriptions. When the GEA Topic Refiner is disabled, our framework demonstrates substantially improved efficiency. In this setting, the runtime of our DVI-DTM is faster than most dynamic topic models. Although BERTopic exhibits shorter runtime due to its clustering-based pipeline without temporal alignment, it performs significantly worse than our method across all metrics. Importantly, even without the GEA Topic Refiner, DVI-DTM still achieves highly competitive topic modeling performance compared with existing baselines, as shown in Table 1 and Table 2 (d).

F Details of Human Evaluation

In this section, we further detail the human evaluation procedure and setup. We randomly sampled 15 topic term evolution results generated from each model, including our model and competitive baselines, based on the StackOverflow dataset. We exclude BERTopic and ANTM from the human evaluation, as they do not guarantee continuous and fixed-length topic term evolution trajectories. To ensure impartiality, the evaluation was conducted in a blind setting, where the source model for each

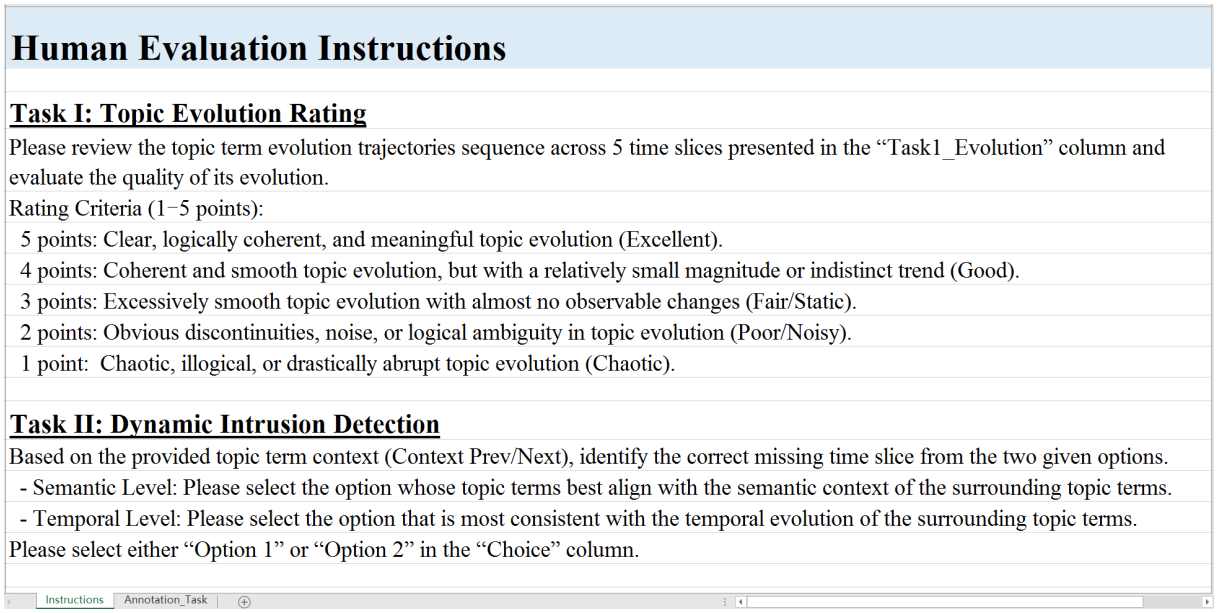

Figure 9: Human evaluation instructions.

topic term evolution result was anonymized and shuffled. Three volunteer graduate students were recruited as annotators to perform the evaluation independently, and we encouraged them to use external resources to aid them.

Figure 9 illustrates the detailed instructions presented to the annotators. Two types of evaluation tasks are designed. The first task, Topic Evolution Rating, directly evaluates the quality of **evolution** and the logical coherence of individual topics over time. Annotators were asked to review the topic term sequence across 5 time slices for each model and assign a score based on a 5-point Likert scale (higher scores indicate better performance). The specific rating criteria are defined as follows:

- **5 (Excellent)**: The topic evolution is clear, logically coherent, and meaningful.
- **4 (Good)**: The evolution is coherent and smooth, though the magnitude of change is relatively small or the trend is slightly indistinct.
- **3 (Fair/Static)**: The evolution is excessively smooth with almost no observable changes, indicating an over-smoothed or static topic.
- **2 (Poor/Noisy)**: The evolution contains obvious discontinuities, noise words, or logical ambiguities.
- **1 (Chaotic)**: The topic evolution is chaotic, illogical, or drastically abrupt.

The second task, Dynamic Intrusion Detection, is designed as an objective metric to assess whether the generated dynamic topic term trajectories exhibit **semantic** distinctiveness and recognizable **temporal** discriminability. Unlike subjective ratings, this task requires annotators to perform a discrimination task, providing an objective metric for topic semantic interpretability and temporal coherence. Annotators were presented with the topic term context at the previous $(t-1)$ and subsequent $(t+1)$ time slices and asked to identify the correct term set for the middle time slice t. For each evaluation, two options were provided: the ground truth generated by the model and a distractor (an "intruder"). The distractor introduced a disruption in either semantic content or temporal consistency. Specifically, distractors consisted of two types: (1) **Semantic level**: terms from a different topic at the same time slice t, to test whether the annotator could distinguish topic semantic boundaries, and (2) **Temporal level**: terms from a distant time slice of the same topic, to test sensitivity to temporal shifts. Annotators were required to select the correct missing time slice, and higher accuracy indicates that the model's topic evolution is more coherent and interpretable to humans.

As illustrated in Table 4, the evaluation results are computed as the average score assigned by each annotator for each model. Overall, our DVI-DTM consistently achieves the highest scores across all evaluation criteria.