Evaluating Dynamic Topic Models

Charu James,* **Mayank Nagda,*** **Nooshin Haji Ghassemi, Marius Kloft, Sophie Fellenz**
RPTU Kaiserslautern-Landau, Germany
{charu,nagda,nooshin,kloft,fellenz}@cs.uni-kl.de

Abstract

There is a lack of quantitative measures to evaluate the progression of topics through time in dynamic topic models (DTMs). Filling this gap, we propose a novel evaluation measure for DTMs that analyzes the changes in the quality of each topic over time. Additionally, we propose an extension combining topic quality with the model's temporal consistency. We demonstrate the utility of the proposed measure by applying it to synthetic data and data from existing DTMs, including DTMs from large language models (LLMs). We also show that the proposed measure correlates well with human judgment. Our findings may help in identifying changing topics, evaluating different DTMs and LLMs, and guiding future research in this area.

1 Introduction

Dynamic Topic Models (DTMs) (Blei and Lafferty, 2006) learn topics and their evolution over time from a time-indexed collection of documents. Variants of DTMs include traditional statistical topic models, neural VAE-based topic models, and topics learned using large language models (LLMs). DTMs have proven useful in various domains, including text mining (McCallum et al., 2005; Wang et al., 2007; Ramage et al., 2011; Gerrish and Blei, 2011), computer vision (Fei-Fei and Perona, 2005; Cao and Fei-Fei, 2007; Chong et al., 2009), and computational biology (Pritchard et al., 2000; Zheng et al., 2006). DTMs enable summarization, browsing, and searching of large document collections by capturing changes in topics over time. However, evaluating DTMs can be challenging due to their unsupervised nature, although it is crucial for effectively detecting trends in time-indexed documents.

With the advent of VAE-based and LLM-based topic models, there is an increasing need for evaluation procedures to compare these models (Ostheimer et al., 2023, 2024). While traditional evaluation measures (Dieng et al., 2019; Blei and Lafferty, 2006) can assess the quality and diversity of topics, they fail to capture the smoothness of topic changes over time. This limitation becomes problematic when a DTM has high topic quality but lacks temporal smoothness. In such cases, existing evaluation measures may incorrectly assign a high score to the model, even when there are rapid and abrupt transitions between topics. For example, if a topic quickly changes from "politics" to "sports", conventional evaluation measures may still rate the model positively. To accurately assess the quality of a DTM, it is crucial to consider the smoothness of topic changes over time, which can help identify gradual topic drifts or sudden shifts. Unfortunately, existing evaluation measures lack the ability to effectively track topic changes over time. To bridge this gap, we propose Temporal Topic Quality (TTQ)—a novel evaluation measure specifically designed for DTMs. TTQ incorporates changes in topic quality into its evaluation, thereby capturing the temporal characteristics of topics in DTMs.

We provide empirical evidence for the effectiveness of the proposed measure by evaluating it on both synthetic and real topics. The results demonstrate a positive correlation between human ratings and the individual components of the TTQ measure. To provide an overall evaluation of DTMs, we propose the Dynamic Topic Quality (DTQ). The DTQ measure aggregates the TTQ measure with the static topic quality score. This aggregation is performed for both year-wise evaluations and temporal topic assessments, as illustrated in Figure 1. In our experiments, we compare the results obtained using the DTQ measure with those obtained using previously employed measures for different topic models, including LLM-based topic models. We show that the DTQ measure effectively indi-

*Equal contribution.

Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), pages 160–176

cates the smoothness of topics in trained DTMs compared to the measures used in the past. We expect that the introduction of the new measure will contribute to improved comparisons between DTMs in future research efforts. Our contributions can be summarized as follows:

- We present a novel evaluation measure for DTMs that integrates both the vertical (year-wise) and the horizontal (temporal) dimension in the quality estimate (See Figure 1).
- We conduct a meta-analysis of prominent (statistical, neural, and LLM-based) DTMs with our novel evaluation measures and present our findings.
- We show a positive correlation between human evaluations and the new evaluation measures, confirming their validity.

2 Related Work

This section presents the previous work on DTMs and their evaluation approaches. We further discuss how the human evaluation for new measures was conducted in this domain.

Dynamic Topic Models are developed to model topics over time. Dynamic latent Dirichlet allocation (D-LDA) (Blei and Lafferty, 2006) extends the original LDA method (Blei et al., 2003) to account for temporal characteristics of sequential text data. The dynamic embedded topic model (D-ETM) combines D-LDA and word embeddings with a recurrent neural network (RNN) (Dieng et al., 2019). Sia et al. (Sia et al., 2020) propose an LLM-based topic model using pre-trained word embeddings and clustering them. While this is not a dynamic model that models topics over time, it is straight-forward to extend to a dynamic model by using an online clustering method. A continuous-time version of DTMs was introduced by Wang *et al.* (Wang et al., 2008; Wang and McCallum, 2006). This model is not applicable to our datasets which have discrete timestamps. Later work focused on the scalability of DTMs due to their computationally intensive training (Jähnichen et al., 2018; Bhadury et al., 2016). Many other DTMs have been proposed for different purposes, and not all are directly based on LDA (Grootendorst, 2022; He et al., 2013; Zhou et al., 2017; Gou et al., 2018; Morinaga and Yamanishi, 2004; Mei and Zhai, 2005; Ahmed and Xing, 2008, 2012; Dubey et al., 2013; Wang and McCallum, 2006). It can be expected that with the recent advent of neural topic models (Burkhardt and Kramer, 2019; Burkhardt et al., 2020; Nagda et al., 2021; Nagda and Fellenz, 2024) and LLM-based topic models, more DTM variants will be published in the future. In this work, we compare D-LDA and D-ETM as the major proponents of the statistical and neural DTMs. Additionally, we evaluate an LLM-based topic model (Sia et al., 2020).

Evaluation Measures for Topic Models Previous work has focused on measures for static topic models such as the *topic coherence* as measured by the normalized pointwise mutual information (NPMI) or C_v score (Newman et al., 2010; Mimno et al., 2011; Röder et al., 2015). Using a language model (LLM) for evaluation (Stammbach et al., 2023) is ineffective if the corpus is domain-specific and the LLM has not been trained on that particular corpus (Manduchi et al., 2024).

With the rise of neural topic models, local minima during training have become an issue, which may lead to component collapse (Burkhardt and Kramer, 2019; Bhat et al., 2023). To capture such and other problems, different topic diversity, redundancy, or overlap measures have been introduced (Burkhardt and Kramer, 2019; Gui et al., 2019; Dieng et al., 2020). Dieng *et al.* combine topic diversity with coherence scores, resulting in *topic quality* scores (Dieng et al., 2020), which is the basis of the definition of our temporal topic quality measure.

Human Evaluation for Topic Models The concepts of coherence and interpretability are "simultaneously important and slippery" (Lipton, 2018; Hoyle et al., 2021). A topic is coherent when a set of terms, viewed together, enables human recognition of an identifiable category (Hoyle et al., 2021). Doogan and Buntine (2021) define an interpretable topic as one that can be easily labeled and has a high level of agreement on its labels. In the case of DTMs, the topic coherence, interpretability, and smoothness across the temporal dimension are vital for its purpose and are the focus of our study.

There are two main ways to carry out human evaluation of topic models: topic intrusion and topic rating. Both were developed specifically to account for the topic coherence in static topic models. In the topic rating task, humans are presented with a topic and are asked to rate it on a scale. Previously,

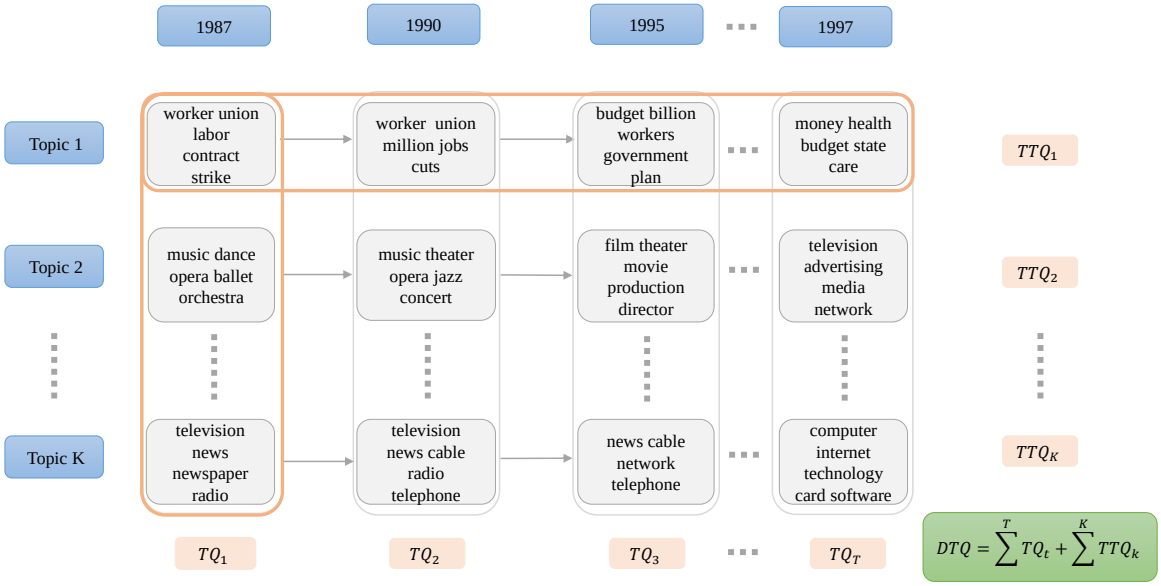

Figure 1: This figure illustrates the core concept presented in this paper. It illustrates the topic structure within DTMs. The vertical box highlights the set of topics for the first year, and the horizontal box shows the evolution of Topic 1 over time. Topic Quality (TQ) evaluates the topics for each year vertically, whereas Temporal Topic Quality (TTQ) evaluates each topic horizontally, capturing both the evolution of the topic over time and the smoothness of topic progression.

authors have used ratings on a three-point ordinal scale (Hoyle et al., 2021; Mimno et al., 2011; Aletras and Stevenson, 2013; Ostheimer et al., 2023, 2024). The rating task is not directly transferable to DTMs since we need to also rate how the topic changes over time. In the topic intrusion task, topics are chosen randomly, and one word in the topic is replaced with a word from another topic. The intruder word (Hoyle et al., 2021) is identified by Human evaluators. Here, we extend and tailor both tasks to the temporal evaluation of DTMs.

3 Background on Topic Evaluation Measures

This section reviews the most common evaluation measures for topic models, which form the basis for our proposed measures: Topic coherence, diversity, and quality.

3.1 Topic Coherence

NPMI (Röder et al., 2015) is the most commonly used coherence measure. For topic k, it is computed as

$$\phi_k = \sum_{j=2}^{N} \sum_{i=1}^{j-1} \frac{\log \frac{P\left(w_i^{(k)}, w_j^{(k)}\right)+\epsilon}{P\left(w_i^{(k)}\right)P\left(w_j^{(k)}\right)}}{-\log(P\left(w_i^{(k)}, w_j^{(k)}\right)+\epsilon)}, \quad (1)$$

where $(w_1^{(k)}, \ldots, w_N^{(k)})$ is a list of the top N words in topic k, and $P(w_i^{(k)}, w_j^{(k)})$ is the probability of words $w_i^{(k)}$ and $w_j^{(k)}$ occurring together in a document, which is approximated by counting the number of documents where both words appear together, divided by the total number of documents (Aletras and Stevenson, 2013). A sliding window is used that determines the words to be considered at a time. The C_v score (Röder et al., 2015) extends the NPMI by creating content vectors using co-occurrences of words, then calculating NPMI and cosine similarity between words.

3.2 Topic Diversity

There exist (at least) three different approaches to measure diversity in topic models. The measure by (Burkhardt and Kramer, 2019) takes into account *how often* a word is repeated across topics and not only *if* it is repeated. Additionally, it allows us to compute the diversity for individual topics and not just for the whole topic model. It computes the diversity for topic k as $\mathrm{d}_{k,C} = 1 - \mathrm{r}_{k,C}$, where $r_{k,C}$ is the redundancy of topic k with respect to the other topics $C = (v_1, \ldots, v_{k-1}, v_{k+1}, \ldots, v_K)$, where v_i denotes the list of words for topic i and can be

obtained as follows

$$\mathrm{r}_{k,C} = \frac{1}{K-1} \sum_{i=1}^{N} \sum_{q \in C} \mathbb{1}(w_i^{(k)}, q), \qquad (2)$$

where $\mathbb{1}(w_i^{(k)}, q)$ is one if the ith word of topic k, $w_i^{(k)}$, occurs in topic q and otherwise zero. K is the number of topics, and $\mathrm{r}_{k,C}$ ranges from 0 to 1. Redundancy close to zero indicates that a topic has words that do not occur in any other topic, and redundancy close to one indicates that most words in a topic also occur in (multiple) other topics. This is the primary measure used in the current work.

A related measure by Gui et al. (Gui et al., 2019) computes the Topic Overlap (TO). A high value in TO indicates that the associated words frequently appear across topics and can therefore be considered background words (Gui et al., 2019). Dieng et al. (Dieng et al., 2019) proposed a third measure, which computes the topic diversity as the percentage of unique words in the top N topics. Having a diversity near zero indicates redundant topics. All three measures rank topics in the same order and thus lead to the same correlation values in our experiments.

3.3 Topic Quality

Topic quality is defined as a combination of topic coherence (TC) and topic diversity (TD). A high diversity ensures that words across topics are different, and a high coherence ensures that words within topics are highly related, resulting in high-quality topics. While Dieng et al. (Dieng et al., 2019) used NPMI for coherence and their own diversity measure, the two components can be exchanged with different coherence and diversity measures. For K different topics, it is computed as

$$\mathrm{TQ} = \frac{1}{K} \sum_{k=1}^{K} \phi_k \cdot d_{k,C},$$

where ϕ_k can be any coherence measure such as NPMI or C_v score.

4 Proposed Measures

None of the existing measures is suitable for evaluating temporal topic changes. Our proposed measure fills this gap. First, we present the temporal topic coherence and smoothness measures. We then show how the two are combined to form temporal topic quality, which measures the quality of topic transitions over time. This measure is then used in an aggregated measure, the dynamic topic quality, which evaluates both crucial aspects in DTMs: the quality of the topic model in each year and the quality of topic transitions over time.

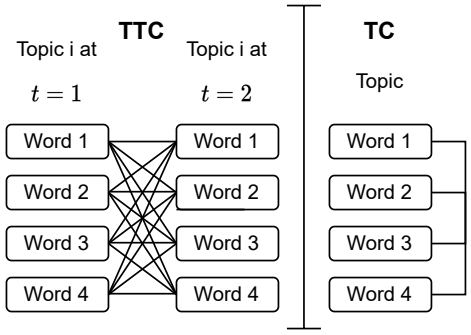

Figure 2: The idea of temporal topic coherence (TTC) in comparison with topic coherence (TC). TC considers word pairs within one topic. TTC only considers word pairs across timestamps of one topic.

4.1 Temporal Topic Coherence

Temporal topic coherence (TTC, see Fig. 2) considers word pairs between two consecutive timestamps of one topic. Otherwise, the principle is the same as in TC: the co-occurrence of each word pair in the reference corpus is counted. Thus, if the topic remains semantically the same, TTC will be high, whereas if words associated with the same topic in consecutive timestamps do not occur together in the reference corpus, TTC will be low. The results of temporal topic coherence are significantly influenced by the reference corpus used. Here we use each dataset as a reference corpus. More formally, we can now define the temporal topic coherence for window size L as

$$\mathrm{TTC}_{k,t} = \sum_{j=1}^{N} \sum_{i=1}^{N} \frac{\log \frac{P\left(w_i^{(k,t)}, w_j^{(k,t+L)}\right) + \epsilon}{P\left(w_i^{(k,t)}\right) P\left(w_j^{(k,t+L)}\right)}}{-\log\left(P\left(w_i^{(k,t)}, w_j^{(k,t+L)}\right) + \epsilon\right)}, \qquad (3)$$

where the variables are defined as in the definition of TC except $w_i^{(k,t)}$ is the ith word in topic k and timestamp t.

4.2 Temporal Topic Smoothness

The idea of temporal topic smoothness (TTS) is to use the diversity measure (introduced in Section 3.2), but instead of applying it vertically for one topic model, we apply it horizontally over time (see

Figure 1). In this case, the goal is to have smooth changes, which corresponds to a low diversity measure. Therefore, smoothness can be considered to be the opposite of diversity d. We apply TTS to one topic over a window of time steps. TTS for topic k in a window with size L can be obtained by

$$\text{TTS}_{k,t} = r_{k,\tilde{C}},$$

where $\tilde{C} = (v^{(k,t)}, \ldots, v^{(k,t+L-1)})$ and $r_{k,\tilde{C}}$ is defined in Equation 2.

4.3 Temporal Topic Quality

Temporal Topic Coherence is calculated with respect to a reference corpus, whereas Temporal Topic Smoothness is only based on the words of the topics themselves. Thus, they are complementary since TTC can be high and TTS low or the other way around. High TTS and low TTC would point to component collapsing (an incoherent topic that is repeated over time), whereas high TTC and low TTS could point for example to changes in vocabulary use over time in topically coherent topics. As a combination of both measures, analogously to the topic quality measure, we propose the temporal topic quality (TTQ).

TTQ for topic k over a sequence of timestamps $t = 1, \cdots, T$ with a window size of L can be computed as

$$\text{TTQ}_k = \frac{1}{T-L+1} \sum_{t=1}^{T-L+1} \text{TTC}_{k,t} \cdot \text{TTS}_{k,t}.$$

The window size parameter L enables to calculate the measure at different resolutions which correspond to detecting rapid changes (small window size) or slow transitions (large window size).

4.4 Dynamic Topic Quality

TTQ enables us to see the changes of a topic over time (see Figure 1 horizontal box), but it tells us nothing about the relation between different topics within one timestamp or their coherence (see Figure 1 vertical box). The role of TQ is to ensure the created topics are coherent and diverse. A DTM should exhibit high TTQ and TQ. Evaluating a model based on both measures gives rise to an aggregated measure called dynamic topic quality, DTQ, which measures the overall quality of a DTM and can be computed as

$$\text{DTQ} = \frac{1}{2}\left[\frac{1}{T}\sum_{t=1}^{T}\text{TQ}_t + \frac{1}{K}\sum_{k=1}^{K}\text{TTQ}_k\right],$$

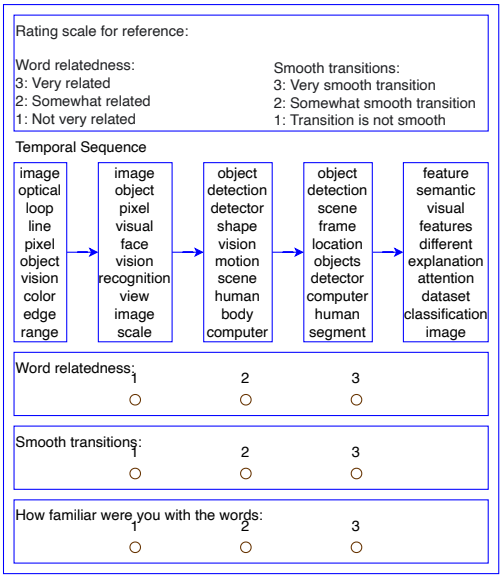

Figure 3: The temporal topic rating task as presented to the human evaluators. We present a sequence of five timestamps with equal spacing to evaluate a temporal topic.

where TQ_t is the average quality of all topics in year t. Using DTQ, DTMs can be compared and ranked based on their performance on sequential text data.

5 Using Human Evaluation to Examine DTMs

Since topic modeling is unsupervised and no ground truth for topic quality is available, human evaluation is needed to validate our measures. In this section, we propose two tasks for the human evaluation of DTMs. The proposed tasks are adapted from the widespread *word intrusion* (Chang et al., 2009) and *topic rating* (Newman et al., 2010) tasks for static topic models.

5.1 Temporal Word Intrusion

Word intrusion (Chang et al., 2009) is a common way of evaluating topic models. Intrusion words are chosen such that they have a low probability of belonging to the target topic, but a high probability of belonging to another topic. Words of existing topics are replaced by the intrusion words. Humans are then asked to detect the intrusion words. In temporal word intrusion, we instead modify a temporal sequence of one topic k, $(v_k^{(1)}, \ldots, v_k^{(T)})$, where $v_k^{(t)}$ corresponds to the list of top words for topic k at time t. We can then analyze how our proposed measure of coherence over time changes based on

	D-LDA							D-ETM						
	year-wise			*temporal* (ours)				*year-wise*			*temporal* (ours)			
Dataset	TC	TD	TQ	TTC	TTS	TTQ	DTQ	TC	TD	TQ	TTC	TTS	TTQ	DTQ
NeurIPS	.08	.97	.08	.17	.94	.16	.12	.07	.97	.07	0.14	.72	.10	.09
NeurIPS*	.08	.97	.08	.00	.07	.00	.04	.07	.97	.07	.00	.06	.00	.03
NYT	.13	.96	.12	.20	.95	.19	.16	.13	.98	0.13	.15	.60	.12	.13
NYT*	.13	.96	.12	.00	.09	.00	.06	.13	.98	.13	.00	.05	.00	.06
UN Debates	.06	.94	.05	.15	.96	.15	.10	.06	.96	.06	.14	.82	.12	.09
UN Debates*	.06	.94	.05	.00	.08	.00	.02	.06	.96	.06	.00	.05	.00	.03
diff	→	→	→	↓	↓	↓	↓	→	→	→	↓	↓	↓	↓

Table 1: This table demonstrates that our temporal measures are able to capture temporal transitions, whereas the year-wise measures are not. NeurIPS* is a synthetic dataset where the original topics from the NeurIPS dataset are shuffled. On the shuffled topics, the temporal measures record lower scores as compared to the original topics, whereas the year-wise measures show unchanged values. This suggests that using only year-wise measures (TQ) to evaluate Dynamic Topic Models (DTM) is insufficient. The performance of D-LDA and D-ETM models are shown in terms of both *year-wise* (TC, TD, TQ) and *temporal* (TTC, TTS, TTQ) measures. These measures are computed based on the NPMI scores on three real-world datasets of NeurIPS, NYT, and UN General Debates. The arrow indicates a change in score when topics are shuffled.

the number of intruder words.[1]

5.2 Temporal Topic Rating

In the topic rating task, humans are presented with a topic and asked to rate it on a three-point ordinal scale. Similarly, we aim to examine the temporal sequence of a topic for *word-relatedness* and *smoothness*. Human annotators are asked to rate a topic sequence instead of one static topic. Word familiarity scores are also collected for the analysis. Fig. 3 shows how the task was presented to the human annotators.

6 Experiments

In this section, we establish the efficacy of the proposed measures using synthetic data and human evaluation. First, we compare static and temporal measures on synthetic topics in Section 6.2. Second, we investigate the sensitivity of the measures to noise in Section 6.3. Then, we conduct human evaluations of dynamic topics and compute the correlations of human ratings with the temporal measures in Section 6.4. As a window size parameter for our proposed measures TTS and TTC, we choose $L = 2$ in all experiments. Choosing higher window sizes would make the measure more sensitive to detecting slower transitions. However, sudden changes are of greater interest to us as they affect the interpretability of a topic over time more.

[1]Note, that in contrast to (Hoyle et al., 2021) we do not use human evaluation here, but study correlation between intrusion level and coherence directly.

6.1 Models and Corpora

We compare our proposed measures for the models D-LDA (Blei and Lafferty, 2006), D-ETM (Dieng et al., 2019) and D-LLM (Sia et al., 2020). D-LDA is a probabilistic model extending the popular LDA model to be dynamic. D-ETM is a neural DTM, which uses embeddings of words and topics. D-LLM is a dynamic version of the model by Sia et al. (Sia et al., 2020). To make the model dynamic, we train separate models on the data for each timestamp, initializing the cluster means with those from the previous time step. We use 50 topics as is common in the literature (Dieng et al., 2019; Hoyle et al., 2021) for all the models. We randomly select 80% of documents for training, 10% for testing, and 10% for validation.

We study our proposed evaluation measures using three commonly used datasets in the domain. The UN General Debates corpus (Jankin Mikhaylov et al., 2017) spans 51 years (1970 to 2020). It contains general debate statements from 1970 to 2020. The second dataset (Sandhaus, 2008) consists of New York Times articles spanning 21 years (1987 to 2007). The third dataset, the NeurIPS corpus (Swami, 2020), contains all NeurIPS papers from 1987 to 2019. Each dataset is preprocessed using standard techniques such as tokenization and removal of all punctuation and stop words (see Appendix A, B, and C for complete corpus statistics, preprocessing, and model training details).

6.2 Efficacy of Temporal Topic Evaluation

Table 1 compares the proposed temporal measures (TTC, TTS, TTQ, and DTQ) to static (year-wise) measures (TC, TD, and TQ). The results for D-LLM are shown in appendix H. We evaluate our proposed measures on three real-world corpora. Additionally, we construct synthetic topic models by shuffling the original topics of each model in each timestamp. This shuffling disrupts the topic transitions. The resultant synthetic data works as a proxy for the output of a DTM with poor topic transitions. An ideal evaluation measure is expected to capture the impact of shuffling on the topic transitions.

The evaluation results for the shuffled topics (corpus with *) are compared to the results for the unshuffled topics in Table 1. This reveals that the year-wise measures are unchanged regardless of the shuffling. However, the temporal measures of the shuffled topics are significantly lower than the temporal measures of the original topics in all three datasets. This suggests that the proposed temporal measures can consistently detect poor topic transitions. The results also emphasize that temporal changes are not reflected in the year-wise measures. Therefore, they are inadequate for evaluating dynamic topics.

Table 1 also shows that D-LDA and D-ETM have similar TQ, but exhibit different temporal behavior. D-LDA generally produces smoother topic transitions (higher TTS) than D-ETM. Furthermore, using the temporal evaluation measures, we can monitor changes in topics over time, as shown in Appendix D. Table 10 in Appendix E also shows why the combined measure DTQ is useful for evaluating a DTM. It examines the case where topic diversity is zero, which affects only DTQ, but not TTQ. D-LLM, a baseline proposed by us, has an overall low TC, but slightly higher TTC and TTS, indicating that there is some temporal consistency while the overall topic quality is low. Details on how these results were obtained can be found in Appendix H. It can be concluded that research on LLM-based dynamic topic models is needed.

6.3 Word Intrusion Assessment of Temporal Topics

We now investigate the effect of the noise intrusion level on our measure using the temporal word intrusion task presented in Section 5.1. For this task, a timestamp is randomly chosen from the temporal sequence of one randomly selected target topic for each model and corpus. Then, one randomly chosen word is replaced by a random intruder word in the selected topic. This process is repeated for ten intruder words. The number of intruder words determines the noise level chosen between one and ten. We compute Spearman's correlation on a ranking of both intrusion levels and temporal measures for each model and dataset.

	D-LDA			D-ETM		
Dataset	**TTC**	**TTS**	**TTQ**	**TTC**	**TTS**	**TTQ**
NeurIPS	0.97	0.98	0.98	0.95	0.91	0.91
NYT	0.95	0.98	0.94	0.98	0.91	0.96
UN	0.87	0.99	0.94	0.89	0.92	0.88

Table 2: This table shows that all our measures correlate well with temporal word intrusion. Shown is the Spearman's correlation for temporal word intrusion for three datasets. All the correlations yield more than 95% confidence intervals.

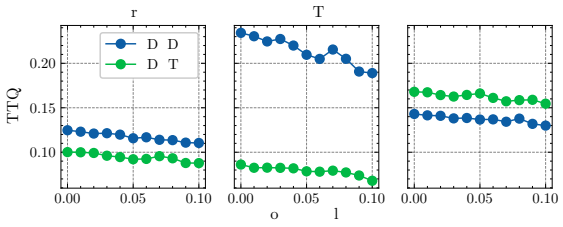

Figure 4: The figure shows temporal topic quality for different noise levels for one random topic (Topic 44) per dataset. The TTQ measure decreases continuously as more intrusion words are added to the topics for all three datasets.

As shown in Table 2, strong correlations are obtained for all datasets. This confirms that the proposed measures decrease as the noise level increases. Figure 4 shows this visually. Overall, the intrusion results underline the success of the proposed measures in distinguishing between low and high quality topics and are sensitive the level of intrusion.

6.4 Human Evaluation of Temporal Topics

In this section, we investigate the correlation between human evaluation scores and the proposed automated measures on a random sample of topics from each dataset. For this purpose, we use the temporal word-relatedness and smoothness tasks described in Section 5 and shown in Figure 3. We also show a correlation with other measures, such as the simple mean similarity across topics, by us-

ing topic embeddings.

We randomly select 20 topics from both models of each dataset for the human survey. For each topic, we present human raters with a sequence of five equally spaced time steps. We conduct a separate study for each corpus. We recruit crowd workers from Prolific.co and compensate them with the equivalent of 12 USD/hr. Each participant is provided with detailed instructions. We follow the protocol by Hoyle et al. (2021) and recruit a large number of crowd workers (18) per task to ensure adequate statistical power. Aggregate human ratings of word relatedness and smoothness for each topic are calculated by averaging across all valid respondents.

We use two criteria to identify valid respondents. The first is based on a control task. Respondents who fail the control task are excluded from the analysis. Topics in the control task are created synthetically by randomly selecting words, resulting in a very low-quality topic in terms of word relatedness and smooth transitions. Second, we monitor the time taken to complete the survey. We filter out outlier respondents based on the median time taken to complete the survey. These criteria for filtering out invalid respondents are consistent with previous studies in the field (Hoyle et al., 2021; Chang et al., 2009).

The scatter plots in Figure 5 show the correlation of human ratings and temporal measures for word-relatedness (TTC, bottom) and smoothness (TTS, top) respectively. Inspection of outlier points reveals that low human ratings and high temporal scores often belong to topics with low familiarity among raters. The figure shows that humans rated D-LDA topics higher as compared to D-ETM. It also shows that human ratings are more varied as compared to the automated measures.

No standard baseline exists for temporal topic evaluations. Existing work (Blei and Lafferty, 2006; Dieng et al., 2019) is limited to qualitative evaluations and delivers no quantitative measures. Following the existing evaluation measures, we derive baselines for topic coherence (B-TC) and topic smoothness (B-TS). The baseline measure for coherence is computed from the average score for topic coherence over time. For smoothness, we calculated $1 -$ diversity of one topic over time as the baseline which corresponds to TTS with maximum window size.

Table 3 shows that TTC correlates better with

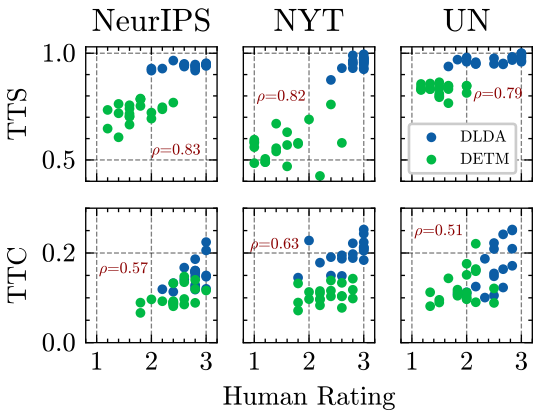

Figure 5: This figure shows that the temporal topic smoothness and coherence correlate well with the human evaluation. Correlation (Spearman's ρ) between human and temporal topic smoothness (top) and coherence (bottom) for random topics from NeurIPS (left), NYT (middle) and UN (right) datasets.

Dataset	TTC	B-TC	TTS	B-TS
NeurIPS	**0.57**	0.21	**0.83**	0.65
NYT	**0.63**	0.17	0.82	**0.83**
UN	**0.51**	-0.02	0.79	**0.81**

Table 3: This table shows that 1) TTC correlates better with human evaluation than B-TC 2) TTS correlates better or to a similar degree as B-TS with human smoothness evaluation. Spearman's correlation coefficients between mean human evaluation and automated measures. These correlations of proposed measures (TTC and TTS) are compared to the baseline measures (B-TC and B-TS). The highest correlations for each pair are shown in bold.

the human perception of word relatedness in DTMs than the baseline TC which does not consider temporal transitions. The proposed temporal measures consistently show a stronger correlation with human ratings than the baseline among all the datasets. The TTS also correlates well. However, the baseline topic smoothness also has a high correlation with human-perceived smoothness which is to be expected since it corresponds to TTS with maximum window size. This suggests that, depending on the dataset, the TTS measure is fairly robust with respect to the window size. All correlations obtained are in 95% confidence intervals.

Figures 7 and 8 in Appendix F show the results of the human and automated evaluation for word-relatedness and smoothness, respectively. These figures indicate that human ratings align with our proposed temporal measures.

Dataset	sim-wr	sim-sm
NeurIPS	0.36	0.44
NYT	0.48	0.87
UN	0.17	0.55

Table 4: This table shows a correlation between mean similarity measures and human evaluations, such as word relatedness (sim-wr) and smoothness (sim-sm).

Table 4 shows the correlation between the mean similarity measure and human evaluations, such as word relatedness (sim-wr) and smoothness (sim-sm). The mean similarity measure is calculated using the topic embedding over time, computed by averaging pretrained word embeddings. The results indicate a weak correlation with word relatedness (sim-wr) and typically a moderate correlation with smoothness (sim-sm). Furthermore, our method significantly outperforms these measures by a large margin.

Additionally, we assessed the inter-rater agreement. For the smoothness task, the mean Spearman correlation scores were 0.59 for Neurips, 0.68 for NYT, and 0.62 for UN Debates, indicating good agreement. In the word relatedness task context, the inter-rater agreement values are 0.18 for Neurips, 0.21 for NYT, and 0.23 for UN Debates.

7 Conclusion

This paper fills a gap in evaluating temporal characteristics of DTMs such as LLM-based dynamic models. We complement the existing year-wise measures by proposing novel temporal measures. Our proposed temporal measures capture different aspects of temporal topic changes in DTMs. We show that our measures are able to better capture temporal characteristics of topic changes than their year-wise counterparts and have positive correlation with human evaluations. We show the efficacy of our measure by evaluating different dynamic topic models and demonstrating their different temporal characteristics. Our proposed evaluation measures will improve future comparisons between DTMs, including LLM-based topic evaluations. In the future, we want to extend this method also to the evaluation of sequential decision-making (Li et al., 2023, 2024), other structured or online topic models (Ahmadi et al., 2017; Burkhardt and Kramer, 2017a,b) as well as the explicit discovery of temporal anomalies (Ruff et al., 2019; Liznerski et al., 2024).

Limitations

There are two main limitations to our approach. The first concerns the human evaluation. Here, we have to rely on the quality of the answers provided by the human annotators. Although we took care to recruit a large number of annotators (18) in order to reduce the variance of our results, it would be preferable to have fewer annotators providing high-quality annotations. This could only be achieved by training people before they are given the task, which would require a training protocol. This is beyond the scope of our study. This issue of human annotators not being experts in the domain of the dataset or the given task also affects other studies and is difficult to solve. Automated measures need to be validated against human annotations, but human annotations are never perfect. Therefore, there will always remain a gap.

The second limitation of automated topic evaluation, in general, is the reference corpus. Automated measures are calculated with respect to a reference corpus. If words or topics are not present in the reference corpus (for whatever reason), the result will be suboptimal. This is especially true for temporal topics, where the number of documents for selected time steps may be small, which could lead to the respective topics not being present in the reference corpus. This could be addressed in the future by selecting the reference corpus more carefully, or possibly by selecting an external reference corpus.

Acknowledgements

The authors were funded by the German Federal Ministry of Education and Research under grant number 01IS20048. Responsibility for the content of this publication lies with the author. Additionally, we acknowledge support from the Carl-Zeiss Foundation and the DFG awards BU 4042/2-1 and BU 4042/1-1.

References

Zahra Ahmadi, Sophie Burkhardt, and Stefan Kramer. 2017. Online topic modeling: Keeping track of news topics for social good. In *Proceedings of the Second Workshop on Data Science for Social Good at the European Conference on Machine Learning and Principles and Practice of Knowledge Discovery in Databases (ECML-PKDD)*.

Amr Ahmed and Eric Xing. 2008. Dynamic non-parametric mixture models and the recurrent chinese restaurant process: with applications to evolutionary

clustering. In *Proceedings of the 2008 SIAM international conference on data mining*, pages 219–230. SIAM.

Amr Ahmed and Eric P. Xing. 2012. Timeline: A dynamic hierarchical dirichlet process model for recovering birth/death and evolution of topics in text stream. *CoRR*, abs/1203.3463.

Nikolaos Aletras and Mark Stevenson. 2013. Evaluating topic coherence using distributional semantics. In *Proceedings of the 10th International Conference on Computational Semantics (IWCS 2013)–Long Papers*, pages 13–22.

Arnab Bhadury, Jianfei Chen, Jun Zhu, and Shixia Liu. 2016. Scaling up dynamic topic models. In *Proceedings of the 25th International Conference on World Wide Web*, pages 381–390.

Asmita Bhat, Nooshin Haji-Ghassemi, Deepak Nagaraj, and Sophie Fellenz. 2023. Constraint-based parameterization and disentanglement of aerodynamic shapes using deep generative models. In *Machine Learning and Knowledge Discovery in Databases (ECML PKDD)*.

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.

Sophie Burkhardt and Stefan Kramer. 2017a. Multi-label classification using stacked hierarchical dirichlet processes with reduced sampling complexity. In *ICBK 2017 - International Conference on Big Knowledge*, pages 1–8, Hefei, China. IEEE.

Sophie Burkhardt and Stefan Kramer. 2017b. Online sparse collapsed hybrid variational-gibbs algorithm for hierarchical dirichlet process topic models. In *Machine Learning and Knowledge Discovery in Databases*, pages 189–204, Cham. Springer International Publishing.

Sophie Burkhardt and Stefan Kramer. 2019. Decoupling sparsity and smoothness in the dirichlet variational autoencoder topic model. *J. Mach. Learn. Res.*, 20(131):1–27.

Sophie Burkhardt, Julia Siekiera, Josua Glodde, Miguel A Andrade-Navarro, and Stefan Kramer. 2020. Towards identifying drug side effects from social media using active learning and crowd sourcing. In *Pacific Symposium of Biocomputing (PSB)*, volume 25, pages 319–330. World Scientific.

Liangliang Cao and Li Fei-Fei. 2007. Spatially coherent latent topic model for concurrent segmentation and classification of objects and scenes. In *2007 IEEE 11th International Conference on Computer Vision*, pages 1–8. IEEE.

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.

Wang Chong, David Blei, and Fei-Fei Li. 2009. Simultaneous image classification and annotation. In *2009 IEEE Conference on computer vision and pattern recognition*, pages 1903–1910. IEEE.

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

Adji B Dieng, Francisco JR Ruiz, and David M Blei. 2020. Topic modeling in embedding spaces. *Transactions of the Association for Computational Linguistics*, 8:439–453.

Caitlin Doogan and Wray Buntine. 2021. Topic model or topic twaddle? re-evaluating semantic interpretability measures. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 3824–3848.

Avinava Dubey, Ahmed Hefny, Sinead Williamson, and Eric P Xing. 2013. A nonparametric mixture model for topic modeling over time. In *Proceedings of the 2013 SIAM international conference on data mining*, pages 530–538. SIAM.

Li Fei-Fei and Pietro Perona. 2005. A bayesian hierarchical model for learning natural scene categories. In *2005 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'05)*, volume 2, pages 524–531. IEEE.

Sean M Gerrish and David M Blei. 2011. Predicting legislative roll calls from text. In *Proceedings of the 28th International Conference on Machine Learning, ICML 2011*.

Zhinan Gou, Lixin Han, Ling Sun, Jun Zhu, and Hong Yan. 2018. Constructing dynamic topic models based on variational autoencoder and factor graph. *IEEE Access*, 6:53102–53111.

Maarten Grootendorst. 2022. BERTopic: Neural topic modeling with a class-based TF-IDF procedure. ArXiv:2203.05794 [cs].

Lin Gui, Jia Leng, Gabriele Pergola, Yu Zhou, Ruifeng Xu, and Yulan He. 2019. Neural Topic Model with Reinforcement Learning. 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 3478–3483, Hong Kong, China. Association for Computational Linguistics.

Yulan He, Chenghua Lin, Wei Gao, and Kam-Fai Wong. 2013. Dynamic joint sentiment-topic model. *ACM Transactions on Intelligent Systems and Technology*, 5(1):1–21.

Matthew Honnibal and Ines Montani. 2017. spaCy 2: Natural language understanding with Bloom embeddings, convolutional neural networks and incremental parsing. To appear.

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.

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.

Slava Jankin Mikhaylov, Alexander Baturo, and Niheer Dasandi. 2017. United Nations General Debate Corpus.

Weichen Li, Rati Devidze, and Sophie Fellenz. 2023. Learning to play text-based adventure games with maximum entropy reinforcement learning. In *Machine Learning and Knowledge Discovery in Databases (ECML PKDD)*.

Weichen Li, Rati Devidze, Waleed Mustafa, and Sophie Fellenz. 2024. Ethics in action: Training reinforcement learning agent for moral decision-making in text-based adventure games. In *Proceedings of International Conference on Artificial Intelligence and Statistics (AISTATS)*.

Zachary C Lipton. 2018. The mythos of model interpretability: In machine learning, the concept of interpretability is both important and slippery. *Queue*, 16(3):31–57.

Philipp Liznerski, Saurabh Varshneya, Ece Calikus, Sophie Fellenz, and Marius Kloft. 2024. Reimagining anomalies: What if anomalies were normal?

Laura Manduchi, Kushagra Pandey, Robert Bamler, Ryan Cotterell, Sina Däubener, Sophie Fellenz, Asja Fischer, Thomas Gärtner, Matthias Kirchler, Marius Kloft, Yingzhen Li, Christoph Lippert, Gerard de Melo, Eric Nalisnick, Björn Ommer, Rajesh Ranganath, Maja Rudolph, Karen Ullrich, Guy Van den Broeck, Julia E Vogt, Yixin Wang, Florian Wenzel, Frank Wood, Stephan Mandt, and Vincent Fortuin. 2024. On the challenges and opportunities in generative ai.

Andrew McCallum, Andrés Corrada-Emmanuel, and Xuerui Wang. 2005. The author-recipient-topic model for topic and role discovery in social networks: Experiments with enron and academic email. *Computer Science Department Faculty Publication Series*, page 44.

Qiaozhu Mei and ChengXiang Zhai. 2005. Discovering evolutionary theme patterns from text: An exploration of temporal text mining. In *Proceedings of the Eleventh ACM SIGKDD International Conference on Knowledge Discovery in Data Mining*, KDD '05, page 198–207, New York, NY, USA. Association for Computing Machinery.

Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg S Corrado, and Jeff Dean. 2013. Distributed representations of words and phrases and their compositionality. *Advances in neural information processing systems*, 26.

David Mimno, Hanna Wallach, Edmund Talley, Miriam Leenders, and Andrew McCallum. 2011. Optimizing semantic coherence in topic models. In *Proceedings of the 2011 conference on empirical methods in natural language processing*, pages 262–272.

Satoshi Morinaga and Kenji Yamanishi. 2004. Tracking dynamics of topic trends using a finite mixture model. In *Proceedings of the Tenth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD '04, page 811–816, New York, NY, USA. Association for Computing Machinery.

Mayank Nagda and Sophie Fellenz. 2024. Putting back the stops: Integrating syntax with neural topic models. In *Proceedings of the 33rd International Joint Conference on Artificial Intelligence (IJCAI)*. to appear.

Mayank Kumar Nagda, Charu James, Marius Kloft, and Sophie Burkhardt. 2021. Hierarchical topic evaluation: Statistical vs. neural models. In *Bayesian Deep Learning Workshop at NeurIPS*.

David Newman, Jey Han Lau, Karl Grieser, and Timothy Baldwin. 2010. Automatic evaluation of topic coherence. In *Human language technologies: The 2010 annual conference of the North American chapter of the association for computational linguistics*, pages 100–108.

Phil Ostheimer, Mayank Nagda, Marius Kloft, and Sophie Fellenz. 2024. Text style transfer evaluation using large language models. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-Coling)*.

Phil Ostheimer, Mayank Kumar Nagda, Marius Kloft, and Sophie Fellenz. 2023. A call for standardization and validation of text style transfer evaluation. In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 10791–10815, Toronto, Canada. Association for Computational Linguistics.

Jonathan K Pritchard, Matthew Stephens, and Peter Donnelly. 2000. Inference of population structure using multilocus genotype data. *Genetics*, 155(2):945–959.

Daniel Ramage, Christopher D Manning, and Susan Dumais. 2011. Partially labeled topic models for interpretable text mining. In *Proceedings of the 17th ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 457–465.

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.

Lukas Ruff, Yury Zemlyanskiy, Robert Vandermeulen, Thomas Schnake, and Marius Kloft. 2019. Self-attentive, multi-context one-class classification for unsupervised anomaly detection on text. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 4061–4071, Florence, Italy. Association for Computational Linguistics.

Evan Sandhaus. 2008. The New York Times Annotated Corpus.

Suzanna Sia, Ayush Dalmia, and Sabrina J. Mielke. 2020. Tired of topic models? clusters of pretrained word embeddings make for fast and good topics too! In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1728–1736, Online. Association for Computational Linguistics.

Dominik Stammbach, Vilém Zouhar, Alexander Hoyle, Mrinmaya Sachan, and Elliott Ash. 2023. Re-visiting automated topic model evaluation with large language models. *arXiv preprint arXiv:2305.12152*.

Rohit Swami. 2020. All neurips (nips) papers.

Chong Wang, David M Blei, and David Heckerman. 2008. Continuous time dynamic topic models. In *UAI'08 Proceedings of the Twenty-Fourth Conference on Uncertainty in Artificial Intelligence*.

Xuerui Wang and Andrew McCallum. 2006. Topics over time: a non-markov continuous-time model of topical trends. In *Proceedings of the 12th ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 424–433.

Xuerui Wang, Andrew McCallum, and Xing Wei. 2007. Topical n-grams: Phrase and topic discovery, with an application to information retrieval. In *Seventh IEEE international conference on data mining (ICDM 2007)*, pages 697–702. IEEE.

Bin Zheng, David C McLean, and Xinghua Lu. 2006. Identifying biological concepts from a protein-related corpus with a probabilistic topic model. *BMC bioinformatics*, 7(1):1–10.

Houkui Zhou, Huimin Yu, and Roland Hu. 2017. Topic evolution based on the probabilistic topic model: a review. *Frontiers of Computer Science*, 11(5):786–802.

A Dataset Details

We study our evaluation measure using three different datasets: NeurIPS, New York Times, and UN General Debates. The table 5 shows corpus statistics for the datasets.

Dataset	**NeurIPS**	**NYT**	**UN**
Domain	Science	News	Politics
Number of Docs	9,679	274,665	8,481
Vocab Size	3,102	8,240	3,005

Table 5: Corpus Statistics. Shows the dataset that varies in the domain, document number, and vocab size. The NeurIPS is from (Swami, 2020), NYT is from (Sandhaus, 2008) and UN General Debates is from (Jankin Mikhaylov et al., 2017).

B Preprocessing Details

The datasets undergo a series of preprocessing steps, which include converting the text to lowercase, eliminating stopwords, and removing punctuation marks. Tokenization is performed using Spacy (Honnibal and Montani, 2017), and further refinement involves removing words that appear in less than a specified percentage (min_df) of the documents, as well as words that occur in more than a specified percentage (max_df) of the documents, using count vectorizer. The min_df for the New York Times (NYT) dataset was determined to be 0.3% after observing that a min_df of 5% yielded a vocabulary size of 564, which was considered insufficient for a large dataset. Table 6 shows the cut-off parameters used for the different datasets.

Dataset	**min_df**	**max_df**	**vocab size**
NeurIPS	5%	95%	3,102
UN Debates	5%	95%	3,005
NYT	0.3%	95%	9,046

Table 6: The table shows the vocab size that results from removing words that occur in less than min_df percent of the documents and words that occur in more than max_df percent of the documents.

C Model Training Details

Expanding subsection 6.1, here we explain how hyperparameters are set for each topic model.

D-LDA For training the model, we use the Gensim python wrapper for dynamic topic models (DTM). We slice all datasets by year. As a result, every time slice contains all documents from that year. For each dataset, we ran 50 iterations with an alpha value of 0.01, which is a hyperparameter affecting the sparsity of the document topics for each time slice in the LDA models. In addition, we used top_chain_var values of 0.005.

D-ETM Using the skip-gram model, a 300-dimensional word embedding is obtained (Mikolov et al., 2013). The batch size for all datasets was 100 documents. We used the perplexity score on the validation set as stopping criteria for all datasets. The learning rate is set to a default value of 0.001. The hyperparameters delta, sigma, and gamma in D-ETM are set to 0.005 as suggested by the authors. A random selection of 80% of documents is used for training, 10% for testing, and 10% for validation.

D Qualitative results

A topic's temporal topic quality is determined by how smoothly its words change over time. The temporal topic quality is calculated in terms of temporal topic coherence (TTC) and temporal topic smoothness (TTS). The concept of TTS is shown in the top row of Figure 6. Topic 19 (top-left) illustrates an instance of topic words exhibiting smooth transitions. During the year 1988-1989, the TTS remains 1.0, indicating a lack of change in the topic words. Table 7 provides a depiction of Topic 19 during this time frame. Furthermore, in the scenario where consecutive TTS score reach 1.0, the TTC score remains unchanged, as the topic words have not undergone any changes.

Table 8 provides empirical evidence within the D-ETM model, demonstrating that Topic 8 experienced a relatively low TTS score between 1998-1999. Notably, despite the decline in TTS score, the corresponding topic remained largely unchanged, as indicated by the nearly same TTC score that did not exhibit a significant decrease. The same analysis applies to Topic 21, which represents a topic with drastic changes in its words. The TTS and TTC scores are observed to be low between 1992-1993, indicating a radical shift in the topic. Table 9 shows topic 21 during this time. As the table shows there is a change from the topic of rule extraction to a topic on image object recognition at this point in time. This change can also be seen in temporal topic quality (TTQ) in Figure 6 wherein 1992-1993 the TTQ score is low.

The temporal topic coherence is calculated based on NPMI for D-LDA and D-ETM. Whereas D-LDA in general shows relatively unchanged temporal topic coherence over time, D-ETM exhibits more variance in TTC.

Year	words in Topic 19	TTS	TTC
1988	connectionist human figure systems research science knowledge performance target rules	1.0	0.098
1989	connectionist figure human rules knowledge target performance science research systems	1.0	0.098
1990	figure connectionist rules human target knowledge performance science research information	0.9	0.093

Table 7: Topic 19 from D-LDA model using NeurIPS dataset, which shows the smoothness in topic during the year 1988-89, when TTS in Figure 6 is 1.0, which is between the current year and previous year.

Year	words in Topic 8	TTS	TTC
1996	position hand task user location based object body target robot	0.7	0.127
1997	object position hand task robot user direction location right coordinates	0.6	0.126
1998	position hand human line movement direction motor task object location	0.3	0.105
1999	spatial localization location light position human temporal subjects robot subject	0.3	0.104

Table 8: Topic 8 from the D-ETM model using the NeurIPS dataset, which shows low smoothness in 1998–1999, but TTC remains nearly the same. This is shown in Figure 6 where the TTS score is 0.3 which is between the current year and previous year.

E Efficacy of year-wise Topic Evaluation

In this section, we establish the efficacy of year-wise topic evaluation measures for evaluating DTMs. To this end, we construct synthetic topics which behave as proxies to output of a poor topic model. For the synthetic topics, we randomly select one topic from each model and dataset and repeat it while removing all other topics. The synthetic topics work as an extreme case of component collapse in a DTM. The results of this experiment are shown in Table 10. For all the synthetic versions of the three datasets, year-wise TQ is zero (because TD is zero). Hence, the overall DTQ is low as compared to TTQ. This establishes the effi-

Year	words in Topic 21	TTS	TTC
1990	rules rule cell extraction group clustering groups cluster expert clusters	0.7	0.099
1991	rules rule extraction extracted group expert groups clustering induction self	0.5	0.068
1992	rules rule children expert extraction features view self image feature	0.2	0.027
1993	image surface view recognition object matching images correspondence views objects	0.8	0.190

Table 9: Topic 21 from D-ETM model using NeurIPS dataset, which shows the change in topic during the year 1992-1993, when TTS in Figure 6 is 0.20 and TTC is 0.027 which is between the current year and previous year.

cacy of combining both TTQ and TQ in the form of DTQ when evaluating DTMs.

F Human Evaluations

In this section, we report the results of the human rating survey. We show the results of the automatic and human ratings of the randomly selected 20 topics from each model and dataset in Figure 7 and 8 for word relatedness and smoothness, respectively. The average human ratings from the survey are consistent and in line with the previous studies (Hoyle et al., 2021; Röder et al., 2015).

Furthermore, the instruction provided to human for rating task is shown in Figure 9. The figure depicts a sequence of words list, serving as a sample for establishing the definitions of word-relatedness and smooth transitions within the context of the study. Prior to the start of the survey, participants were briefed on what data would be collected and for what purpose it would be used. And during the course of this study, no personal data was collected.

G Word Intrusion Assessment of TTQ

In this section, we continue the assessment of temporal topic quality w.r.t the intrusion levels as discussed in Section 6.3. The result is shown in Figure 4. A consistent decrease in TTQ is observed for both models, with an increase in intrusion levels. This relationship is also backed by the correlations discussed in Table 2. In all the cases, a strong correlation can be observed. From the intrusion task, we conclude that the TTQ measure is adequate in measuring even small changes in temporal topic quality.

H Dynamic Cluster Model utilizing a pre-trained LLM

D-LLM The Dynamic Large Language Model (D-LLM) was formulated based on the methodology outlined by Sia et al. (Sia et al., 2020). However, we adapted the static model to a dynamic setting, where the initialization of a model in each timestamp is done using the cluster mean from the previous timestamp. Following Sia et al., we used word embedding representations derived from a large language model (we used "paraphrase-distilroberta-base-v1", a variant of the sentence-transformer architecture). We then applied a Gaussian mixture model (GMM) to obtain $k = 50$ clusters and used TF-IDF for reranking. With a convergence threshold set at 1e-3, the GMM algorithm executes 100 iterations for the Expectation-Maximization (EM) procedure. In contrast to the methodology employed by Sia et al., our approach uniformly applies 50 topics across all datasets. Furthermore, with a window size of 10, we utilize Normalized Pointwise Mutual Information (NPMI) for our evaluation metics. As a result of adapting the static model to a dynamic context, the number of documents accessible for training per year has been reduced compared to static model.

Table 11 shows the results. In particular, the temporal topic smoothness is lower compared to D-LDA and D-ETM. Additionally, the overall DTQ score is also low. TTC and TTS show a limited temporal consistency of topics, which is confirmed by the qualitative examination of the topics. As these results show, our proposed baseline for LLM-based dynamic topic models, the D-LLM model, is not competitive. Clearly, further research is needed to be able to extract dynamic topics from LLMs.

I Abbreviations

The abbreviations used throughout this paper are detailed in Table 12. This table provides a comprehensive description of each abbreviation to ensure clarity and ease of understanding for the reader.

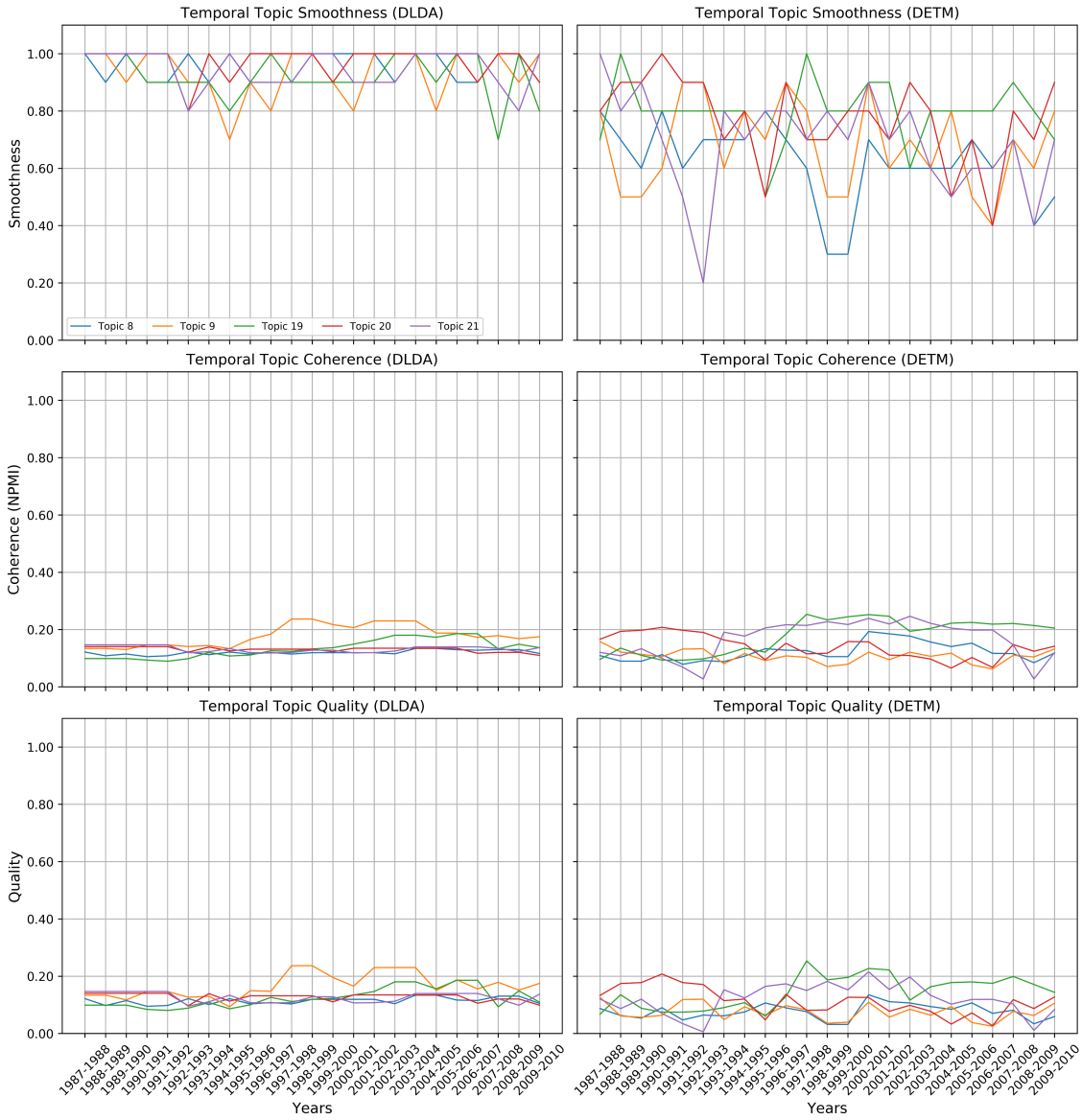

Figure 6: Shows how **temporal topic smoothness** changes over year for five topics generated by D-LDA (top-left) and D-ETM (top-right) for NeurIPS dataset. Shows how **temporal topic coherence** changes over year for five topics generated by D-LDA (middle-left) and D-ETM (middle-right) based on **NPMI score**. Shows how **temporal topic quality** (TTQ) changes over year for five topics generated by D-LDA (bottom-left) and D-ETM (bottom-right).

	D-LDA							D-ETM						
	year-wise			*temporal* (ours)				*year-wise*			*temporal* (ours)			
Dataset	TC	TD	TQ	TTC	TTS	TTQ	DTQ	TC	TD	TQ	TTC	TTS	TTQ	DTQ
NeurIPS	.08	.97	.08	.17	.94	.16	.12	.07	0.97	.07	.14	.72	.10	.09
NeurIPS*	.05	.00	.00	.13	.94	.12	.06	.17	.00	.00	.11	.78	.08	.04
NYT	.13	.96	.12	.20	.95	.19	.16	.13	.98	.13	.15	.60	.12	.13
NYT*	.18	.00	.00	.21	.94	.20	.10	.19	.00	.00	.09	.71	.07	.03
UN Debates	.06	.94	.05	.15	.96	.15	.10	.06	.96	.06	.14	.82	.12	.09
UN Debates*	.01	.00	.00	.06	.95	.06	.03	.12	.00	.00	.09	.79	.08	.04
diff	→	↓	↓	→	→	→	↓	→	↓	↓	→	→	→	↓

Table 10: This table demonstrates the need of having year-wise measures in the DTQ. We construct synthetic datasets (marked with *) where we randomly select a topic and repeat it. The synthetic topics now work as proxy for the extreme case of component collapse in case of DTMs. On the synthetic topics, the year-wise measures record lower TQ scores as compared to the original topics, whereas the temporal measures show similar values.

	D-LLM						
	year-wise			*temporal* (ours)			
Dataset	TC	TD	TQ	TTC	TTS	TTQ	DTQ
NeurIPS	-.017	.904	-.015	.004	.256	.011	-.002
NeurIPS*	-.017	.904	-.015	.000	.060	.000	-.007
NYT	-.009	.964	-.009	-.007	.197	.014	.003
NYT*	-.009	.964	-.009	.000	.072	.000	-.004
UN Debates	-.015	.883	-.013	.004	.232	.009	-.002
UN Debates*	-.015	.883	-.013	.000	.008	.000	-.006
diff	→	→	→	↓	↓	↓	↓

Table 11: The table shows the performance of the Dynamic Large Language Model (D-LLM) in terms of both *year-wise* (TC, TD, TQ) and *temporal* (TTC, TTS, TTQ) measures. These measures are computed based on the NPMI scores on three real-world datasets of NeurIPS, NYT, and UN General Debates. The synthetic datasets is marked with *, where we randaomly select a topic and repeat it. Compared to the original topics, the temporal measures result in lower scores on the shuffled topics, whereas the year-wise measures remain unchanged.

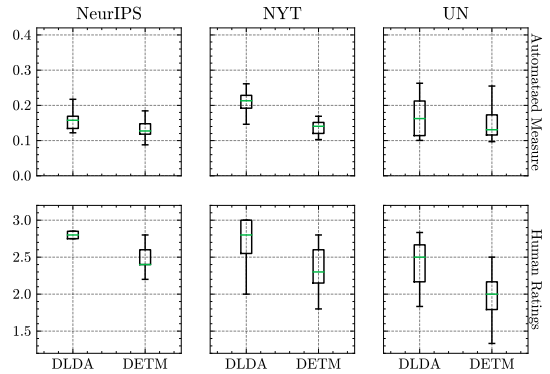

Figure 7: The mean and variance for the automated measure of TTC and human evaluation results for the three datasets of (right) NeurIPS, (middle) NYT and (right) UN.

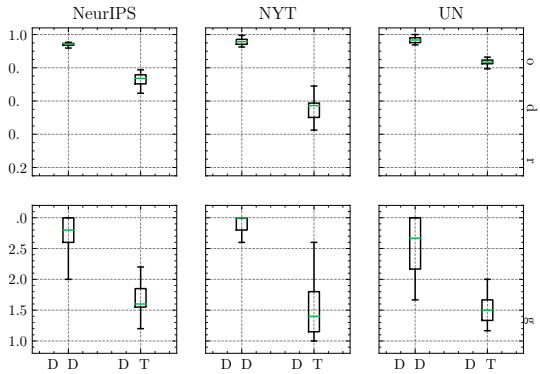

Figure 8: The mean and variance for the automated measure of TTS and human evaluation results for the three datasets of (right) NeurIPS, (middle) NYT and (right) UN.

Abbreviation	Description
DTM	Dynamic Topic Model
DTQ	Dynamic Topic Quality
D-ETM	Dymaic Embedded Topic Model
D-LDA	Dynamic Latent Dirichlet Allocation
D-LLM	Dynamic Large Language Model
NPMI	Normalized Pointwise Mutual Information
TC	Topic Coherence
TD	Topic Diversity
TO	Topic Overlap
TQ	Topic Quality
TTC	Temporal Topic Coherence
TTS	Temporal Topic Smoothness
TTQ	Temporal Topic Quality

Table 12: Abbreviations and their Descriptions

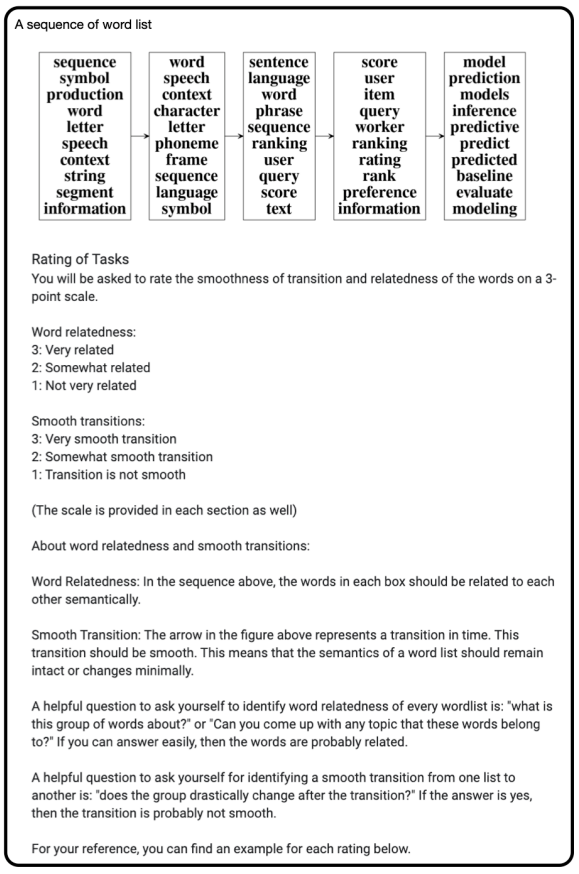

Figure 9: The instructions provided to human participants engaged in topic rating tasks.