

Don't Generate, Just Score: Enhancing Topic Coherence with Masked Semantic Alignment

Tien-Phat Nguyen^{1*}, Minh Chu Xuan^{1*}, Phung Gia Huy^{1*}, Linh Ngo Van^{1†},
Duc Anh Nguyen¹, Trung Le², Thien Huu Nguyen³

¹Hanoi University of Science and Technology, Hanoi, Vietnam

²Department of Data Science & AI, Monash University, Australia

³Department of Computer Science, University of Oregon, USA

Abstract

Neural topic models are widely used to uncover latent thematic structure in text corpora, but often produce topics with limited semantic coherence and weak alignment between topics and documents. Recent refinement approaches based on large language models (LLMs) have shown promising improvements, yet typically rely on repeated generative inference, resulting in high computational cost and practical issues such as sensitivity to prompting, stochastic variability, and the introduction of out-of-vocabulary content. We propose **BERT-NTM**, an encoder-based semantic refinement framework for neural topic modeling that avoids generative inference altogether. Instead of generating or rewriting topic words, BERT-NTM leverages a pretrained BERT encoder to provide discriminative semantic signals during a refinement phase: topic-word distributions are refined by scoring candidate words within a topic using masked language modeling, and document-topic representations are improved by aligning document embeddings with mixtures of topic embeddings via a Maximum Mean Discrepancy (MMD) objective. Both refinement components are fully parallelizable, stable, and operate strictly within the model vocabulary. Experiments on standard topic modeling benchmarks demonstrate that BERT-NTM consistently improves topic coherence and document-topic alignment while substantially reducing runtime and GPU memory usage compared to LLM-based refinement methods. Our code is publicly available at <https://github.com/phunggiahuy159/bert-ntm>.

1 Introduction

Topic models uncover latent thematic structure in document collections by representing each topic as a distribution over words and each document

as a mixture of topics. This provides an interpretable view of text corpora and supports applications such as exploratory analysis, document clustering, information retrieval, and summarization. Classical probabilistic models, most notably Latent Dirichlet Allocation (LDA) (Blei et al., 2003), formalize this framework through generative assumptions over topics, words, and documents. Neural topic models extend this foundation with neural networks, typically using amortized variational inference and flexible encoder-decoder architectures (Srivastava and Sutton, 2017; Dieng et al., 2020). They preserve the core topic-document-word structure while improving scalability and modeling capacity, and have shown strong performance across diverse domains and large-scale corpora (Bianchi et al., 2021a). A substantial line of subsequent work strengthens neural topic models from the inside, by adding regularizers or auxiliary objectives during training (Pham et al., 2024b; Le et al., 2025; Nguyen et al., 2025d), by replacing the reconstruction geometry with optimal transport and contrastive terms (Zhao et al., 2021; Vuong et al., 2025), or by adapting the model to specific data regimes such as short texts (Nguyen et al., 2025a; Vu et al., 2025) and multilingual corpora (Nguyen et al., 2025b; Xuan et al., 2026). Our work is complementary to this line: rather than changing how a topic model is trained, we ask how the topics it has already produced can be refined afterwards. We survey these directions in detail in Appendix A.

While large language models (LLMs) have substantially advanced semantic understanding in natural language processing, topic models continue to play an important role in corpus-level text analysis due to their interpretability and ability to capture global thematic structure. To address the limited semantic expressiveness of traditional topic models, recent studies have explored incorporating LLMs into the topic modeling pipeline, leveraging their rich contextual and world knowledge to enhance

*Equal contribution.

†Corresponding author: linhnv@soict.hust.edu.vn.

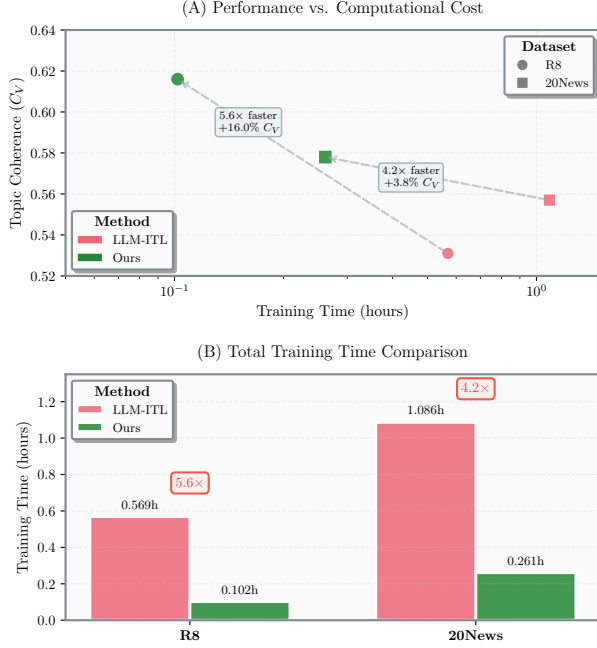


Figure 1: BERT-NTM achieves better topic coherence with significantly lower training cost compared to LLM-ITL. Results are reported on R8 and 20News in terms of C_V and total training time in GPU A40.

topic representations (Wang et al., 2023a; Pham et al., 2024a; Mu et al., 2024; Doi et al., 2024). By introducing external semantic signals, these approaches have demonstrated notable improvements in topic coherence and interpretability. Consequently, integrating LLMs into topic modeling has emerged as a prominent and actively explored research direction. However, despite their strong results, LLM-based topic modeling approaches involve important practical trade-offs. Early methods typically refine topics at the document level, requiring separate inference for individual documents and relying on local context for topic generation or refinement (Rijcken et al., 2023; Wang et al., 2023a). This is computationally expensive, scales poorly to large corpora, and makes it harder to capture coherent global topics. More recent frameworks shift refinement to the topic level, reducing the number of LLM calls and better preserving corpus-level structure (Yang et al., 2025b). Still, they depend on repeated inference with large generative models, which remains costly in practice. In addition, LLM-based refinement is sensitive to prompt design and contextual variation, while stochastic decoding can lead to unstable outputs (Pham et al., 2024a). Because topic models assume a fixed vocabulary, generative refinement

may also introduce out-of-vocabulary words that fall outside the model’s lexical space, as observed in recent iterative topic labeling methods (Yang et al., 2025b). These limitations motivate alternative refinement strategies that are both more efficient and more stable. We quantify LLM-ITL’s vocabulary mismatch before its filtering step, together with invalid refusal-style generations that require regeneration, in Appendix G.2. To address the high cost and instability of generative LLM-based refinement, we propose **BERT-NTM**, an encoder-based refinement framework for neural topic modeling. Instead of relying on text generation, BERT-NTM uses a pretrained encoder language model to inject semantic information through discriminative scoring. This enables efficient, fully parallel evaluation over the topic vocabulary, substantially reducing computational overhead and GPU memory usage compared with generative refinement, as shown in Figure 1.

At a high level, BERT-NTM performs semantic refinement in a dedicated phase by jointly refining topic representations and document–topic assignments. Encoder-derived semantic scores guide topic–word distributions toward greater coherence, while document–topic proportions are encouraged to align with more consistent topic assignments. Because refinement is restricted to the original topic-model vocabulary, BERT-NTM avoids introducing out-of-vocabulary or corpus-external terms that may distort topic structure. Overall, it provides a more efficient and stable alternative to generative LLM-based refinement for large-scale topic modeling.

We summarize our contributions as follows:

- We propose **BERT-NTM**, a neural topic modeling framework that improves topic coherence via BERT-based masked language modeling.
- We introduce a **BERT-based Semantic Scoring** that quantifies how well candidate topic words fit a topic-induced context, addressing polysemy and semantic drift.
- We formulate a lightweight semantic alignment objective that guides topic representations during neural topic model training.
- Extensive experiments show our approach reduces GPU usage and runtime while improving topic interpretability.

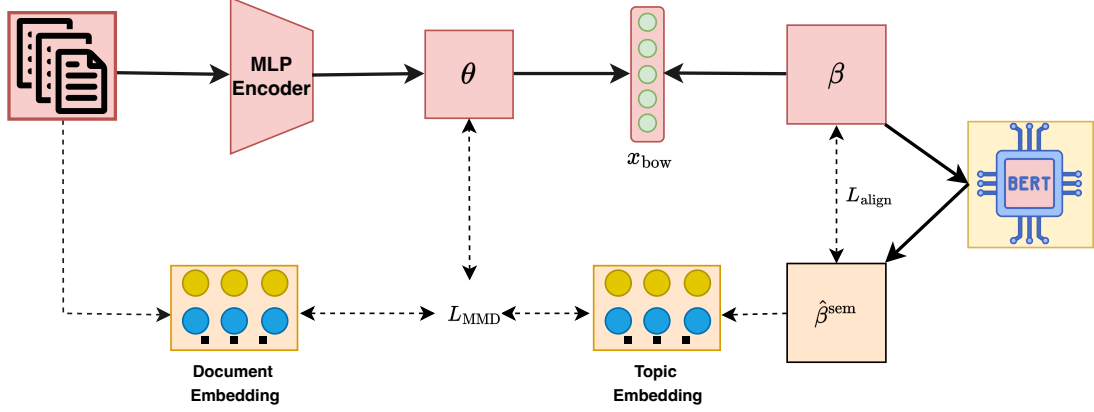


Figure 2: Overview of the **BERT-NTM** framework. A neural topic model encodes documents into topic proportions θ and topic-word distributions β . A pretrained BERT model provides topic-conditioned semantic signals to produce a semantic topic-word distribution $\hat{\beta}^{\text{sem}}$, which is aligned with β at the topic level. In parallel, document and topic embeddings are aligned via an MMD-based loss to encourage semantically consistent document-topic assignments.

2 Preliminaries

2.1 Notations

We consider a corpus consisting of D documents with the goal of discovering K latent topics. The corpus is represented as a collection $X = \{x_d\}_{d=1}^D$ of Bag-of-Words (BoW) representations, where each document $x_d \in \mathbb{R}^{|V|}$ corresponds to a document in the corpus and V denotes the vocabulary of size $|V|$. The topic-word distribution is denoted by $\beta \in \mathbb{R}^{|V| \times K} = (\beta_1, \dots, \beta_K)$, where each $\beta_k \in \mathbb{R}^{|V|}$ represents the word distribution of topic k over the vocabulary V , satisfying $\sum_{v \in V} \beta_{v,k} = 1$. Each document x_d is associated with a topic proportion vector $\theta_d \in \mathbb{R}^K$, where $\sum_{k=1}^K \theta_{d,k} = 1$.

2.2 VAE-based Topic Model

We use a variational autoencoder (VAE) as the backbone of the topic model. Each document, represented as a Bag-of-Words vector x_d , is encoded into a Gaussian variational posterior $q(z | x_d) = \mathcal{N}(z | \mu, \Sigma)$. A latent variable z is sampled via the reparameterization trick (Kingma and Welling, 2013) under a Gaussian prior. The latent representation is mapped to document-specific topic proportions θ_d through a softmax transformation. The decoder reconstructs the document using the topic-word distributions $\beta \in \mathbb{R}^{|V| \times K}$, which may be directly learned or parameterized with word embeddings (Srivastava and Sutton, 2017; Dieng et al., 2020; Wu et al., 2023), assuming a multinomial likelihood. The training objective follows the standard VAE formulation for neural topic models, consisting of a reconstruction term and a regular-

ization term:

$$\mathcal{L}_{\text{TM}} = \frac{1}{D} \sum_{d=1}^D \left(-x_d^\top \log \text{softmax}(\beta \theta_d) + \text{KL}(q(z | x_d) \| p(z)) \right).$$

2.3 Maximum Mean Discrepancy (MMD)

Maximum Mean Discrepancy (MMD) measures distributional differences by embedding probability distributions into a reproducing kernel Hilbert space (RKHS) and comparing their kernel mean embeddings (Gretton et al., 2012; Muandet et al., 2017). Given a characteristic kernel k , the squared MMD between distributions P and Q is defined as

$$\begin{aligned} \text{MMD}_k^2(P, Q) &= \mathbb{E}_{x, x' \sim P} k(x, x') \\ &+ \mathbb{E}_{y, y' \sim Q} k(y, y') \\ &- 2\mathbb{E}_{x \sim P, y \sim Q} k(x, y). \end{aligned}$$

which is equivalent to the squared distance between their kernel mean embeddings in RKHS.

For weighted empirical samples, MMD admits a differentiable quadratic estimator with respect to both sample locations and weights, making it suitable as an alignment objective in neural models. In embedding spaces, Gaussian kernels with bandwidth selected via the median heuristic or multi-scale variants are commonly used for robustness (Garreau et al., 2017). MMD is also closely related to energy distance through distance-induced kernels (Sejdinovic et al., 2013). MMD has been used in topic modeling before: W-LDA (Nan et al., 2019) matches the aggregated posterior of the document-topic proportions to a Dirichlet prior

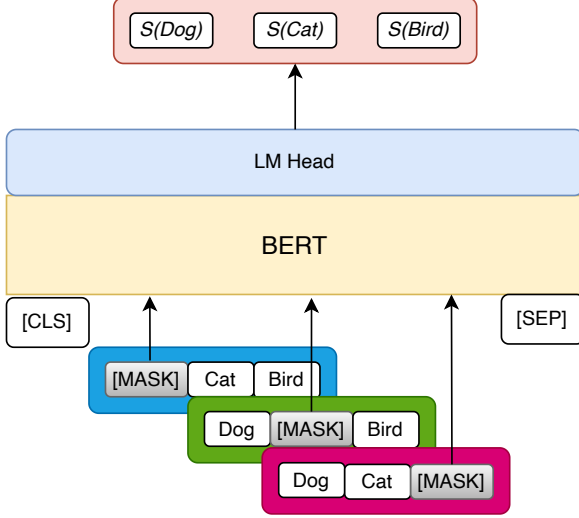


Figure 3: BERT-based semantic scoring for a topic. Each candidate word is masked within a topic-induced context and scored using BERT’s masked language modeling head to obtain a topic-conditioned semantic compatibility score.

on the simplex. Our use differs in target and space — we match a per-document embedding against its own topic-mixture measure in a pretrained semantic space — and we discuss the distinction in Appendix H.6.

3 Method

We begin with a standard neural topic model that captures corpus-level co-occurrence patterns through topic–word distributions and document–topic proportions. Building on these representations, we introduce a semantic refinement pipeline that brings contextual semantic knowledge into the topic modeling process. The pipeline first derives topic-conditioned semantic scores from a pretrained encoder language model, providing a semantic view of how words are expected to appear within a topic context. These scores are then used to guide the refinement of topic–word distributions and to align document–topic assignments, allowing semantic information to be propagated throughout the model. An overview of the proposed framework is illustrated in Figure 2. The remainder of this section details each component of this refinement pipeline.

3.1 BERT-based Semantic Scoring

We first describe how we extract topic-conditioned semantic signals from a pretrained BERT model and represent them as a semantic distribution over

the vocabulary. This signal will later be used to guide refinement of both topic–word and document–topic representations. Figure 3 illustrates the BERT-based semantic scoring procedure used in this subsection.

Topic-induced context. Given a topic k with topic–word distribution β_k , we select a set of representative context words $\mathcal{C}_k$ (e.g., the top- N words ranked by β_k). To evaluate candidate words under a topic-specific semantic context, we construct a deterministic input template consisting of a fixed prefix and the topic words:

$$S_k(\cdot) = [\text{CLS}] \text{ Words in this topic:}$$

$$c_{k,1}, \dots, c_{k,i-1}, [\text{MASK}], \\ c_{k,i+1}, \dots, c_{k,N} [\text{SEP}].$$

where one position in the sequence is replaced by a mask token to score a candidate word, and the ordering of context words is fixed (e.g., by descending β_k).

Token-level MLM scoring. Let w be a candidate vocabulary word. BERT’s tokenizer maps w to a sequence of m subword tokens $\tau(w) = (t_1, \dots, t_m)$. We score w by replacing the masked position in (3.1) with m consecutive [MASK] tokens (one for each subword in $\tau(w)$) and querying BERT’s masked language modeling head. For each subword token t_i , we obtain its probability at the i -th mask position. Denote by $p_{\text{BERT}}(t \mid S_k(\cdot))$ the MLM probability of token t at its corresponding masked position conditioned on the topic context. We define a token-level log-score

$$\ell_k(t) = \log p_{\text{BERT}}(t \mid S_k(\cdot)).$$

From subwords to word scores. To obtain word-level scores that are comparable across words with different numbers of subword tokens, we aggregate token-level scores using length normalization:

$$s_k(w) = \frac{1}{|\tau(w)|} \sum_{t \in \tau(w)} \ell_k(t).$$

This aggregation yields a length-invariant word score and avoids systematic bias toward words that are tokenized into fewer subwords.

Semantic distribution over candidate words. Finally, we convert the word scores into a topic-conditioned semantic distribution defined over a restricted candidate set $\mathcal{V}$, which consists of the

top- N words ranked by the topic–word distribution β_k . We apply a temperature-scaled softmax:

$$\hat{\beta}_k^{\text{sem}}(w) = \frac{\exp(s_k(w)/\tau)}{\sum_{w' \in \mathcal{V}} \exp(s_k(w')/\tau)},$$

where $\tau > 0$ controls the sharpness of the distribution. The resulting $\hat{\beta}_k^{\text{sem}}$ captures BERT’s topic-conditioned semantic preferences over the candidate set and serves as the semantic reference signal for subsequent refinement components.

3.2 Semantic-Guided Refinement

We use the topic-conditioned semantic distribution $\hat{\beta}_k^{\text{sem}}$ as a reference signal to guide the refinement of topic representations. In contrast to prior approaches that directly generate topic words and are therefore susceptible to vocabulary mismatches or prompt-dependent variations (Yang et al., 2025b; Wang et al., 2023a; Rijcken et al., 2023), our method refines corpus-driven estimates within a fixed candidate space.

Concretely, we restrict the original topic–word distribution β_k to the same top- N candidate set used to construct $\hat{\beta}_k^{\text{sem}}$ and renormalize it to obtain a truncated distribution, denoted by $\beta_k^{(N)}$. We then encourage alignment between the two distributions by minimizing the forward Kullback–Leibler (KL) divergence:

$$\mathcal{L}_{\text{sem}}(k) = \text{KL}(\hat{\beta}_k^{\text{sem}} \parallel \beta_k^{(N)}),$$

where both $\hat{\beta}_k^{\text{sem}}$ and $\beta_k^{(N)}$ are defined over the same set of top- N candidate words.

3.3 Semantic Refinement of Document–Topic Representations

While the semantic refinement in the previous subsection focuses on adjusting the topic–word distributions on the decoder side, we now turn to semantic refinement on the encoder side by improving document–topic representations. Specifically, we refine the document–topic proportions θ through semantic alignment in a shared embedding space.

Topic embedding construction. Let $\mathcal{W}_k^{(M)} = \{w_{k,1}, \dots, w_{k,M}\}$ denote the resulting subset of the top- N topic candidates, selected as the M highest-scoring words according to $s_k(w)$ and ordered by decreasing semantic score. We concatenate the selected words into a single sequence and

encode them using a pretrained encoder $f(\cdot)$ to obtain a topic embedding (Zhang et al., 2025). The topic embedding is then obtained as

$$\mathbf{z}_k = f(w_{k,1}, \dots, w_{k,M}).$$

where $\mathbf{z}_k$ corresponds to the encoder output associated with the sequence-level representation (e.g., the [CLS] token).

Document embeddings. Given a document d , we obtain its embedding $\mathbf{z}_d = f(d)$ using the same encoder, ensuring that document and topic representations lie in a shared semantic space.

MMD-based alignment. We refine document–topic representations by aligning document embeddings with their corresponding topic mixtures in the embedding space. Specifically, we define the semantic document distribution as a Dirac measure

$$P_d = \delta(\mathbf{z}_d),$$

and the semantic topic mixture distribution as

$$Q_d = \sum_{k=1}^K \theta_{d,k} \delta(\mathbf{z}_k),$$

where $\mathbf{z}_k$ denotes the topic embedding constructed from semantically salient topic words.

The encoder-side semantic refinement objective is then defined by minimizing the maximum mean discrepancy between these two distributions:

$$\mathcal{L}_{\text{enc}} = \sum_d \text{MMD}(P_d, Q_d),$$

where the MMD formulation follows the definition introduced in the background section.

This formulation aligns document representations with semantically informed topic mixtures at the distribution level, without requiring explicit word generation or vocabulary expansion. Together with decoder-side refinement of topic–word distributions, this encoder-side alignment injects semantic structure into both inference and generative components of the topic model.

3.4 Training Objective

Training is performed in two phases. In the first phase, we train the base topic model using its original objective $\mathcal{L}_{\text{TM}}$. In the second phase, we further refine the model by incorporating the proposed semantic refinement losses on both the decoder and encoder sides.

Model	20News		R8		DBpedia		AGNews	
	C_V	PN	C_V	PN	C_V	PN	C_V	PN
LDA (Blei et al., 2003)	0.529	0.489	0.384	0.700	0.537	0.762	0.544	0.596
BERTopic (Grootendorst, 2022)	0.404	0.342	0.323	0.697	0.545	0.720	0.506	0.450
TopicGPT (Pham et al., 2024a)	NA	0.363	NA	0.410	NA	0.706	NA	0.634
NVDM (Miao et al., 2017)	0.243	0.140	0.283	0.371	0.306	0.177	0.317	0.250
+ LLM-ITL	0.314 ± 0.025	0.133 ± 0.013	0.386 ± 0.022	0.360 ± 0.011	0.449 ± 0.019	0.182 ± 0.010	0.435 ± 0.021	0.255 ± 0.011
+ BERT-NTM	0.422 ± 0.022	0.132 ± 0.012	0.500 ± 0.019	0.358 ± 0.010	0.516 ± 0.017	0.183 ± 0.009	0.541 ± 0.018	0.256 ± 0.010
SCHOLAR (Card et al., 2018)	0.511	0.580	0.388	0.696	0.617	0.832	0.573	0.641
+ LLM-ITL	0.560 ± 0.013	0.590 ± 0.012	0.419 ± 0.012	0.684 ± 0.010	0.667 ± 0.011	0.832 ± 0.009	0.662 ± 0.012	0.639 ± 0.010
+ BERT-NTM	0.643 ± 0.011	0.581 ± 0.011	0.478 ± 0.010	0.685 ± 0.009	0.708 ± 0.009	0.846 ± 0.008	0.680 ± 0.010	0.641 ± 0.009
ETM (Dieng et al., 2020)	0.490	0.444	0.419	0.670	0.505	0.759	0.535	0.574
+ LLM-ITL	0.557 ± 0.013	0.387 ± 0.012	0.531 ± 0.012	0.690 ± 0.010	0.654 ± 0.011	0.724 ± 0.009	0.639 ± 0.012	0.568 ± 0.010
+ BERT-NTM	0.578 ± 0.011	0.410 ± 0.010	0.616 ± 0.010	0.727 ± 0.008	0.676 ± 0.009	0.766 ± 0.007	0.644 ± 0.010	0.588 ± 0.008
NSTM (Zhao et al., 2021)	0.432	0.371	0.408	0.686	0.648	0.764	0.578	0.598
+ LLM-ITL	0.452 ± 0.013	0.361 ± 0.011	0.565 ± 0.012	0.697 ± 0.009	0.678 ± 0.011	0.756 ± 0.008	0.673 ± 0.012	0.596 ± 0.009
+ BERT-NTM	0.564 ± 0.011	0.369 ± 0.010	0.587 ± 0.010	0.688 ± 0.008	0.788 ± 0.009	0.797 ± 0.007	0.662 ± 0.010	0.599 ± 0.008
CLNTM (Nguyen and Luu, 2021)	0.466	0.579	0.376	0.679	0.506	0.721	0.561	0.574
+ LLM-ITL	0.615 ± 0.012	0.578 ± 0.010	0.435 ± 0.011	0.679 ± 0.009	0.640 ± 0.010	0.667 ± 0.008	0.692 ± 0.011	0.577 ± 0.009
+ BERT-NTM	0.640 ± 0.010	0.581 ± 0.009	0.437 ± 0.010	0.695 ± 0.008	0.694 ± 0.008	0.679 ± 0.007	0.674 ± 0.009	0.595 ± 0.008
WeTe (dongsheng wang et al., 2022)	0.534	0.309	0.507	0.658	0.550	0.744	0.597	0.522
+ LLM-ITL	0.619 ± 0.012	0.283 ± 0.009	0.594 ± 0.011	0.662 ± 0.008	0.663 ± 0.010	0.726 ± 0.007	0.647 ± 0.011	0.524 ± 0.008
+ BERT-NTM	0.620 ± 0.011	0.272 ± 0.008	0.632 ± 0.010	0.645 ± 0.007	0.694 ± 0.009	0.712 ± 0.006	0.675 ± 0.010	0.529 ± 0.007
ECRTM (Wu et al., 2023)	0.320	0.516	0.300	0.663	0.588	0.764	0.422	0.586
+ LLM-ITL	0.550 ± 0.028	0.519 ± 0.013	0.354 ± 0.025	0.664 ± 0.012	0.686 ± 0.039	0.775 ± 0.010	0.472 ± 0.026	0.588 ± 0.011
+ BERT-NTM	0.613 ± 0.025	0.534 ± 0.011	0.480 ± 0.022	0.647 ± 0.010	0.738 ± 0.035	0.789 ± 0.008	0.663 ± 0.022	0.600 ± 0.009
FASTopic (Wu et al., 2024b)	0.5147	0.4746	0.3920	0.6546	0.5794	0.7695	0.5382	0.6117
+ LLM-ITL	0.5612 ± 0.012	0.4763 ± 0.010	0.4415 ± 0.011	0.6487 ± 0.009	0.6429 ± 0.010	0.7720 ± 0.008	0.5753 ± 0.011	0.6072 ± 0.009
+ BERT-NTM	0.5891 ± 0.010	0.4805 ± 0.009	0.4803 ± 0.009	0.6420 ± 0.008	0.6798 ± 0.008	0.7701 ± 0.007	0.5932 ± 0.009	0.6019 ± 0.008
GINopic (Adhya and Sanyal, 2024)	0.3913	0.4607	0.3443	0.6511	0.4610	0.7703	0.4375	0.5936
+ LLM-ITL	0.4382 ± 0.011	0.4622 ± 0.009	0.3985 ± 0.010	0.6466 ± 0.008	0.6030 ± 0.009	0.7763 ± 0.007	0.5567 ± 0.010	0.5924 ± 0.008
+ BERT-NTM	0.4407 ± 0.010	0.4632 ± 0.008	0.4026 ± 0.009	0.6498 ± 0.007	0.6920 ± 0.008	0.7819 ± 0.006	0.6012 ± 0.009	0.5935 ± 0.007

Table 1: Comparison of topic modeling performance on four datasets (20News, R8, DBpedia, and AGNews). We report topic coherence (C_V) and topic alignment (PN) for base models, their LLM-ITL-refined variants, and **BERT-NTM**, including recent neural topic model backbones FASTopic and GINopic. **Bold** values indicate the highest score within each model family on a given dataset, and “NA” denotes not applicable. Results for **LDA** (Blei et al., 2003) and **TopicGPT** (Pham et al., 2024a) are taken from **LLM-ITL** (Yang et al., 2025b). For these quality-metric comparisons, we use the same released preprocessed datasets and evaluation setting as **LLM-ITL**; all other methods are run five times.

The overall objective in the second phase is given by

$$\mathcal{L} = \mathcal{L}_{\text{TM}} + \lambda_{\text{dec}} \mathcal{L}_{\text{sem}} + \lambda_{\text{enc}} \mathcal{L}_{\text{enc}},$$

where $\mathcal{L}_{\text{sem}} = \sum_k \text{KL}(\hat{\beta}_k^{\text{sem}} \| \beta_k^{(N)})$ refines the topic-word distributions as described in Section 3.2, and $\mathcal{L}_{\text{enc}} = \sum_d \text{MMD}(P_d, Q_d)$ refines the document-topic representations as described in Section 3.3. Detail algorithm is reported in Appendix B

4 Experiments and Results

Datasets

We evaluate our approach on four standard publicly available benchmarks for topic modeling: 20News-group (Lang, 1995) (**20News**), Reuters-21578 (Debole and Sebastiani, 2005) (**R8**), DBpedia (Auer et al., 2007), and AGNews (Zhang et al., 2015). Detailed dataset statistics are provided in Appendix C. For datasets with longer documents (20News and R8), we use 50 topics, while for shorter-document datasets (DBpedia and AGNews) we set the number of topics to 25. We further evaluate performance across multiple values of K , with full results reported in Appendix G.

Baseline Models

Following prior work on LLM-based iterative topic labeling (Yang et al., 2025b), we conduct experiments using the **BERT-NTM** framework integrated with a range of neural topic models. In addition to neural topic models, we include representative baselines from other modeling paradigms, including the probabilistic topic model LDA (Blei et al., 2003), the clustering-based approach BERTopic (Grootendorst, 2022), and the recent LLM-based topic modeling method TopicGPT (Pham et al., 2024a). To strengthen comparison with recent neural topic models, we additionally evaluate FASTopic (Wu et al., 2024b) and GINopic (Adhya and Sanyal, 2024) within the same refinement setting. We further perform a direct comparison between **BERT-NTM** and the **LLM-ITL** (Yang et al., 2025b) framework under the same base topic models, enabling a controlled evaluation of encoder-based semantic refinement versus generative LLM-based refinement. This comparison isolates the effect of the refinement mechanism itself, independent of the underlying topic modeling backbone. Detailed configurations and hyperparameter settings for all baseline models

are provided in Appendix D.

Evaluation Metrics

We evaluate **BERT-NTM** along two dimensions: topic quality and document representation quality. Topic quality is measured using the standard topic coherence metric C_V (Röder et al., 2015), computed with Palmetto over the top-10 words of each topic against an external Wikipedia reference index and reported as the average across all topics; the full protocol is documented in Appendix H.1. Document representation quality is assessed via *topic alignment*, which evaluates how well document–topic assignments correspond to ground-truth labels (Chuang et al., 2013; Pham et al., 2024a). We report the average of Purity and NMI (PN) as the primary alignment metric, following Yang et al. (2025b) for comparability; the two components are also reported separately, and we discuss the limits of this aggregate in Appendix H.4. We also provide an LLM-based topic interpretability evaluation (Stammach et al., 2023) in Appendix G.1 as a complementary assessment. Other metrics, including topic diversity (TD) and individual Purity and NMI scores, are provided in Appendix G, with TD in Appendix G.8 and Purity/NMI in Appendix G.9. Automated coherence metrics are known to diverge from human judgement (Doogan and Buntine, 2021; Hoyle et al., 2021); we therefore rely on agreement across several metrics rather than on any one of them, and discuss this in Appendix H.2.

4.1 Topic Quality Analysis

Topic Coherence and Alignment. Table 1 reports topic coherence (C_v) and topic alignment (PN) for different models. Both LLM-ITL and **BERT-NTM** improve coherence over the base topic models, with **BERT-NTM** consistently achieving the highest C_v when applied to the same backbone. For example, on **20News** with NVDM and **R8** with ETM, BERT-NTM produces more coherent topics than both the base model and LLM-ITL. At the same time, PN remains broadly comparable across methods, indicating that coherence gains do not substantially disrupt document–topic associations. This pattern also holds on larger datasets such as **DBpedia**, where BERT-NTM improves coherence while preserving strong alignment.

Overall, these results show that **BERT-NTM** achieves the best topic coherence among all com-

pared approaches while maintaining competitive topic alignment.

To strengthen comparison with recent neural topic models, we also report results on FASTopic (Wu et al., 2024b) and GINopic (Adhya and Sanyal, 2024) in Table 1. BERT-NTM consistently improves C_v on both backbones across 20News, R8, DBpedia, and AGNews. Notably, FASTopic already includes a Dual Semantic-relation Reconstruction (DSR) mechanism, yet BERT-NTM still brings further coherence gains, suggesting complementarity rather than redundancy. PN remains generally competitive, with modest gains on some datasets and small trade-offs on others. This trend is also consistent with an auxiliary LLM-based topic interpretability evaluation reported in Appendix G.1, where BERT-NTM receives higher average coherence scores than the corresponding base models and LLM-ITL.

Additional qualitative examples illustrating the effect of semantic refinement on topic–word distributions are provided in Appendix G.5.

4.2 Ablation Study

To assess the contribution of individual components in **BERT-NTM**, we conduct ablation studies on the ETM baseline using the 20 Newsgroups and R8 datasets (Table 2). Removing semantic scoring causes a clear drop in topic coherence, with C_V decreasing from 0.578 to 0.512 on 20News and from 0.616 to 0.493 on R8, confirming its central role in topic refinement. Removing document–topic alignment mainly affects topic alignment, reducing PN from 0.727 to 0.667 on R8. Overall, the full model improves C_V by +0.088 on 20News and +0.197 on R8 relative to ETM, showing that the two components are complementary.

We also compare several alternatives to the KL divergence used in the decoder-side semantic objective $\mathcal{L}_{\text{sem}}$, including cosine similarity, one-dimensional optimal transport (1D-OT), and mean squared error (MSE). Although some alternatives achieve competitive topic coherence, they often reduce topic alignment or behave less stably, as shown by the sharp PN drop of 1D-OT on 20News (0.237). In contrast, KL divergence in the full model gives the best overall balance, achieving the highest PN on R8 (0.727) while maintaining strong topic coherence. The analogous study for the encoder-side objective $\mathcal{L}_{\text{enc}}$, where MMD is compared against cosine similarity and MSE, is reported in Appendix G.13.

Variant	20News		R8	
	C_V	PN	C_V	PN
ETM (Base)	0.490	0.444	0.419	0.670
<i>Component Ablation</i>				
w/o Doc-Topic Align.	0.535	0.410	0.596	0.667
w/o Semantic Scoring	0.512	0.408	0.493	0.693
<i>Decoder-Side Semantic Objective ($\mathcal{L}_{sem}$)</i>				
w/ Cosine	0.514	0.410	0.604	0.622
w/ 1D-OT	0.571	0.237	0.604	0.622
w/ MSE	0.514	0.410	0.513	0.691
BERT-NTM (Full)	0.578	0.410	0.616	0.727

Table 2: Ablation study on ETM baseline. Top: component ablation showing the impact of removing individual refinement components. Bottom: comparison of alternatives to the KL divergence used in the *decoder-side* semantic objective $\mathcal{L}_{sem}$, which aligns β_k with the masked-language-model semantic distribution. KL Divergence (our default) achieves the best performance. The corresponding study for the *encoder-side* document-topic alignment objective $\mathcal{L}_{enc}$ is reported separately in Table E12.

4.3 Robustness and Sensitivity Analysis

We study the robustness of **BERT-NTM** to refinement design choices and the sensitivity of its main hyperparameters, including template wrapper, keyword order, subword aggregation, Top- M , refinement frequency, and Top- N .

Table 3 shows that **BERT-NTM** is highly stable across natural-language template wrappers (T1–T3) on AGNews and R8, with only a small drop for the raw-list variant (T4), suggesting that gains come mainly from MLM compatibility among topic words rather than template artifacts. Preserving, reversing, or shuffling keyword order yields negligible differences, indicating limited sensitivity to order. For subword aggregation, length-normalized averaging is the most reliable, while *sum*, *min*, and first-subword variants remain competitive but slightly less stable. Sensitivity to Top- M is reported in Appendix G.3, where a moderate Top- M gives the best overall balance.

Figure 4 shows the sensitivity of **BERT-NTM** with ETM on R8 to refinement frequency and Top- N . More frequent refinement improves C_V while keeping PN stable, but increases training cost; less frequent refinement is faster but weakens topic quality. Likewise, very small Top- N reduces computation but provides weaker semantic guidance, whereas larger Top- N improves C_V and PN at the cost of higher overhead and slightly lower

Variant	AGNews		R8	
	C_V	PN	C_V	PN
<i>Template Wrapper</i>				
T1 ("Words in this topic: ...")	0.6800	0.6410	0.4780	0.6850
T2 ("Related words include: ...")	0.6789	0.6419	0.4769	0.6852
T3 ("Topic words are: ...")	0.6798	0.6411	0.4778	0.6856
T4 (raw list)	0.6705	0.6362	0.4697	0.6829
<i>Keyword Order</i>				
Preserve	0.6800	0.6410	0.4780	0.6850
Reverse	0.6792	0.6412	0.4776	0.6840
Shuffle	0.6807	0.6415	0.4774	0.6846
<i>Subword Aggregation</i>				
Mean (length-normalized)	0.6801	0.6414	0.4784	0.6853
Sum	0.6685	0.6429	0.4770	0.6851
Min	0.6703	0.6382	0.4780	0.6850
First subword	0.6712	0.6398	0.4737	0.6848

Table 3: Robustness under template wrapper, keyword order, and subword aggregation variants on SCHOLAR + **BERT-NTM**. T1–T3 correspond to the three natural-language wrappers described in the text.

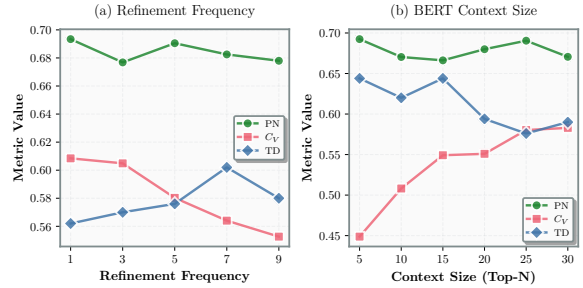


Figure 4: Sensitivity analysis of **BERT-NTM** with respect to refinement frequency and topic context size (Top- N). We report topic alignment (PN), topic coherence (C_V), and topic diversity (TD) under different settings.

TD. Overall, **BERT-NTM** is robust to the formulation of MLM-based refinement, while moderate refinement frequency and intermediate Top- N /Top- M provide the best trade-off between topic quality and efficiency. Exact values for the Top- N sweep are tabulated in Appendix G.14, a controlled comparison of the document-topic alignment objective (none / cosine / MSE / MMD) is given in Appendix G.13, and the parameter budgets of the two refinement methods are compared in Appendix G.18. Additional results and implementation details are provided in the appendix.

4.4 VRAM Usage and Training Time

We compare the computational efficiency of **BERT-NTM** with LLM-ITL in terms of GPU memory usage and training time on DBpedia and AGNews. As shown in Figure 5, **BERT-NTM** consistently operates in a substantially more efficient regime,

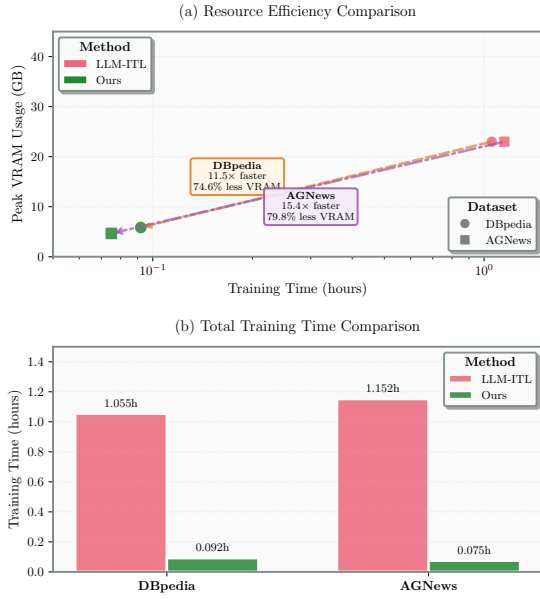


Figure 5: Resource efficiency comparison between **BERT-NTM** and LLM-ITL built on the ETM base model, evaluated on DBpedia and AGNews in terms of VRAM usage and training time.

achieving a 74.6% reduction in peak VRAM usage and an $11.5\times$ speedup on DBpedia, and a 79.8% reduction in memory usage with a $15.4\times$ speedup on AGNews. Across both datasets, **BERT-NTM** completes end-to-end training within minutes, whereas LLM-ITL requires over an hour, demonstrating the practical efficiency advantages of encoder-based semantic refinement for large-scale and resource-constrained topic modeling.

5 Conclusion

We presented **BERT-NTM**, an encoder-based semantic refinement framework for neural topic modeling. By leveraging discriminative semantic signals from pretrained encoders, **BERT-NTM** improves topic coherence and document–topic alignment without generative inference, while remaining computationally efficient across multiple backbones.

Limitations

While **BERT-NTM** provides an efficient and stable alternative to generative LLM-based refinement, it has several limitations. First, the framework relies on a pretrained encoder language model to provide semantic guidance, and the effectiveness of refinement therefore depends on the representational quality and coverage of the underlying encoder. Second, although **BERT-NTM** is compat-

ible with a wide range of neural topic models, its performance is influenced by the quality of the base topic model after initial training, and weak initial topic representations may limit the benefits of subsequent refinement.

Ethical Considerations

We adhere to the ACL Code of Ethics and the terms of each codebase license. Our method aims to advance the field of topic modeling, and we are confident that, when used properly and with care, it poses no significant social risks.

Acknowledgments

Trung Le was supported by the Air Force Office of Scientific Research under award number FA2386-25-1-4023 and the ARC Discovery Project grant DP250100262. Thien Huu Nguyen was partially supported by NSF Grant #2239570. The views and conclusions contained herein are those of the authors and should not be interpreted as necessarily representing the official policies, either expressed or implied, of ODNI, IARPA, or the U.S. Government.

References

- Suman Adhya and Debarshi Kumar Sanyal. 2024. [Ginopic: Topic modeling with graph isomorphism network](#). In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, NAACL 2024, Mexico City, Mexico, June 16-21, 2024, pages 6171–6183. Association for Computational Linguistics.
- Dimo Angelov. 2020. [Top2vec: Distributed representations of topics](#). *CoRR*, abs/2008.09470.
- Sören Auer, Christian Bizer, Georgi Kobilarov, Jens Lehmann, Richard Cyganiak, and Zachary Ives. 2007. Dbpedia: A nucleus for a web of open data. In *The Semantic Web*, pages 722–735, Berlin, Heidelberg. Springer Berlin Heidelberg.
- Federico Bianchi, Silvia Terragni, and Dirk Hovy. 2021a. [Pre-training is a hot topic: Contextualized document embeddings improve topic coherence](#). In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing, ACL/IJCNLP 2021, (Volume 2: Short Papers)*, Virtual Event, August 1-6, 2021, pages 759–766. Association for Computational Linguistics.
- Federico Bianchi, Silvia Terragni, Dirk Hovy, Debora Nozza, and Elisabetta Fersini. 2021b. [Cross-lingual](#)

- contextualized topic models with zero-shot learning. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume, EACL 2021, Online, April 19 - 23, 2021*, pages 1676–1683. Association for Computational Linguistics.
- David M. Blei and John D. Lafferty. 2006. [Dynamic topic models](#). In *Proceedings of the 23rd International Conference on Machine Learning, ICML '06*, page 113–120, New York, NY, USA. Association for Computing Machinery.
- David M Blei, Andrew Y Ng, and Michael I Jordan. 2003. Latent dirichlet allocation. *Journal of Machine Learning Research*, pages 993–1022.
- Dallas Card, Chenhao Tan, and Noah A. Smith. 2018. [Neural models for documents with metadata](#). In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2031–2040, Melbourne, Australia. Association for Computational Linguistics.
- Jason Chuang, Sonal Gupta, Christopher D. Manning, and Jeffrey Heer. 2013. Topic model diagnostics: assessing domain relevance via topical alignment. In *Proceedings of the 30th International Conference on International Conference on Machine Learning - Volume 28, ICML'13*, page III–612–III–620. JMLR.org.
- Quoc Phong Dao, Hoang Son Nguyen, Chi Pham Khanh, Linh Ngo Van, Nguyen Thi Ngoc Diep, Thien Huu Nguyen, and Trung Le. 2026. TALAS: teacher-anchored layer alignment with adaptive sharpness-aware minimization for embedding distillation. In *Proceedings of the 64th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 32705–32723. Association for Computational Linguistics.
- Franca Debole and Fabrizio Sebastiani. 2005. [An analysis of the relative hardness of reuters-21578 subsets](#). *J. Assoc. Inf. Sci. Technol.*, 56(6):584–596.
- Adji B Dieng, Francisco JR Ruiz, and David M Blei. 2020. Topic modeling in embedding spaces. *Transactions of the Association for Computational Linguistics*, pages 439–453.
- 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, ACL 2024 - Student Research Workshop, Bangkok, Thailand, August 11-16, 2024*, pages 21–33. Association for Computational Linguistics.
- dongsheng wang, Dan dan Guo, He Zhao, Huangjie Zheng, Korawat Tanwisuth, Bo Chen, and Mingyuan Zhou. 2022. [Representing mixtures of word embeddings with mixtures of topic embeddings](#). In *International Conference on Learning Representations*.
- 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.
- Damien Garreau, Wittawat Jitkrittum, and Motonobu Kanagawa. 2017. [Large sample analysis of the median heuristic](#). *arXiv preprint arXiv:1707.07269*.
- Arthur Gretton, Karsten M. Borgwardt, Malte J. Rasch, Bernhard Schölkopf, and Alexander J. Smola. 2012. [A kernel two-sample test](#). *J. Mach. Learn. Res.*, 13:723–773.
- Maarten Grootendorst. 2022. Bertopic: Neural topic modeling with a class-based tf-idf procedure. *arXiv preprint arXiv:2203.05794*.
- Nguyen Manh Hieu, Vu Lam Anh, Hung Pham Van, Nam Le Hai, Diep Thi-Ngoc Nguyen, Linh Ngo Van, and Thien Huu Nguyen. 2025. MaGiX: A multi-granular adaptive graph intelligence framework for enhancing cross-lingual RAG. In *Findings of the Association for Computational Linguistics: EMNLP 2025*, pages 5202–5219. Association for Computational Linguistics.
- Alexander Hoyle, Pranav Goel, Denis Peskov, Andrew Hian-Cheong, Jordan Boyd-Graber, and Philip Resnik. 2021. [Is automated topic model evaluation broken? the incoherence of coherence](#). In *Advances in Neural Information Processing Systems 34: Annual Conference on Neural Information Processing Systems 2021, NeurIPS 2021*, volume 34.
- Alexander Miserlis Hoyle, Pranav Goel, and Philip Resnik. 2020. [Improving neural topic models using knowledge distillation](#). In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing, EMNLP 2020, Online, November 16-20, 2020*, pages 1752–1771. Association for Computational Linguistics.
- Phung Gia Huy, Hai An Vu, Minh-Phuc Truong, Thang Duc Tran, Linh Ngo Van, Thanh Hong Nguyen, and Trung Le. 2026. MIPIC: matryoshka representation learning via self-distilled intra-relational and progressive information chaining. In *Findings of the Association for Computational Linguistics: ACL 2026*, pages 13824–13838. Association for Computational Linguistics.
- Chi Pham Khanh, Tue Le, Anh-Cuong Le, and Linh Ngo Van. 2026a. A multi-objective surrogate framework for training neural topic models. *Neurocomputing*, 682:133432.
- Chi Pham Khanh, Hoang Tran Vuong, Sang Dinh Viet, and Linh Ngo Van. 2026b. A framework for neural topic modeling using hierarchical clustering and contrastive learning with optimal transport. *Neurocomputing*, 670:132530.

- Diederik P Kingma and Max Welling. 2013. Auto-encoding variational bayes. In *2nd International Conference on Learning Representations*.
- Ken Lang. 1995. Newsweeder: learning to filter net-news. In *Proceedings of the Twelfth International Conference on International Conference on Machine Learning, ICML'95*, page 331–339, San Francisco, CA, USA. Morgan Kaufmann Publishers Inc.
- Tue Le, Hoang Tran Vuong, Tung Nguyen, Linh Ngo Van, Dinh Viet Sang, Trung Le, and Thien Huu Nguyen. 2025. Multi-surrogate-objective optimization for neural topic models. In *Findings of the Association for Computational Linguistics: EMNLP 2025, Suzhou, China, November 4-9, 2025*, pages 135–151. Association for Computational Linguistics.
- Vinh Le, Nguyen Hong Dang, Tu Vu, Linh Ngo Van, Duc Anh Nguyen, and Trung Le. 2026. HieRD: hierarchical relational distillation for vision-language embedding models. In *Forty-third International Conference on Machine Learning*.
- Chun-Liang Li, Wei-Cheng Chang, Yu Cheng, Yiming Yang, and Barnabás Póczos. 2017. *Mmd gan: Towards deeper understanding of moment matching network*. In *Advances in Neural Information Processing Systems (NeurIPS)*.
- Yujia Li, Kevin Swersky, and Richard S. Zemel. 2015. *Generative moment matching networks*. *CoRR*, abs/1502.02761.
- Mingsheng Long, Yue Cao, Jianmin Wang, and Michael I. Jordan. 2015. *Learning transferable features with deep adaptation networks*. In *Proceedings of the 32nd International Conference on Machine Learning (ICML)*, pages 97–105.
- Andrew Kachites McCallum. 2002. MALLET: A machine learning for language toolkit. <http://mallet.cs.umass.edu>.
- Yishu Miao, Edward Grefenstette, and Phil Blunsom. 2017. Discovering discrete latent topics with neural variational inference. In *Proceedings of the 34th International Conference on Machine Learning - Volume 70, ICML'17*, page 2410–2419. JMLR.org.
- Yida Mu, Chun Dong, Kalina Bontcheva, and Xingyi Song. 2024. *Large language models offer an alternative to the traditional approach of topic modelling*. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation, LREC/COLING 2024, 20-25 May, 2024, Torino, Italy*, pages 10160–10171. ELRA and ICCL.
- Krikamol Muandet, Kenji Fukumizu, Bharath K. Sriperumbudur, and Bernhard Schölkopf. 2017. *Kernel mean embedding of distributions: A review and beyond*. *Foundations and Trends in Machine Learning*, 10(1–2):1–141.
- Feng Nan, Ran Ding, Ramesh Nallapati, and Bing Xiang. 2019. Topic modeling with wasserstein autoencoders. *arXiv preprint arXiv:1907.12374*.
- Quang Duc Nguyen, Tung Nguyen, Duc Anh Nguyen, Linh Ngo Van, Sang Dinh, and Thien Huu Nguyen. 2025a. *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, NAACL 2025 - Volume 1: Long Papers*, pages 1109–1124. Association for Computational Linguistics.
- Thanh Vinh Nguyen, Ngo Van Dong, Minh Chu Xuan, Tung Nguyen, Linh Ngo Van, Dinh Viet Sang, and Trung Le. 2026a. Beyond coherence: Improving temporal consistency and interpretability in dynamic topic models. In *Findings of the Association for Computational Linguistics: EACL*, pages 3609–3629. Association for Computational Linguistics.
- Thong Nguyen and Anh Tuan Luu. 2021. Contrastive learning for neural topic model. In *Advances in Neural Information Processing Systems*, pages 11974–11986.
- Thong Thanh Nguyen, Xiaobao Wu, Xinshuai Dong, Cong-Duy T. Nguyen, See-Kiong Ng, and Anh Tuan Luu. 2024. *Topic modeling as multi-objective contrastive optimization*. In *The Twelfth International Conference on Learning Representations, ICLR 2024, Vienna, Austria, May 7-11, 2024*. OpenReview.net.
- Tien-Phat Nguyen, Vu Minh Ngo, Tung Nguyen, Linh Ngo Van, Duc Anh Nguyen, Dinh Viet Sang, and Trung Le. 2025b. XTRA: cross-lingual topic modeling with topic and representation alignments. In *Findings of the Association for Computational Linguistics: EMNLP 2025*, pages 5561–5575. Association for Computational Linguistics.
- Tien-Phat Nguyen, Giang Pham Thi Phuong, Thuan Do Phan, Anh-Cuong Le, and Linh Ngo Van. 2026b. A framework for cross-lingual topic modeling using global context spaces and unified decoding with semantic knowledge distillation. *Neurocomputing*, page 134278.
- Tung Nguyen, Duy-Tung Doan, Huu-Thanh Luong, and Linh Ngo Van. 2025c. Sub-document neural topic model. *Knowledge-Based Systems*, page 114762.
- Tung Nguyen, Tue Le, Hoang Tran Vuong, Quang Duc Nguyen, Duc Anh Nguyen, Linh Ngo Van, Sang Dinh, and Thien Huu Nguyen. 2025d. *Sharpness-aware minimization for topic models with high-quality document representations*. In *Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies, NAACL 2025 - Volume 1: Long Papers*, pages 4507–4524. Association for Computational Linguistics.
- Tung Nguyen, Linh Ngo Van, Anh Nguyen Duc, and Sang Dinh Viet. 2026c. Global and local context in

- short text neural topic model. *Artificial Intelligence*, 353:104502.
- OpenAI. 2023. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*.
- Chau Pham, Alexander Miserlis Hoyle, Simeng Sun, Philip Resnik, and Mohit Iyyer. 2024a. [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)*, NAACL 2024, Mexico City, Mexico, June 16-21, 2024, pages 2956–2984. Association for Computational Linguistics.
- Duy-Tung Pham, Thien Trang Nguyen Vu, Tung Nguyen, Linh Ngo Van, Duc Anh Nguyen, and Thien Huu Nguyen. 2024b. [Neuromax: Enhancing neural topic modeling via maximizing mutual information and group topic regularization](#). In *Findings of the Association for Computational Linguistics: EMNLP 2024, Miami, Florida, USA, November 12-16, 2024*, pages 7758–7772. Association for Computational Linguistics.
- Nguyen Tien Phat, Ngo Vu Minh, Linh Ngo Van, Nguyen Thi Ngoc Diep, and Thien Huu Nguyen. 2026. Gloctm: Cross-lingual topic modeling via a global context space. In *Fortieth AAAI Conference on Artificial Intelligence, AAAI*, pages 32710–32718. AAAI Press.
- Hamed Rahimi, David Mimno, Jacob Hoover, Hubert Naacke, Camelia Constantin, and Bernd Amann. 2024. [Contextualized topic coherence metrics](#). In *Findings of the Association for Computational Linguistics: EACL 2024*, pages 1760–1773, St. Julian’s, Malta. Association for Computational Linguistics.
- Emil Rijcken, Floortje Scheepers, Kalliopi Zervanou, Marco Spruit, Pablo Mosteiro, and Uzay Kaymak. 2023. [Towards interpreting topic models with chatgpt](#). In *Proceedings of the 20th World Congress of the International Fuzzy Systems Association (IFSA 2023)*, pages 269–275, Daegu, Korea, Republic of. Paper presented at IFSA 2023.
- 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, WSDM ’15*, page 399–408, New York, NY, USA. Association for Computing Machinery.
- Michal Rosen-Zvi, Thomas Griffiths, Mark Steyvers, and Padhraic Smyth. 2004. The author-topic model for authors and documents. In *Proceedings of the 20th Conference on Uncertainty in Artificial Intelligence, UAI ’04*, page 487–494, Arlington, Virginia, USA. AUAI Press.
- Akash Srivastava and Charles Sutton. 2017. Autoencoding variational inference for topic models. In *International Conference on Learning Representations*.
- Dominik Stammbach, Vilém Zouhar, Alexander Miserlis 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, EMNLP 2023, Singapore, December 6-10, 2023*, pages 9348–9357. Association for Computational Linguistics.
- Laure Thompson and David Mimno. 2020. [Topic modeling with contextualized word representation clusters](#). *CoRR*, abs/2010.12626.
- Thang Duc Tran, Minh-Phuc Truong, Thuan Do Phan, and Linh Ngo Van. 2026. SAMD: span-aware matryoshka distillation for cross-tokenizer embedding models. *Machine Learning*, 115(6):143.
- Minh-Phuc Truong, Hai An Vu, Tu Vu, Nguyen Thi Ngoc Diep, Linh Ngo Van, Thien Huu Nguyen, and Trung Le. 2025. EMO: embedding model distillation via intra-model relation and optimal transport alignments. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pages 7605–7617. Association for Computational Linguistics.
- Hai An Vu, Minh-Phuc Truong, Tu Vu, and Linh Ngo Van. 2026. MoL: mixture of layers in cross-tokenizer embedding model distillation. *Knowledge-Based Systems*, page 116001.
- Tu Vu, Manh Do, Tung Nguyen, Ngo Van Linh, 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, Vienna, Austria, July 27 - August 1, 2025*, pages 7666–7680. Association for Computational Linguistics.
- Hoang Tran Vuong, Tue Le, Tu Vu, Tung Nguyen, Linh Ngo Van, Sang Dinh, and Thien Huu Nguyen. 2025. [Hicot: Improving neural topic models via optimal transport and contrastive learning](#). In *Findings of the Association for Computational Linguistics: ACL 2025, Vienna, Austria, July 27 - August 1, 2025*, pages 13894–13920. Association for Computational Linguistics.
- Han Wang, Nirmalendu Prakash, Nguyen-Khoi Hoang, Ming Shan Hee, Usman Naseem, and Roy Ka-Wei Lee. 2023a. [Prompting large language models for topic modeling](#). In *IEEE International Conference on Big Data, BigData 2023, Sorrento, Italy, December 15-18, 2023*, pages 1236–1241. IEEE.
- Xinyi Wang, Wanrong Zhu, and William Yang Wang. 2023b. [Large language models are implicitly topic models: Explaining and finding good demonstrations for in-context learning](#). *CoRR*, abs/2301.11916.

- 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, ICML 2023, 23-29 July 2023, Honolulu, Hawaii, USA*, volume 202 of *Proceedings of Machine Learning Research*, pages 37335–37357. PMLR.
- Xiaobao Wu, Anh Tuan Luu, and Xinshuai Dong. 2022. [Mitigating data sparsity for short text topic modeling by topic-semantic contrastive learning](#). In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing, EMNLP 2022, Abu Dhabi, United Arab Emirates, December 7-11, 2022*, pages 2748–2760. Association for Computational Linguistics.
- Xiaobao Wu, Thong Nguyen, and Anh Tuan Luu. 2024a. [A survey on neural topic models: methods, applications, and challenges](#). *Artif. Intell. Rev.*, 57(2):18.
- Xiaobao Wu, Thong Nguyen, Delvin Zhang, William Yang Wang, and Anh Tuan Luu. 2024b. [Fastopic: Pretrained transformer is a fast, adaptive, stable, and transferable topic model](#). In *Advances in Neural Information Processing Systems 38: Annual Conference on Neural Information Processing Systems 2024, NeurIPS 2024, Vancouver, BC, Canada, December 10 - 15, 2024*.
- Minh Chu Xuan, Tien Phat Nguyen, Linh Ngo Van, Dinh Viet Sang, Nguyen Thi Ngoc Diep, and Trung Le. 2026. LLM-XTM: enhancing cross-lingual topic models with large language models. In *Proceedings of the 64th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 3719–3737. Association for Computational Linguistics.
- Hongliang Yan, Yukang Ding, Peihua Li, Qilong Wang, Yong Xu, and Wangmeng Zuo. 2017. [Mind the class weight bias: Weighted maximum mean discrepancy for unsupervised domain adaptation](#). In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 2272–2281.
- Xiaohui Yan, Jiafeng Guo, Yanyan Lan, and Xueqi Cheng. 2013. [A bitern topic model for short texts](#). In *Proceedings of the 22nd International Conference on World Wide Web, WWW '13*, page 1445–1456, New York, NY, USA. Association for Computing Machinery.
- Xiaohao Yang, He Zhao, Dinh Phung, Wray Buntine, and Lan Du. 2025a. [Llm reading tea leaves: Automatically evaluating topic models with large language models](#). *Transactions of the Association for Computational Linguistics*, 13:357–375.
- Xiaohao Yang, He Zhao, Weijie Xu, Yuanyuan Qi, Jueqing Lu, Dinh Phung, and Lan Du. 2025b. [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), ACL 2025, Vienna, Austria, July 27 - August 1, 2025*, pages 1377–1401. Association for Computational Linguistics.
- Xiang Zhang, Junbo Zhao, and Yann LeCun. 2015. Character-level convolutional networks for text classification. In *Proceedings of the 29th International Conference on Neural Information Processing Systems - Volume 1, NIPS'15*, page 649–657, Cambridge, MA, USA. MIT Press.
- Yanzhao Zhang, Mingxin Li, Dingkun Long, Xin Zhang, Huan Lin, Baosong Yang, Pengjun Xie, An Yang, Dayiheng Liu, Junyang Lin, Fei Huang, and Jingren Zhou. 2025. [Qwen3 embedding: Advancing text embedding and reranking through foundation models](#). *Preprint*, arXiv:2506.05176.
- He Zhao, Dinh Phung, Viet Huynh, Trung Le, and Wray L. Buntine. 2021. [Neural topic model via optimal transport](#). In *9th International Conference on Learning Representations, ICLR 2021, Virtual Event, Austria, May 3-7, 2021*. OpenReview.net.

A Related Work

Topic Modeling. Classical topic models, such as Latent Dirichlet Allocation (LDA) (Blei et al., 2003) and its variants (Blei and Lafferty, 2006; Rosen-Zvi et al., 2004; Yan et al., 2013), formulate documents as mixtures of latent topics under Bayesian generative assumptions. While effective and interpretable, these models rely primarily on word co-occurrence statistics and are limited in capturing contextual semantics. Recent surveys (Wu et al., 2024a) provide a comprehensive overview of these limitations and the evolution toward neural approaches.

Neural Topic Models (NTMs) address this limitation by incorporating deep neural networks, often within variational autoencoder frameworks (Miao et al., 2017; Srivastava and Sutton, 2017; Card et al., 2018; Dieng et al., 2020). Subsequent work further enhances topic representations using pre-trained embeddings and contextualized language models (Bianchi et al., 2021a,b; Hoyle et al., 2020), as well as alternative objectives such as optimal transport (Zhao et al., 2021; Wu et al., 2023) and contrastive learning to mitigate data sparsity in short texts (Nguyen and Luu, 2021; Nguyen et al., 2024; Wu et al., 2022). Related studies also explore inducing topics by clustering contextualized word representations (Thompson and Mimno, 2020). In parallel, clustering-based approaches such as Top2Vec (Angelov, 2020) and BERTopic (Groo-tendorst, 2022) derive topics directly from dense embedding spaces, offering flexibility at the cost of explicit probabilistic modeling.

A large and rapidly growing body of work improves NTMs along axes that are orthogonal to the refinement problem we study, and it is useful to situate **BERT-NTM** against these directions. One line of work modifies the *training objective or optimization geometry* of an otherwise unchanged NTM: NeuroMax (Pham et al., 2024b) maximizes the mutual information between the topic representation produced by the NTM encoder and a representation derived from a pretrained language model, together with a group topic regularizer; sharpness-aware minimization has been used to steer NTM training toward flatter minima and thereby yield higher-quality document representations (Nguyen et al., 2025d); and multi-objective and surrogate-objective formulations reconcile the competing terms that arise when auxiliary losses are added to the evidence lower bound (Le et al., 2025; Khanh

et al., 2026a). A second line replaces the reconstruction geometry itself, using optimal transport together with contrastive learning and pretrained embeddings to reshape document–topic relationships (Vuong et al., 2025; Khanh et al., 2026b). A third line targets *data regimes* in which word co-occurrence is unreliable, most prominently short texts, by borrowing context from document clusters or from global corpus statistics (Nguyen et al., 2025a; Vu et al., 2025; Nguyen et al., 2026c), or by changing the unit of modeling from the document to finer-grained sub-document spans (Nguyen et al., 2025c). A fourth line extends topic models across languages, aligning topic and representation spaces or sharing a global context space so that topics remain comparable between corpora (Nguyen et al., 2025b; Phat et al., 2026; Nguyen et al., 2026b), including recent work that uses LLMs in a black-box manner with self-consistency-based uncertainty quantification to refine cross-lingual topics (Xuan et al., 2026). Finally, dynamic topic models have been revisited with objectives that explicitly promote temporal consistency and interpretability rather than per-slice coherence alone (Nguyen et al., 2026a).

Two observations follow from this survey. First, almost all of these methods intervene *during* the training of the base topic model, by adding regularizers, changing the transport cost, or altering the optimizer; consequently they are complementary to **BERT-NTM**, which operates as a post-hoc refinement phase on top of an already-trained model and therefore composes with any of them. Second, where these methods do inject knowledge from pretrained language models (Pham et al., 2024b; Vuong et al., 2025; Nguyen et al., 2025d), they do so through *representation-level* signals — embeddings of documents or words that are matched, contrasted, or transported. **BERT-NTM** instead extracts a *distributional* signal from the encoder’s masked language modeling head: it scores existing vocabulary items in a topic-conditioned context. This is what allows the semantic signal to be applied directly to the topic–word distribution β rather than only to the latent document representation, and it is why the refinement never introduces out-of-vocabulary content.

Pretrained Encoders and Embedding Models. **BERT-NTM** relies on two capabilities of pretrained encoders: a masked language modeling head that assigns probability to a masked position

given a topic-conditioned context, and a sentence-level embedding space in which document semantics can be compared. Both capabilities have themselves become the subject of a substantial literature. Work on embedding-model distillation compresses large encoders into smaller ones while preserving the geometry of the embedding space, using intra-model relational constraints and optimal-transport alignment (Truong et al., 2025), teacher-anchored layer alignment combined with adaptive sharpness-aware minimization (Dao et al., 2026), mixtures over teacher layers when the teacher and student use different tokenizers (Vu et al., 2026), or span-aware and Matryoshka-style objectives that make a single model usable at multiple embedding dimensionalities (Huy et al., 2026; Tran et al., 2026); related ideas transfer to multimodal embedding models through hierarchical relational distillation (Le et al., 2026). Downstream, the quality of these embedding spaces directly determines retrieval behavior, including in cross-lingual and graph-structured retrieval-augmented pipelines (Hieu et al., 2025). This literature is relevant to us in two ways. It supports our design choice of treating the encoder as a fixed scoring and embedding module rather than fine-tuning it, since the encoder’s value here lies precisely in the pretrained geometry that these methods work to preserve. It also indicates a practical path to reducing the cost of our document–topic alignment term further: because Section G.4 shows that **BERT-NTM** is robust to the choice of embedding backbone, a distilled encoder can be substituted for a larger one with little expected loss.

Incorporating LLMs in Topic Modeling. Large language models (LLMs) have recently been incorporated into topic modeling in multiple ways. Prior work explores using LLMs such as ChatGPT to generate natural-language descriptions for topics, improving interpretability (OpenAI, 2023; Rijcken et al., 2023). LLMs have also been applied as evaluators for topic quality, including word intrusion and topic rating tasks (Rahimi et al., 2024; Stammbach et al., 2023), as well as for document-level keyword generation (Yang et al., 2025a).

More recent studies propose LLM-based topic modeling frameworks that prompt LLMs to generate topic representations and assign topics to documents (Wang et al., 2023a; Pham et al., 2024a; Mu et al., 2024; Doi et al., 2024). These methods typically rely on document-level LLM inference and incur substantial computational cost. LLM-

ITL (Yang et al., 2025b) mitigates this issue by shifting refinement to the topic level, demonstrating that LLMs can effectively improve topic coherence through iterative refinement.

Complementary perspectives further suggest that LLMs may implicitly perform topic inference during in-context learning, providing a conceptual bridge between topic modeling and general-purpose language modeling (Wang et al., 2023b).

MMD and Distribution Alignment. Maximum Mean Discrepancy (MMD) measures distributional differences by comparing kernel mean embeddings in a reproducing kernel Hilbert space, providing a principled framework for two-sample testing and distribution alignment and exhibiting close connections to energy-distance statistics (Gretton et al., 2012; Muandet et al., 2017; Sejdinovic et al., 2013). In practice, kernel bandwidth selection plays a critical role; common strategies include the median heuristic and multi-kernel formulations that aggregate multiple scales to improve robustness (Garreau et al., 2017; Long et al., 2015).

MMD has been widely adopted as an alignment objective in representation learning, particularly for unsupervised domain adaptation, where minimizing MMD between feature distributions encourages domain-invariant representations and can be extended via weighted variants to account for distributional imbalance (Long et al., 2015; Yan et al., 2017). Beyond alignment, MMD has also been used as a training objective for generative models, such as Generative Moment Matching Networks and MMD-GANs, demonstrating its effectiveness for distribution matching in deep learning (Li et al., 2015, 2017).

B Algorithm

In this section, we present the **BERT-NTM** training procedure 1:

C Details of Dataset

Dataset	# Docs Train	# Docs Test	Vocab Size	Avg. Doc Length	# Labels
20News	11,778	2,944	13,925	150	20
R8	5,485	2,189	5,338	102	8
DBpedia	15,598	3,899	8,550	51	14
AGNews	16,000	4,000	8,389	38	4

Table D1: Statistics of the datasets after preprocessing.

We conduct experiments on four widely used benchmark datasets for topic modeling, namely 20News, R8, DBpedia, and AGNews. For DBpedia and AGNews, we randomly sample 20,000

Algorithm 1 BERT-NTM Training Procedure

Require: Documents $\mathcal{D}$ with BoW representations $X = \{x_d\}_{d=1}^D$, number of topics K , candidate size N , pretrained encoder $f(\cdot)$, pretrained BERT MLM, loss weights $\lambda_{\text{dec}}, \lambda_{\text{enc}}$

Ensure: Refined topic model parameters Θ^*

- 1: Initialize neural topic model parameters Θ
- 2: **/* Phase 1: Base Topic Model Training */**
- 3: **for** epoch = 1 to T_1 **do**
- 4: Update Θ by minimizing the base topic model loss $\mathcal{L}_{\text{TM}}$
- 5: **end for**
- 6: **/* Phase 2: Semantic Refinement */**
- 7: **for** epoch = 1 to T_2 **do**
- 8: Infer topic–word distributions $\{\beta_k\}_{k=1}^K$ and document–topic proportions $\{\theta_d\}_{d=1}^D$
- 9: **for** each topic $k = 1$ to K **do**
- 10: Select top- N words C_k from β_k
- 11: Construct topic-induced masked context $S_k(\cdot)$
- 12: Compute BERT MLM scores $s_k(w)$ for $w \in C_k$
- 13: Form semantic distribution $\hat{\beta}_k^{\text{sem}}$ via softmax
- 14: Compute decoder-side loss $\mathcal{L}_{\text{sem}}^{(k)}$
- 15: **end for**
- 16: **for** each document d **do**
- 17: Encode document embedding $\mathbf{z}_d = f(d)$
- 18: Encode topic embeddings $\{\mathbf{z}_k\}_{k=1}^K$
- 19: Compute encoder-side MMD loss $\mathcal{L}_{\text{enc}}^{(d)}$
- 20: **end for**
- 21: Update Θ by minimizing:

$$\mathcal{L} = \mathcal{L}_{\text{TM}} + \lambda_{\text{dec}} \sum_k \mathcal{L}_{\text{sem}}^{(k)} + \lambda_{\text{enc}} \sum_d \mathcal{L}_{\text{enc}}^{(d)}$$

22: **end for**

documents to construct manageable experimental subsets.

For models that operate directly on textual input, we use the original documents without modification. We use the same preprocessed versions of the datasets released by prior work on LLM-based iterative topic labeling (Yang et al., 2025b). All preprocessing steps and dataset statistics are kept unchanged, and the resulting statistics are reported in Table D1. This keeps the corpus and evaluation setting aligned with Yang et al. (Yang et al., 2025b) for comparisons involving values cited from their

paper.

D Baselines and Experimental Protocol

We compare **BERT-NTM** with a diverse set of topic modeling baselines spanning classical probabilistic models, neural topic models, embedding-based approaches, and recent LLM-based methods. To ensure a fair and consistent comparison, we partially follow the evaluation protocol of the LLM-ITL framework for models whose results are directly reported from prior work.

Specifically, we report the results of Latent Dirichlet Allocation (**LDA**) (Blei et al., 2003) and **TopicGPT** (Pham et al., 2024a) as provided by Yang et al. (Yang et al., 2025b), without re-running these models. These values are included for quality-metric comparison because our experiments use the same released preprocessed datasets and corresponding evaluation setting. LDA is included as a classical probabilistic baseline, while TopicGPT represents a recent LLM-based topic modeling approach that relies on large language models for topic generation and document-level topic assignment. Neither cited row is used in our hardware-dependent runtime or memory comparisons.

For all remaining baselines, we conduct our own experiments using publicly available implementations. These include the Neural Variational Document Model (**NVDM**) (Miao et al., 2017), **SCHOLAR** (Card et al., 2018), the Embedded Topic Model (**ETM**) (Dieng et al., 2020), Neural Sinkhorn Topic Model (**NSTM**) (Zhao et al., 2021), Contrastive Learning Neural Topic Model (**CLNTM**) (Nguyen and Luu, 2021), WeTe (dongsheng wang et al., 2022), Embedding Clustering Regularization Topic Model (**ECRTM**) (Wu et al., 2023), and **BERTopic** (Grootendorst, 2022). For these models, hyperparameters are tuned separately for each dataset following standard practice.

In addition, for each neural topic model listed above, we evaluate both its original version and its refined variant under the **BERT-NTM** framework. We further compare these results with their counterparts refined using the **LLM-ITL** (Yang et al., 2025b) framework, enabling a direct and controlled comparison between encoder-based semantic refinement and generative LLM-based refinement under identical base topic models.

Except for the cited LDA and TopicGPT rows, all methods evaluated in our environment, including LLM-ITL (Yang et al., 2025b) and **BERT-**

NTM, are run five times with different random seeds, and we report the mean of the evaluation metrics.

E Detailed Prompts for LLM Evaluation

This appendix provides the detailed system prompts used for the LLM-based evaluation tasks described in the main text. The prompts are reported in Table D2.

F Implementation Details

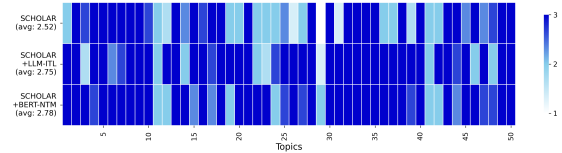
We run experiments on a single NVIDIA A40 GPU. Training is divided into two phases: a **base training** phase followed by a **semantic refinement** phase of 50 epochs.

The two phases share a single, fixed epoch budget: the total number of epochs is set per (backbone, dataset) pair and does not depend on whether refinement is enabled. Concretely, the base phase runs for $T - 50$ epochs and the refinement phase occupies the last 50 epochs of the same total T ; refinement does not add 50 epochs on top of the base schedule. The unrefined rows of Table 1 are therefore trained for exactly the same number of epochs as the corresponding refined rows, and the reported improvements cannot be attributed to additional optimization steps. LLM-ITL is run in our environment under the same total epoch budget, with its own refinement phase occupying the same final window. We discuss this further in Appendix G.19.

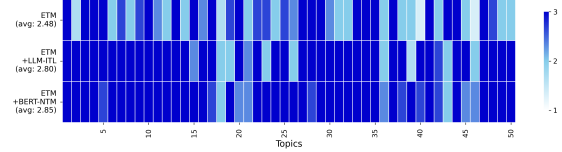
For base topic model training, we use dataset-specific hyperparameters following standard practice. Learning rates range from 0.0001 to 0.005, batch sizes from 200 to 2000, and total training epochs from 150 to 550 depending on the model and dataset combination.

For semantic refinement, we use the following settings:

- BERT model: bert-base-uncased for MLM-based semantic scoring
- Top- N candidate words: 25 words per topic for semantic alignment
- BERT distribution caching: recompute every 5 epochs to improve efficiency
- Refinement weights: $\lambda_{\text{dec}} \in \{1.0, 1.5, 1.75\}$ and $\lambda_{\text{enc}} \in \{0.25, 0.5, 1.0, 1.5\}$ tuned per dataset
- Temperature parameters: $\tau \in \{0.5, 1.0, 1.5\}$ for BERT softmax



(a) 20 Newsgroups



(b) Reuters-21578 (R8)

Figure 6: LLM-based topic coherence evaluation. Each heatmap visualizes LLM-assigned coherence scores for topic words across different models on two datasets.

- Embedding model for MMD: Qwen3-Embedding-0.6B with RBF kernel bandwidth $\sigma = 1.0$

G More Results

G.1 LLM-based Topic Quality Evaluation

After establishing internal robustness, we further evaluate topic quality using a large language model (LLM) as a semantic evaluator that assigns coherence scores to the top words of each topic on a discrete scale, providing a complementary signal to automatic coherence metrics; detailed prompts are given in Appendix E. Figure 6 shows the resulting LLM-based scores across topics for representative base models and their refined variants. Figure 6a and Figure 6b correspond to 20 Newsgroups and Reuters-21578 (R8), respectively. On 20 Newsgroups, the average score increases from 2.52 for SCHOLAR to 2.75 with LLM-ITL and to 2.78 with **BERT-NTM**, while on Reuters-21578 (R8), ETM improves from 2.48 to 2.80 with LLM-ITL and to 2.85 with **BERT-NTM**. Across both datasets, **BERT-NTM** achieves higher average scores and reduces the prevalence of low-coherence topics, indicating more consistent topic quality as judged by the LLM.

G.2 Vocabulary Mismatch and Invalid LLM Refinements

To examine practical failure modes of generative refinement, we analyze the raw topic-word outputs produced by LLM-ITL before its in-vocabulary filtering step. LLM-ITL filters words outside the original corpus vocabulary before using them for refinement; therefore, OOV words are not directly

Dataset	Prompt
20News	You are a helpful assistant evaluating the top words of a topic model output. The dataset is 20 Newsgroups , which consists of online forum posts covering 20 discussion topics (e.g., politics, religion, sports, science, and computer technology). Please rate how semantically related the following words are to each other as a coherent topic on a scale from 1 to 3 ("1" = not very related, "2" = moderately related, "3" = very related). Reply with a single number indicating the overall coherence of the topic.
R8	You are a helpful assistant evaluating the top words of a topic model output. The dataset is Reuters-21578 (R8) , a collection of Reuters news articles mainly about finance and economics (e.g., earnings, trade, money markets, and acquisitions). Please rate how semantically related the following words are to each other as a coherent topic on a scale from 1 to 3 ("1" = not very related, "2" = moderately related, "3" = very related). Reply with a single number indicating the overall coherence of the topic.
DBpedia	You are a helpful assistant evaluating the top words of a topic model output. The dataset is DBpedia , which contains short descriptions of entities (e.g., people, places, organizations, works) derived from Wikipedia/DBpedia. Please rate how semantically related the following words are to each other as a coherent topic on a scale from 1 to 3 ("1" = not very related, "2" = moderately related, "3" = very related). Reply with a single number indicating the overall coherence of the topic.
AGNews	You are a helpful assistant evaluating the top words of a topic model output. The dataset is AG News , a news classification corpus with four categories: World, Sports, Business, and Science/Technology. Please rate how semantically related the following words are to each other as a coherent topic on a scale from 1 to 3 ("1" = not very related, "2" = moderately related, "3" = very related). Reply with a single number indicating the overall coherence of the topic.

Table D2: Coherence prompts for LLM-based topic evaluation.

passed into the topic model. Nevertheless, filtering can leave fewer usable semantic guidance words than requested.

In our measurements, approximately 20% of generated refined-topic outputs contain at least one OOV word before filtering. In these cases, removed words reduce the size of the usable refined-word set, potentially weakening the semantic signal available for refinement. We additionally observe that approximately 5% of generated outputs are invalid: either no in-vocabulary refined word remains after filtering or the generative LLM returns a safety-refusal response rather than a topic-word list. These outputs cannot be used and are discarded and regenerated, adding generation calls, runtime overhead, and output variability.

In contrast, **BERT-NTM** scores only candidate words drawn from the original topic-model vocabulary. Its OOV rate is therefore zero by construction, and it does not require post-generation filtering or regeneration due to invalid generative responses.

Top-M	20News		R8	
	C_v	PN	C_v	PN
5	0.571	0.405	0.604	0.728
10	0.578	0.410	0.616	0.727
20	0.566	0.401	0.590	0.700
30	0.554	0.392	0.582	0.686

Table E1: Top- M sensitivity on ETM + **BERT-NTM**. A moderate Top- M gives the best overall balance.

G.3 Top- M Sensitivity Analysis

Table E1 shows that performance is strongest at a moderate Top- M . Top- $M=10$ gives the best overall balance, while Top- $M=5$ remains competitive, suggesting that the core semantic signal is concentrated in a relatively small set of top topic words. Increasing Top- M beyond 10 gradually degrades both C_v and PN, likely because lower-ranked words introduce noise into the topic embedding.

Encoder for MMD	20News		R8	
	C_V	PN	C_V	PN
w/o Doc-Topic Align.	0.535	0.410	0.596	0.667
bert-base-uncased	0.574	0.407	0.612	0.696
all-MiniLM-L6-v2	0.577	0.405	0.609	0.713
e5-small-v2	0.573	0.409	0.614	0.719
Qwen3-Embedding-0.6B	0.578	0.410	0.616	0.727

Table E2: Sensitivity to the embedding encoder used for MMD-based document–topic alignment under the ETM backbone.

G.4 Encoder Sensitivity Analysis

BERT-NTM uses pretrained encoders in two distinct components. The MLM-based semantic scoring component uses bert-base-uncased by default to score candidate topic words in a masked topic context. Separately, the document–topic alignment component uses an embedding encoder to construct document and topic representations for the MMD objective; our default is Qwen3-Embedding-0.6B. We study the sensitivity of both choices using the ETM backbone on 20News and R8.

Encoder for document–topic alignment. We keep the MMD objective and all other settings fixed, varying only the embedding encoder used to construct document and topic representations. Table E2 shows that MMD-based alignment improves topic coherence across encoder choices. Compared with removing document–topic alignment, all tested encoders improve C_V on both datasets. On R8, alignment also improves PN from 0.667 to 0.696–0.727. Lightweight encoders already yield clear gains, while Qwen3-Embedding-0.6B achieves the best overall performance.

Encoder for MLM-based semantic scoring. We next keep Qwen3-Embedding-0.6B as the MMD encoder and vary only the masked language model used for topic-conditioned candidate-word scoring. Table E3 shows small variation across MLM scorers: all variants maintain strong topic coherence on both datasets, and deberta-v3-base yields modest additional gains. Together with the no-semantic-scoring result in Table 2, these results indicate that the main gain comes from the masked semantic scoring mechanism rather than a single MLM checkpoint.

Overall, the framework is not tied to a single pretrained encoder checkpoint: both the MMD embedding encoder and the MLM scorer can be replaced

MLM Scorer	20News		R8	
	C_V	PN	C_V	PN
bert-base-uncased	0.578	0.410	0.616	0.727
roberta-base	0.578	0.406	0.619	0.722
deberta-v3-base	0.584	0.411	0.622	0.729

Table E3: Sensitivity to the masked language model used for topic-conditioned semantic scoring under the ETM backbone, with Qwen3-Embedding-0.6B fixed for MMD alignment.

while retaining the refinement benefit. Nevertheless, encoder quality affects the final performance, with stronger encoders providing modest improvements.

G.5 Qualitative Effect of Semantic Refinement on Topic–Word Distributions

To illustrate how semantic refinement influences topic–word distributions, we present qualitative examples from the AGNews dataset using the SCHOLAR backbone.

In neural topic modeling, topic–word distributions are primarily determined by corpus-level co-occurrence statistics. In BERT-NTM, the pretrained encoder provides an additional semantic compatibility signal during refinement, which adjusts the relative importance of candidate words without replacing the corpus-driven learning process.

Refinement preserves the core semantic theme of each topic while improving interpretability through small, targeted adjustments, as illustrated in Table E4. In both examples, approximately 60% of the original top words remain after refinement, indicating that corpus-induced topic structure is largely retained. The main change involves suppressing weakly compatible outliers (e.g., named entities unrelated to the dominant theme) and introducing more general on-theme tokens, resulting in clearer and more coherent topic representations.

G.6 Topic Coherence at Different K

We illustrate the topic coherence performance (C_V) of topic models across different numbers of topics (i.e., K) in Table E5. Overall, **BERT-NTM** consistently achieves the highest topic coherence across all datasets and values of K , demonstrating that semantic refinement remains effective across different topic granularities.

The narrower margin over LLM-ITL in some large- K settings should be interpreted in terms of

	Topic: Legal / Trials (Topic 21)	Topic: Elections / Politics (Topic 10)
Phase-1 top words	court, trial, charge, accused, judge, abu, case, prison, kobe, bryant	presidential, kerry, bush, election, hamid, john, democratic, karzai, president, vote
BERT compatibility	0.205, 0.545, 0.054, 0.041, 0.098, 0.000, 0.021, 0.036, 0.001, 0.000	0.020, 0.026, 0.700, 0.003, 0.000, 0.008, 0.012, 0.000, 0.230, 0.001
After refinement phase	trial, court, sentence, prison, accused, charge, judge, guilty, appeal, jail	bush, republican, democratic, president, kerry, democrat, presidential, election, candidate, party

Table E4: Qualitative comparison of topic–word distributions before and after semantic refinement. Compatibility scores indicate semantic consistency within each topic context.

semantic faithfulness rather than as a weakness of **BERT-NTM**. When K exceeds the effective semantic granularity of a corpus, particularly for datasets with relatively few underlying categories such as AGNews and R8, the corpus-induced topics become increasingly fragmented and overlapping. In this regime, the top words of a base topic may no longer provide a well-defined semantic context. A generative refinement method can introduce alternative in-vocabulary words that produce a more coherent-looking word list, but this can shift the refined representation away from the content actually supported by the original corpus-induced topic. By contrast, **BERT-NTM** deliberately preserves topic content: it scores and reweights corpus-supported candidates and aligns representations using the base topic context, rather than rewriting the topic with newly proposed words. Thus, when the base context is already semantically weak under an over-specified topic space, a smaller additional gain is consistent with preserving topic fidelity. Importantly, **BERT-NTM** remains competitive with, and often outperforms, LLM-ITL in these large- K settings, showing that it can improve coherence without replacing the semantic content induced from the corpus.

G.7 Topic Alignment at Different K

We illustrate the topic alignment performance (PN) of topic models across different numbers of topics (i.e., K) in Table E6. Overall, both refinement frameworks maintain topic alignment comparable to or better than base models, with **BERT-NTM** showing slight advantages in most settings.

G.8 Topic Diversity (TD)

We evaluate the performance of topic diversity (TD) defined by Dieng et al. (2020), which evaluates the distinctiveness of the top-words within the learned topics. The results across different numbers of topics (i.e., K) is illustrated in Table E7. Over-

all, topic diversity exhibits a natural trade-off with topic coherence, where both refinement methods reduce diversity compared to base models, though **BERT-NTM** often achieves a better balance than LLM-ITL.

G.9 Purity & NMI

As PN is defined as the mean of Purity and NMI for document–topic evaluation, we provide detailed Purity and NMI results across different numbers of topics in Table E8 and Table E10, respectively.

G.10 Reporting NPMI alongside C_V

For completeness and comparability with prior work, we additionally report Normalized Pointwise Mutual Information (NPMI), a commonly used topic coherence metric, alongside C_V . As shown in Table E9, **BERT-NTM** improves NPMI over both ETM and SCHOLAR across all datasets. These results indicate that the coherence gains are consistent across multiple evaluation metrics.

Model	20News	AGNews	DBpedia	R8
ETM (Dieng et al., 2020)	0.0316	0.0591	0.0458	-0.0099
+BERT-NTM	0.0641	0.1104	0.1125	0.0630
+LLM-ITL	0.0601	0.1102	0.1104	0.0470
SCHOLAR (Card et al., 2018)	-0.0021	0.0637	0.1284	-0.0390
+BERT-NTM	0.1040	0.1367	0.1412	-0.0016
+LLM-ITL	0.0528	0.0784	0.0654	-0.0181

Table E9: NPMI performance across datasets for ETM and SCHOLAR backbones. Higher values indicate better topic coherence.

G.11 NMI

Table E10 reports detailed NMI results across different numbers of topics for document–topic alignment evaluation.

G.12 Ablation on Word Order

We examine the effect of word ordering when constructing topic embeddings. We compare preserving the original ranking, reversing the order, and randomly shuffling topic words.

Model	20News			R8			DBpedia			AGNews		
	K=25	K=75	K=100	K=25	K=75	K=100	K=50	K=75	K=100	K=50	K=75	K=100
NVDM (Miao et al., 2017)	0.299	0.241	0.232	0.324	0.283	0.270	0.274	0.260	0.259	0.289	0.266	0.256
+ LLM-ITL	0.439	0.298	0.292	0.438	0.358	0.340	0.365	0.353	0.346	0.357	0.326	0.325
+ BERT-NTM	0.421	0.329	0.328	0.448	0.389	0.367	0.384	0.365	0.351	0.382	0.325	0.316
SCHOLAR (Card et al., 2018)	0.520	0.478	0.472	0.399	0.385	0.376	0.434	0.422	0.403	0.577	0.581	0.541
+ LLM-ITL	0.622	0.584	0.562	0.462	0.437	0.424	0.564	0.565	0.556	0.643	0.647	0.627
+ BERT-NTM	0.638	0.615	0.579	0.471	0.435	0.427	0.559	0.535	0.533	0.606	0.556	0.543
ETM (Dieng et al., 2020)	0.511	0.466	0.461	0.447	0.432	0.431	0.507	0.501	0.496	0.497	0.497	0.480
+ LLM-ITL	0.627	0.566	0.553	0.570	0.565	0.580	0.657	0.631	0.626	0.626	0.620	0.618
+ BERT-NTM	0.535	0.527	0.524	0.502	0.511	0.554	0.609	0.600	0.618	0.587	0.585	0.598
NSTM (Zhao et al., 2021)	0.450	0.425	0.424	0.415	0.407	0.402	0.636	0.621	0.616	0.596	0.583	0.575
+ LLM-ITL	0.510	0.524	0.518	0.549	0.544	0.534	0.702	0.688	0.690	0.661	0.670	0.649
+ BERT-NTM	0.578	0.554	0.511	0.589	0.573	0.572	0.735	0.715	0.716	0.667	0.637	0.652

Table E5: Topic coherence (C_V) performance across different numbers of topics (K). For 20News and R8, we show $K=25, 75, 100$ ($K=50$ is in main results). For DBpedia and AGNews, we show $K=50, 75, 100$ ($K=25$ is in main results). **Bold** values denote the highest score achieved within each model family on a given dataset.

Model	20News			R8			DBpedia			AGNews		
	K=25	K=75	K=100	K=25	K=75	K=100	K=50	K=75	K=100	K=50	K=75	K=100
NVDM (Miao et al., 2017)	0.146	0.145	0.153	0.380	0.353	0.359	0.150	0.151	0.156	0.223	0.212	0.211
+ LLM-ITL	0.143	0.142	0.153	0.377	0.355	0.358	0.150	0.150	0.152	0.222	0.212	0.211
+ BERT-NTM	0.141	0.148	0.155	0.378	0.358	0.365	0.152	0.149	0.150	0.227	0.218	0.214
SCHOLAR (Card et al., 2018)	0.572	0.580	0.588	0.717	0.686	0.681	0.827	0.794	0.680	0.587	0.426	0.389
+ LLM-ITL	0.562	0.578	0.587	0.716	0.682	0.675	0.828	0.773	0.701	0.567	0.410	0.382
+ BERT-NTM	0.580	0.593	0.598	0.732	0.699	0.693	0.842	0.821	0.699	0.578	0.419	0.395
ETM (Dieng et al., 2020)	0.363	0.434	0.456	0.679	0.666	0.664	0.751	0.747	0.739	0.553	0.555	0.549
+ LLM-ITL	0.352	0.426	0.441	0.690	0.673	0.666	0.740	0.738	0.744	0.558	0.560	0.553
+ BERT-NTM	0.357	0.441	0.465	0.693	0.679	0.673	0.765	0.761	0.755	0.571	0.568	0.560
NSTM (Zhao et al., 2021)	0.365	0.382	0.405	0.674	0.662	0.662	0.739	0.739	0.738	0.576	0.570	0.569
+ LLM-ITL	0.351	0.366	0.384	0.681	0.669	0.669	0.741	0.738	0.736	0.576	0.566	0.566
+ BERT-NTM	0.356	0.385	0.411	0.689	0.677	0.675	0.754	0.752	0.751	0.589	0.581	0.579

Table E6: Topic alignment (PN) performance across different numbers of topics (K). For 20News and R8, we show $K=25, 75, 100$ ($K=50$ is in main results). For DBpedia and AGNews, we show $K=50, 75, 100$ ($K=25$ is in main results). **Bold** values denote the highest score achieved within each model family on a given dataset.

Model	20News				R8				DBpedia				AGNews			
	K=25	K=50	K=75	K=100	K=25	K=50	K=75	K=100	K=25	K=50	K=75	K=100	K=25	K=50	K=75	K=100
NVDM (Miao et al., 2017)	0.993	0.985	0.956	0.931	0.802	0.801	0.789	0.790	0.982	0.964	0.919	0.875	0.962	0.940	0.896	0.847
+ LLM-ITL	0.970	0.966	0.930	0.896	0.834	0.811	0.765	0.714	0.981	0.940	0.878	0.829	0.962	0.929	0.871	0.813
+ BERT-NTM	0.978	0.972	0.948	0.918	0.818	0.798	0.782	0.776	0.978	0.952	0.905	0.868	0.958	0.936	0.888	0.841
SCHOLAR (Card et al., 2018)	0.986	0.920	0.860	0.814	0.695	0.509	0.411	0.363	0.965	0.491	0.319	0.246	0.889	0.404	0.272	0.220
+ LLM-ITL	0.936	0.846	0.779	0.718	0.699	0.489	0.386	0.345	0.926	0.520	0.392	0.330	0.820	0.427	0.300	0.274
+ BERT-NTM	0.962	0.895	0.831	0.796	0.691	0.518	0.419	0.378	0.952	0.512	0.348	0.281	0.928	0.526	0.408	0.362
ETM (Dieng et al., 2020)	0.890	0.786	0.687	0.604	0.688	0.521	0.444	0.373	0.907	0.864	0.762	0.644	0.932	0.828	0.738	0.668
+ LLM-ITL	0.940	0.832	0.754	0.695	0.676	0.518	0.441	0.377	0.974	0.909	0.837	0.770	0.909	0.733	0.636	0.578
+ BERT-NTM	0.912	0.821	0.743	0.672	0.712	0.548	0.471	0.412	0.936	0.892	0.823	0.751	0.945	0.814	0.756	0.694
NSTM (Zhao et al., 2021)	0.648	0.507	0.385	0.330	0.460	0.320	0.244	0.227	0.854	0.768	0.679	0.621	0.831	0.743	0.701	0.703
+ LLM-ITL	0.718	0.593	0.564	0.513	0.446	0.311	0.296	0.224	0.866	0.734	0.683	0.602	0.843	0.749	0.731	0.649
+ BERT-NTM	0.672	0.568	0.512	0.468	0.532	0.385	0.338	0.295	0.862	0.782	0.702	0.648	0.825	0.761	0.718	0.692

Table E7: Topic Diversity (TD) performance across different numbers of topics (K) for all datasets. **Bold** values denote the highest score achieved within each model family on a given dataset.

As shown in Table E11, performance remains nearly unchanged across all variants on both AGNews and R8, indicating that semantic refinement is largely insensitive to word order.

G.13 Choice of the Document–Topic Alignment Objective

Section 3 aligns the document representation with the topic mixture through a Maximum Mean Discrepancy (MMD) term. Because this is the one

Model	20News				R8				DBpedia				AGNews			
	K=25	K=50	K=75	K=100	K=25	K=50	K=75	K=100	K=25	K=50	K=75	K=100	K=25	K=50	K=75	K=100
NVDM (Miao et al., 2017)	0.166	0.165	0.166	0.170	0.610	0.600	0.596	0.603	0.210	0.193	0.199	0.204	0.432	0.402	0.389	0.389
+ LLM-ITL	0.162	0.163	0.166	0.172	0.610	0.603	0.598	0.602	0.206	0.195	0.197	0.200	0.428	0.401	0.388	0.389
+ BERT-NTM	0.164	0.168	0.175	0.178	0.615	0.612	0.603	0.615	0.218	0.202	0.198	0.197	0.444	0.415	0.406	0.401
SCHOLAR (Card et al., 2018)	0.597	0.627	0.634	0.650	0.921	0.911	0.930	0.928	0.872	0.864	0.822	0.690	0.855	0.790	0.570	0.533
+ LLM-ITL	0.580	0.607	0.631	0.649	0.920	0.911	0.927	0.923	0.875	0.865	0.798	0.709	0.855	0.765	0.546	0.523
+ BERT-NTM	0.609	0.651	0.658	0.669	0.937	0.941	0.951	0.944	0.909	0.882	0.867	0.742	0.870	0.689	0.565	0.548
ETM (Dieng et al., 2020)	0.355	0.410	0.444	0.471	0.872	0.875	0.887	0.877	0.794	0.813	0.817	0.812	0.783	0.781	0.789	0.787
+ LLM-ITL	0.348	0.403	0.436	0.457	0.860	0.875	0.870	0.861	0.760	0.781	0.798	0.811	0.784	0.784	0.794	0.790
+ BERT-NTM	0.368	0.448	0.471	0.488	0.889	0.891	0.898	0.886	0.811	0.825	0.836	0.859	0.809	0.803	0.808	0.810
NSTM (Zhao et al., 2021)	0.356	0.375	0.390	0.423	0.883	0.895	0.903	0.908	0.818	0.825	0.840	0.848	0.818	0.819	0.821	0.825
+ LLM-ITL	0.341	0.361	0.371	0.395	0.884	0.896	0.903	0.905	0.797	0.825	0.836	0.843	0.807	0.819	0.817	0.822
+ BERT-NTM	0.368	0.398	0.423	0.441	0.902	0.921	0.925	0.923	0.847	0.863	0.859	0.873	0.841	0.835	0.833	0.836

Table E8: Purity performance across different numbers of topics (K) for all datasets. **Bold** values denote the highest score achieved within each model family on a given dataset.

Model	20News				R8				DBpedia				AGNews			
	K=25	K=50	K=75	K=100	K=25	K=50	K=75	K=100	K=25	K=50	K=75	K=100	K=25	K=50	K=75	K=100
NVDM (Miao et al., 2017)	0.126	0.117	0.124	0.136	0.149	0.118	0.111	0.115	0.140	0.107	0.104	0.107	0.071	0.045	0.035	0.033
+ LLM-ITL	0.123	0.113	0.119	0.135	0.145	0.117	0.111	0.114	0.134	0.104	0.102	0.104	0.069	0.044	0.035	0.034
+ BERT-NTM	0.118	0.128	0.135	0.132	0.141	0.104	0.113	0.115	0.086	0.096	0.100	0.103	0.010	0.021	0.030	0.027
SCHOLAR (Card et al., 2018)	0.547	0.538	0.525	0.525	0.512	0.449	0.443	0.433	0.778	0.790	0.766	0.671	0.420	0.383	0.282	0.244
+ LLM-ITL	0.545	0.529	0.525	0.525	0.511	0.449	0.438	0.426	0.781	0.791	0.746	0.692	0.423	0.368	0.274	0.241
+ BERT-NTM	0.551	0.535	0.538	0.527	0.527	0.457	0.447	0.442	0.775	0.760	0.775	0.656	0.286	0.149	0.273	0.242
ETM (Dieng et al., 2020)	0.371	0.405	0.424	0.442	0.485	0.463	0.460	0.438	0.728	0.690	0.678	0.667	0.352	0.325	0.321	0.311
+ LLM-ITL	0.357	0.394	0.416	0.425	0.519	0.498	0.476	0.471	0.723	0.698	0.678	0.676	0.354	0.332	0.327	0.315
+ BERT-NTM	0.346	0.434	0.459	0.442	0.497	0.467	0.460	0.460	0.719	0.697	0.686	0.651	0.333	0.333	0.328	0.310
NSTM (Zhao et al., 2021)	0.375	0.371	0.375	0.388	0.465	0.435	0.422	0.415	0.717	0.653	0.638	0.628	0.368	0.334	0.319	0.313
+ LLM-ITL	0.362	0.359	0.362	0.373	0.477	0.450	0.435	0.433	0.706	0.656	0.639	0.630	0.362	0.333	0.316	0.311
+ BERT-NTM	0.344	0.372	0.399	0.381	0.476	0.433	0.429	0.427	0.661	0.641	0.645	0.629	0.337	0.327	0.329	0.322

Table E10: NMI performance across different numbers of topics (K) for all datasets. **Bold** values denote the highest score achieved within each model family on a given dataset.

component of BERT-NTM that acts on the encoder rather than on the topic–word distributions, we isolate its contribution here. Starting from the full model, we replace *only* the alignment loss $\mathcal{L}_{\text{enc}}$ and leave every other component — the base topic model, the masked–language–model scoring procedure, the semantic loss $\mathcal{L}_{\text{sem}}$, the loss weights and the training schedule — untouched. We compare four variants: (i) no document–topic alignment at all ($\lambda_{\text{enc}} = 0$); (ii) a cosine-similarity objective that pulls the document embedding towards the convex combination of topic embeddings; (iii) a mean-squared-error objective between the same two quantities; and (iv) the MMD objective used by BERT-NTM.

Table E12 reports the results on 20News and R8. Every alignment objective improves over having none, which confirms that the encoder-side signal is doing real work and is not merely a regularizer that could be dropped. The distributional objective is nevertheless the strongest of the three on both datasets and on both metrics: MMD gains +0.043 C_V on 20News and +0.020 C_V on R8 over the no-alignment variant, and +0.020/+0.007 C_V over the best point-wise alternative. The gap in topic

alignment (PN) on R8 is larger still (0.727 versus 0.716 for cosine and 0.667 without alignment). We attribute this to the fact that cosine similarity and MSE both compare two *points* in embedding space and therefore collapse the topic mixture to a single vector, whereas MMD compares the empirical measure of the document embedding against the mixture measure induced by θ_d over the topic embeddings, so it retains information about how the probability mass is distributed across topics.

The no-alignment row of Table E12 is the same configuration as the *w/o Doc-Topic Align.* row of Table E2; the two experiments are therefore complementary. Table E2 varies the encoder while holding the MMD objective fixed, and shows that the gain survives every encoder we tried; Table E12 varies the objective while holding the encoder fixed, and shows that the distributional form of the objective matters independently of which encoder produces the embeddings.

Relation to Table 2. The rows labelled *w/ Cosine* and *w/ MSE* in Table 2 concern a different substitution and are not comparable to the corresponding rows here. There, cosine similarity and

Dataset	Word Order	Topic Coherence (C_v)	Topic Alignment (PN)
AGNews	Preserve (Default)	0.6800	0.6410
	Reverse	0.6795	0.6407
	Shuffle (Random Permutation)	0.6803	0.6415
R8	Preserve (Default)	0.4780	0.6850
	Reverse	0.4778	0.6847
	Shuffle (Random Permutation)	0.4775	0.6849

Table E11: Ablation on word ordering for topic embedding construction.

Alignment objective	20News		R8	
	C_V	PN	C_V	PN
None ($\lambda_{\text{enc}} = 0$)	0.535	0.410	0.596	0.667
Cosine similarity	0.558	0.409	0.609	0.716
Mean squared error	0.549	0.407	0.603	0.688
MMD (full BERT-NTM)	0.578	0.410	0.616	0.727

Table E12: Effect of the document–topic alignment objective. Only $\mathcal{L}_{\text{enc}}$ is changed; all other components are held fixed. The distributional (MMD) objective dominates the point-wise alternatives on both datasets.

MSE replace the KL divergence in the *decoder-side* objective $\mathcal{L}_{\text{sem}}$, which compares two distributions over candidate words; here they replace the MMD in the *encoder-side* objective $\mathcal{L}_{\text{enc}}$, which compares a document embedding with a topic mixture. The two experiments therefore modify different terms of the same total loss, which is why the numbers differ.

G.14 Top- N Sensitivity Analysis

Top- N controls the size of the candidate pool over which the masked language model is queried and over which the semantic loss $\mathcal{L}_{\text{sem}}$ is defined; it should not be confused with Top- M (Appendix G.3), which controls how many words enter the topic embedding used by the alignment term. Figure 4 plots the sensitivity to Top- N ; because the exact values are easier to compare in numeric form, we tabulate the same runs in Table E13.

Two trends are visible. First, C_V increases monotonically with Top- N , from 0.449 at $N=5$ to 0.583 at $N=30$, with the bulk of the gain realised by $N=25$; a larger candidate pool gives the masked language model more topical context to condition on, and gives the semantic loss more words to reweight. Second, PN is largely flat, moving within a 0.026 band across the whole sweep, so document–topic assignment quality is not what drives the choice of N . Topic diversity behaves in the opposite direction to C_V : it is highest for

Top- N	PN	C_V	TD
5	0.692	0.449	0.644
10	0.670	0.508	0.620
15	0.666	0.549	0.644
20	0.680	0.551	0.594
25 (default)	0.690	0.580	0.576
30	0.671	0.583	0.590

Table E13: Sensitivity to the candidate-pool size Top- N , tabulating the Top- N panel of Figure 4. C_V improves with a larger candidate pool and saturates around the default $N=25$, while PN stays within a narrow band and TD moves in the opposite direction.

small pools and decreases as N grows, because refinement concentrates probability on a semantically compatible core that is partly shared between related topics. The default $N=25$ therefore sits at the point where C_V has essentially saturated but TD has not yet degraded further; moving to $N=30$ buys +0.003 C_V at the cost of 0.019 PN and additional MLM forward passes. The important observation for the claims of the paper is that the behaviour is smooth and monotone in N rather than peaked at one narrowly tuned value: no single candidate size is load-bearing, and any N in the upper half of the sweep yields comparable coherence.

G.15 The Candidate Set Is Dynamic, Not Fixed

A natural concern about restricting the semantic loss to the current Top- N words is that refinement could only ever reweight an initial candidate pool, and would therefore inherit whatever mistakes the base model made in selecting that pool. This is not what happens, for two reasons.

First, the candidate pool is not frozen. The base topic-model objective $\mathcal{L}_{\text{TM}}$ remains active throughout the refinement phase, so the *full* topic–word distribution β_k over the entire vocabulary continues to be updated by corpus statistics; only the auxiliary semantic term is restricted to the current candidates. The masked-language-model seman-

tic distribution is recomputed from the current β_k every five epochs, at which point the Top- N candidate set is re-drawn. A word that lies outside the candidate set early in refinement can therefore be promoted into it later by corpus-driven updates and then receive semantic supervision. The refresh interval is a hyperparameter; Figure 4 reports the effect of refining more or less frequently.

Second, the resulting movement in the visible topics is substantial rather than cosmetic. In both qualitative examples of Appendix G.5 (Table E4), only 6 of the original top-10 words survive refinement, i.e. roughly 40% of the final top words were not in the initial top-10 list. The legal/trials topic acquires *sentence*, *guilty*, *appeal* and *jail* while dropping the off-theme named entities *abu*, *kobe* and *bryant*; the elections topic acquires *republican*, *democrat*, *candidate* and *party* while dropping *hamid*, *karzai* and *john*. Refinement is thus reranking and repopulating the head of the distribution, not merely permuting it.

We nonetheless regard the vocabulary restriction as a deliberate design trade-off rather than a solved problem, and discuss its limits in Appendix G.17. What the two observations above establish is narrower: semantic supervision is local to the current candidate set at any given step, but the final topic vocabulary is not permanently confined to the initial Top- N .

G.16 Masking Protocol: Independent Candidate Scoring

Because the template in (3.1) contains several [MASK] tokens for multi-subword candidates, it is worth stating explicitly what is and is not masked simultaneously. Candidate words are scored *independently of one another*: each candidate w gets its own forward pass in which exactly one slot of the template is masked, and the remaining topic context words are always present as surface tokens. Consecutive masks appear only *within* that single slot, when the tokenizer splits one candidate into $m > 1$ subwords; the m masks then stand for the m subwords of the same word, not for m different candidates. The score of one candidate therefore never depends on which other candidates are being scored, and there is no inter-candidate mask interaction of the kind that would arise if the whole Top- N list were masked at once.

Two further biases are worth separating from this design. Positional effects and subword-frequency effects could in principle contaminate $s_k(w)$ even

under independent scoring. The robustness study in Table 3 addresses the first: preserving, reversing and shuffling the context order changes C_V and PN by at most 0.002 on either dataset (Appendix G.12 repeats this on AGNews and R8), so the score is not an artefact of where in the template a word sits. The length-normalized aggregation in (3.1) addresses the second, and Table 3 shows that the alternative aggregation rules (*sum*, *min*, first-subword) remain competitive, i.e. the signal is not an artefact of one particular normalization.

An alternative would be strictly sequential scoring, in which a candidate’s subwords are filled in one at a time with previously predicted subwords revealed. We did not adopt it: it multiplies the number of forward passes by the average subword length of a candidate, which erodes the efficiency advantage that motivates an encoder-based refiner in the first place (Figure 5), and the robustness results above give little reason to expect a systematic gain. We note it as a possible refinement of the scoring procedure rather than as a limitation of the objective.

G.17 Scope of the Vocabulary Constraint

BERT-NTM scores only words that already belong to the topic model’s vocabulary. We make explicit here what this does and does not cost.

The constraint is shared, not distinctive. Generative refiners are subject to the same restriction in practice, because the words they emit must be mapped back onto β_k to be usable. LLM-ITL filters generated words against the original corpus vocabulary before refinement, so out-of-vocabulary suggestions are discarded there as well. The difference is where the restriction is enforced. In BERT-NTM it holds by construction, and the out-of-vocabulary rate is exactly zero; in a generate-then-filter pipeline it is enforced after the fact, which is what produces the failure modes quantified in Appendix G.2: about 20% of generated outputs contain at least one out-of-vocabulary word and are silently thinned by filtering, and about 5% are unusable altogether and must be regenerated.

What the constraint genuinely rules out. Neither method can introduce a descriptor that the pre-processing pipeline removed from the vocabulary in the first place — for instance a canonical domain term pruned by a minimum-document-frequency cutoff. For such a term to appear in a refined topic,

the vocabulary itself has to be enlarged, which is orthogonal to how the refinement signal is computed: our objective would apply unchanged to a larger $\mathcal{V}$.

Domain-specific and low-resource corpora.

Two distinct risks arise here, and only one of them is about the candidate set. The first is that a small or highly technical vocabulary may simply not contain good descriptors, in which case no reweighting scheme — discriminative or generative — can recover them. The second is that a general-domain masked language model may assign poor compatibility scores to specialist terminology, so the semantic signal itself degrades. The second risk is addressable within our framework without changing the objective, because the scorer is a swappable component: Table E3 shows that bert-base-uncased, roberta-base and deberta-v3-base are interchangeable to within $0.006 C_V$, which suggests that substituting a domain-adapted masked language model is a matter of configuration rather than redesign. We have not run this experiment: all four of our corpora are general-domain, and a controlled study on a specialist corpus with a domain-adapted scorer remains future work. We therefore present the vocabulary restriction as a deliberate trade-off — corpus fidelity and zero out-of-vocabulary output in exchange for an upper bound on what refinement can express — and not as a property we have validated outside the general-domain setting.

G.18 Parameter Budget of the Comparison with LLM-ITL

Since BERT-NTM and LLM-ITL use different pretrained models, it is fair to ask whether the gap in Table 1 reflects the refinement mechanism or merely the capacity of the pretrained model each method leans on. Table E14 lists the budgets. BERT-NTM uses bert-base-uncased (110M) for masked scoring and Qwen3-Embedding-0.6B (0.6B, embedding only) for alignment, and neither component generates text; LLM-ITL uses instruction-tuned generative models in the 7B–8B range. The size difference therefore runs *against* BERT-NTM: it attains higher coherence with roughly an order of magnitude fewer parameters and no autoregressive decoding, so the improvement cannot be attributed to a larger pretrained model.

The encoder ablations settle the question more directly, since they vary capacity while holding

Method	Pretrained model	Params	Gen.
BERT-NTM	bert-base-uncased	110M	no
	Qwen3-Embedding-0.6B	0.6B	no
LLM-ITL	Llama-3-8B-Instruct	8B	yes
	Mistral-7B-Instruct-v0.3	7B	yes

Table E14: Pretrained models used by each refinement method. “Gen.” indicates whether the component performs autoregressive generation. BERT-NTM uses substantially smaller models and never generates text.

the mechanism fixed. On the alignment side (Table E2), replacing the 0.6B embedding encoder with all-MiniLM-L6-v2 (22M) — a $27\times$ reduction — costs $0.001 C_V$ on 20News and 0.007 on R8, i.e. a 22M encoder recovers nearly all of the gain. On the scoring side (Table E3), the spread between bert-base-uncased (110M), roberta-base (125M) and deberta-v3-base (184M) is at most $0.006 C_V$, whereas removing semantic scoring altogether costs $0.066 C_V$ on 20News and 0.123 on R8 (Table 2). The variation attributable to encoder capacity is thus an order of magnitude smaller than the effect of the mechanism itself.

On a size-matched generative control. The cleanest control would be LLM-ITL instantiated with a generator of the same size as our encoders. This is not straightforward, because LLM-ITL requires a model that can follow an instruction and emit a list of topic words: an encoder such as bert-base-uncased cannot play that role, so “identical backbones” is not a well-defined configuration across the two methods. A comparison against LLM-ITL driven by a small instruction-tuned decoder would be a legitimate and informative control, and we leave it to future work; we note only that the evidence above already rules out capacity as the explanation for our gains, and that BERT-NTM improves C_V over the unrefined backbone in all 36 backbone–dataset combinations we report, including the weakest backbone (NVDM), which is not the behaviour of a method whose advantage comes from model scale.

G.19 Two-Stage Refinement and the Matched Training Budget

Why refinement is a second phase. The semantic signal is *conditional on the current topic*: the template in (3.1) is filled with the current top-ranked words of β_k , and the masked language model scores a candidate against that context. Early

in training, β_k is close to uninformative, so the context words are arbitrary and the resulting compatibility scores carry no topical information — the auxiliary loss would then inject noise into β at precisely the point where the corpus signal is weakest. Refinement therefore starts only once the topic contexts are stable enough to be worth conditioning on. This is the same reason the candidate set is refreshed periodically rather than fixed once (Appendix G.15): the semantic signal is recomputed as the context it depends on improves. A fully joint objective from epoch zero would require the semantic term to be reliable before there is anything for it to be reliable about.

The budget is matched. Because refinement is a second phase, it is important to be explicit that it does not buy extra optimization. The total number of epochs T is fixed per (backbone, dataset) pair and is independent of whether refinement is enabled; the switch to refinement happens at epoch $T - 50$, so the base phase is shortened by exactly as many epochs as refinement adds. The unrefined and refined rows of Table 1 thus consume identical epoch budgets, and the coherence gains cannot be explained by longer training. Under this schedule the refined model is in fact trained for *fewer* epochs on the base objective alone than its unrefined counterpart, which makes the comparison conservative rather than favourable to us.

Division of labour between the two terms. The component ablation in Table 2 shows that the two refinement terms are not interchangeable. Removing semantic scoring costs 0.066 C_V on 20News and 0.123 on R8, while its effect on PN is much smaller ($0.410 \rightarrow 0.408$ on 20News, $0.727 \rightarrow 0.693$ on R8). Removing document–topic alignment (Table E12) has the opposite profile: a smaller coherence cost (0.043 on 20News, 0.020 on R8) but the largest PN drop of any ablation on R8 ($0.727 \rightarrow 0.667$). The decoder-side term is thus primarily responsible for topic coherence and the encoder-side term for document–topic assignment, which is what the two objectives were designed to do.

H Evaluation Protocol and Discussion of Metrics

Automated coherence scores depend strongly on implementation choices — the reference corpus used to estimate word probabilities, the window

size, the number of top words scored — so we document ours in full, and then discuss what these metrics do and do not license us to conclude.

H.1 Computation of C_V and NPMI

We compute C_V with Palmetto 0.1.5 (Röder et al., 2015), invoked as an external executable rather than reimplemented. Word probabilities are estimated from the English Wikipedia reference index distributed with Palmetto (the `wikipedia_bd` index); no reference statistics are taken from the evaluation corpora themselves, so coherence is measured against an external corpus and cannot be inflated by overfitting to 20News, R8, DBpedia or AGNews. Each topic is represented by its **top-10** words ranked by β_k , one topic per line, and the reported figure is the arithmetic mean of the per-topic scores returned by Palmetto. The remaining components of C_V are those fixed by its definition (Röder et al., 2015): one-set segmentation of the top words, a boolean sliding window of size 110 for probability estimation, NPMI as the confirmation measure, and cosine similarity between each top-word NPMI vector and the sum of all such vectors, aggregated by arithmetic mean. All methods in Table 1 that we run ourselves are scored with exactly this invocation, and the LDA and TopicGPT rows cited from Yang et al. (Yang et al., 2025b) were produced on the same preprocessed corpora (Appendix D).

NPMI is reported in Appendix G.10 under the same top-10 convention. Because C_V and NPMI use different definitions, they are not expected to rank models identically, and we do not treat disagreement between them as an error in either.

H.2 What Automated Coherence Metrics Do and Do Not Show

We report C_V and NPMI because they are the standard basis of comparison in this literature and because the baseline numbers we compare against were produced with them. We do not, however, treat them as a direct measurement of interpretability, and the evidence for that caution is by now substantial. Doogan and Buntine (Doogan and Buntine, 2021) show that automated coherence rankings can diverge from expert judgement and are sensitive to the reference corpus and parameterization, to the point where they can mislead model selection. Hoyle et al. (Hoyle et al., 2021) report that automated coherence and human evaluation disagree on which of two models produces better topics, and that validation of these metrics against

human judgement has been much weaker than their widespread use implies. Both findings apply to our numbers as much as to anyone else’s: a C_V improvement is evidence that the top words of a topic co-occur more tightly in Wikipedia, and it is an inference — not an observation — that a human reader would find the refined topic more interpretable.

This is a real limitation on the strength of our claims, and we make three choices to keep it in view. First, we never rest a conclusion on C_V alone: we report NPMI (Appendix G.10), topic diversity (Appendix G.8), Purity and NMI separately (Appendix G.9), and LLM-based coherence (Appendix G.1), which measure different and partly independent properties. Second, we include qualitative topic pairs (Appendix G.5) so that a reader can inspect what refinement actually changed rather than trusting the aggregate. Third, we state where the evidence is weakest: none of these signals is a human judgement, and a small controlled word-intrusion study on one dataset would test the interpretability claim far more directly than any of them. We did not run such a study for this paper, and we identify it as the most valuable single addition to this evaluation in the Limitations section.

Two consequences for how our results should be read. Improvements that are consistent across metrics with different failure modes — as ours are, since C_V , NPMI, LLM-based coherence and PN move in the same direction across the 36 backbone–dataset settings — are more credible than an improvement visible only in C_V . Conversely, small differences *within* a single metric, such as the 0.003 C_V gap between Top- $N=25$ and $N=30$ in Table E13, should not be over-interpreted; we use them only to argue that behaviour is stable, never to claim one configuration is better.

H.3 LLM-Based Coherence Evaluation

The LLM-based scores in Appendix G.1 follow the protocol of Stambach et al. (Stambach et al., 2023): the evaluator is shown the top words of a single topic together with a one-sentence description of the source corpus, and returns an integer coherence rating on a 1–3 scale. The full prompts are given verbatim in Appendix E, one per dataset, and every model’s topics are scored with the same prompt and the same evaluator so the comparison is within-evaluator.

Two caveats apply regardless. An LLM judge is not a human judge: it is a model-based relat-

edness estimate, which is why we present it as a complement to C_V rather than as a resolution of the concerns above. And if the evaluator shares pretraining data or architecture family with the encoders used for refinement, some agreement is expected *a priori*; the LLM-based results should be read as corroborating the automated metrics, not as independent confirmation.

H.4 On Reporting PN as the Mean of Purity and NMI

Following the protocol of Yang et al. (Yang et al., 2025b), we summarise document–topic alignment as PN, the arithmetic mean of Purity and NMI. We keep this aggregate for a single reason — several rows of Table 1 are quoted from their paper, and changing the summary statistic would make those rows incomparable — but the aggregate has a genuine weakness that we do not want to paper over.

Purity and NMI are not commensurable quantities. Both lie in $[0, 1]$ and both are higher-is-better, which is what makes averaging them look reasonable, but Purity is a cluster-assignment accuracy that increases mechanically with the number of clusters, whereas NMI is an information-theoretic quantity normalized against the entropies of both partitions and does not share that bias. Their mean therefore has no interpretation of its own, and it can conceal a trade-off in which one component improves while the other degrades. A model that gains Purity by fragmenting topics can be rewarded by PN even though its NMI falls.

We mitigate this rather than defend it. Purity and NMI are reported separately for every dataset and every value of K in Appendix G.9 (Tables E8 and E10), so no information is lost by the aggregation, and any reader who prefers one component can read it directly. Where we make claims about document–topic alignment we have checked them against the disaggregated tables. Readers comparing against future work should prefer the separate Purity and NMI numbers; PN is retained for continuity with the LLM-ITL protocol, not because we regard it as the better statistic.

H.5 The LDA Baseline

The LDA row of Table 1 is quoted from Yang et al. (Yang et al., 2025b) and was not re-run by us, as stated in Appendix D; this is possible because we use their released preprocessed corpora and the same evaluation setting. Two clarifications are warranted. First, the row should be read as a reference

point for the reader’s orientation and not as part of a controlled comparison: the controlled comparison in this paper is always between a base neural topic model and the same model after refinement, where every other factor is held fixed. Second, the choice of LDA estimator matters more than it may appear. LDA estimated by collapsed Gibbs sampling — as implemented in MALLET (McCallum, 2002) — has repeatedly been found competitive with, or superior to, both variational LDA and several neural topic models on automated coherence, and a well-tuned Gibbs LDA would very likely score higher than the row we quote (Doogan and Buntine, 2021). We therefore do not claim any advantage over LDA as such.

H.6 Relation to Prior Work Not Included as a Baseline

Three lines of work are closely enough related that their absence from Table 1 deserves an explicit account.

Knowledge distillation from BERT into a topic model. Hoyle et al. (Hoyle et al., 2020) also improve a neural topic model using a BERT-derived signal, and their method is likewise modular with respect to the backbone. The mechanism differs in what is transferred and when. They train a student topic model against a teacher distribution derived from a pretrained language model, so the pretrained signal shapes the model’s reconstruction target throughout training, and obtaining a refined model requires training it from scratch under the teacher. BERT-NTM leaves the base model’s training untouched and acts on an *already trained* model, using a topic-conditioned score defined over each topic’s current candidate words rather than a document-level target distribution; this is what allows the same procedure to be applied unchanged to nine different backbones and to be compared directly against a generative refiner on the same checkpoints. The two are nevertheless complementary in principle — a distilled model could itself be refined afterwards — and a direct empirical comparison would be informative. We did not run it, and we note it as the most relevant missing baseline.

Contextualized topic models. The contextualized topic model family (Bianchi et al., 2021a,b) feeds pretrained sentence embeddings into the inference network, improving topic coherence through better document representations. This is a change

to the model architecture rather than a refinement procedure: the pretrained signal enters through the encoder input and the topic–word distributions are never scored against a language model. It is therefore a different kind of contribution and sits alongside the backbones we refine rather than competing with our refinement step; in principle it is a candidate backbone for BERT-NTM. Our own encoder-side term is closer in spirit to this line than our decoder-side term is, and the comparison in Table E12 between point-wise and distributional alignment is the relevant point of contact.

MMD in topic modeling. Using MMD inside a topic model is not new, and we want to be explicit about the precedent. W-LDA (Nan et al., 2019) trains a Wasserstein autoencoder for topic modeling in which an MMD penalty matches the *aggregated posterior* of the document–topic proportions to a Dirichlet prior over the simplex. Our use of MMD has a different target and a different space: we match a per-document Dirac measure at the document’s embedding against the topic-mixture measure $\sum_k \theta_{d,k} \delta(\mathbf{z}_k)$ induced by that same document’s proportions, in the representation space of a pretrained encoder. The penalty is per-document rather than aggregate, and the reference is a semantic quantity derived from the current topics rather than a fixed prior. What we claim as novel is this construction, not the use of MMD.