LLM-XTM: Enhancing Cross-Lingual Topic Models with Large Language Models

Minh Chu Xuan[1*], Tien-Phat Nguyen[1*], Dinh Viet Sang [1], Linh Ngo Van[1,†], Diep Thi-Ngoc Nguyen[2], Trung Le[3]

[1]Hanoi University of Science and Technology,
[2]VNU University of Engineering and Technology, [3]Monash University

Abstract

Cross-lingual topic modeling aims to discover shared semantic structures across languages, yet existing models depend on sparse bilingual resources and often yield incoherent or weakly aligned topics. Recent LLM-based refinements improve interpretability but are costly, document-level, and prone to hallucination, with prior white-box approaches requiring inaccessible token probabilities. We propose LLM-XTM, a framework that integrates LLM-guided topic refinement with self-consistency uncertainty quantification, enabling black-box, stable, and scalable enhancement of cross-lingual topic models. Experiments on multilingual corpora show that LLM-XTM achieves superior topic coherence and alignment while reducing reliance on bilingual dictionaries and expensive LLM calls. The code is publicly available at `https://github.com/tienphat140205/LLM-XTM`.

1 Introduction

Discovering latent themes within large text collections has long been a central problem in computational linguistics. Topic modeling (TM) provides the core framework for uncovering such hidden structures in an unsupervised manner (Hofmann, 1999; Blei et al., 2003). When extended to multilingual corpora, this paradigm becomes Cross-Lingual Topic Modeling (CLTM), which seeks to map semantically equivalent documents from different languages to comparable topic distributions (Ni et al., 2009; Mimno et al., 2009; Yuan et al., 2018; Wu et al., 2020, 2023a; Nguyen et al., 2025b; Phat et al., 2026). In addition to aligning topic proportions, CLTM must ensure that paired topics across languages convey consistent meanings, preserving interpretability and comparability across linguistic contexts. Consequently, CLTM serves as an essential framework for analyzing multilingual data and supporting cross-cultural semantic understanding.

En Topic#1:	rating	gauge	height	mile	shoe
ZH Topic#1:	投资者	财经	基金	股市	大盘
Translation#1:	investor	finance	fund	stock mkt.	index
En Topic#2:	buy	nice	awesome	excellent	seller
ZH Topic#2:	好看	用品	说明书	精品	保养
Translation#2:	good-looking	product	manual	premium	maintenance
En Topic#3:	software	windows	desktop	tech	computer
ZH Topic#3:	体系	答	治疗	转换	整理
Translation#3:	system	answer	treatment	convert	organize
En Topic#4:	news	hype	hint	reviews	media
ZH Topic#4:	炒作	皇帝	帅	想象	效应
Translation#4:	hype	emperor	handsome	imagination	effect

Table 1: Example of misaligned topics identified from the analysis. The words grouped under each topic differ in semantic coherence between the two languages by InfoCTM (Wu et al., 2023a).

Despite notable progress, most cross-lingual topic models still depend heavily on external bilingual resources—such as document embeddings, seed dictionaries, or parallel corpora—to establish alignment across languages (Mimno et al., 2009; Wu et al., 2023a, 2020; Chang et al., 2024; Nguyen et al., 2025b). While effective to some extent, these resources are inherently limited in coverage and often capture only surface-level or corpus-specific correspondences, providing shallow semantic signals that may not suffice to ensure deeper thematic consistency across languages (Vulic and Moens, 2015; Yuan et al., 2018). In practice, the situation is further complicated by the quality of the corpora themselves: parallel or comparable datasets often contain mistranslations, domain mismatches, and lexical ambiguities, introducing substantial noise into the alignment process (Mimno et al., 2009; Ruder et al., 2019; Artetxe and Schwenk, 2019). Such imperfections make it difficult for models to discover truly coherent and semantically equivalent topics across languages. As a result, topics that are nominally aligned may diverge in meaning, and documents that express the same underlying content can still yield inconsistent topic distributions, as illustrated in Table 1. Meanwhile, recent

*Equal contribution
†Corresponding author: linhnv@soict.hust.edu.vn

Proceedings of the 64th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), pages 3719–3737
July 2-7, 2026 ©2026 Association for Computational Linguistics

progress in large language models (LLMs) offers a promising remedy. Pretrained on massive multilingual corpora, LLMs internalize deep semantic correspondences that extend well beyond lexical overlap. Their ability to reason across languages makes them powerful semantic experts for refining topic-word distributions and strengthening cross-lingual alignment.

While recent advances have explored the use of Large Language Models (LLMs) to refine or even directly generate topics, these approaches also face important limitations. Most existing methods treat LLM outputs as ground truth, prompting the model to produce topic words or document-level labels independently for each document (Rijcken et al., 2023; Wang et al., 2023; Pham et al., 2024a; Mu et al., 2024; Doi et al., 2024). This design overlooks the global structure of the corpus, often leading to incomplete topic coverage and misalignment across documents. Moreover, running LLM inference for every document in large multilingual corpora is computationally prohibitive, making such approaches impractical at scale. Even when applied selectively, LLMs are prone to hallucination, producing irrelevant or misleading topic words that undermine consistency (Ji et al., 2023). Attempts to address these issues, such as LLM-in-the-Loop (LLM-ITL) (Yang et al., 2025b), integrate LLM refinement during training, but rely on white-box access to token probabilities to estimate confidence. This requirement is costly and infeasible for low-resource settings or closed-source models. Consequently, there remains a need for a more efficient, uncertainty-aware way to incorporate LLM refinements into cross-lingual topic modeling.

In this work, we propose LLM-XTM, a framework that integrates LLM refinement into cross-lingual topic modeling through a scalable, uncertainty-aware design. The first component focuses on refining topic-word distributions: candidate topic words produced by the base model are passed to an LLM for refinement, and multiple refinement paths are sampled. Inspired by SelfCheckGPT (Manakul et al., 2023), which detects hallucinations in LLM outputs via self-consistency across generations, we adapt this technique to quantify uncertainty in topic refinement. Only words with high agreement across refinement paths are retained, producing stable and trustworthy refinements. These are then aligned with the base topic-word distributions using a distributional loss based on Maximum Mean Discrepancy (MMD) (Gretton et al., 2012; Li et al., 2015),ensuring that LLM feedback improves interpretability while remaining consistent with corpus-driven signals. This component enhances topic-word coherence and robustness, laying the foundation for reliable cross-lingual alignment.

While LLM-guided topic-word refinement indirectly shapes document-topic distributions via the reconstruction loss, it does not explicitly enforce cross-lingual consistency at the document level. Due to lexical divergence, semantically equivalent documents in different languages may still yield diverging topic proportions θ_d, as their surface word distributions differ significantly. To address this, we introduce a document-level alignment loss inspired by question answering: each document is encoded into a multilingual semantic space (e.g., using BGE-M3 (Chen et al., 2024)), and matched against topic embeddings derived from LLM-refined bilingual word sets, treating documents as "questions" and topics as "answers." The resulting similarity-based topic distribution $\hat{\theta}_d$ reflects semantic relevance, and we minimize the KL-divergence $\mathrm{KL}(\theta_d \parallel \hat{\theta}_d)$ to align latent topic proportions with semantic intent. This encourages semantically similar documents—regardless of language—to share similar topic distributions, thereby enforcing robust cross-lingual alignment at the document level. Our contributions are threefold:

- We introduce **LLM-XTM**, the first framework that enhances cross-lingual topic modeling with LLM-guided refinement while remaining efficient and scalable.

- We adapt **self-consistency uncertainty quantification** to the topic refinement setting, enabling hallucination-resistant integration of LLM feedback, and align refined topic-word distributions with corpus-driven signals via an **MMD loss**.

- We propose a novel document-level alignment mechanism: refined topics are embedded into a multilingual space and matched with document embeddings, with a **KL-divergence objective** ensuring consistent alignment between semantic similarity vectors and θ distributions.

- Our extensive empirical results and qualitative analyses on diverse multilingual corpora not only show that LLM-XTM surpasses strong baselines but also provide insights into the effectiveness and robustness of each proposed component.

2 Preliminaries

2.1 Notations

We model a multilingual corpus comprising D documents in two languages, denoted L_1 and L_2, with the objective of discovering K shared 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 belongs to either L_1 or L_2. The vocabulary for L_1 is V_1 of size $|V_1|$, and for L_2 is V_2 of size $|V_2|$. The BoW representation of document d is $x_d \in \mathbb{R}^{|V_\ell|}$, where $\ell \in \{1, 2\}$ indicates the document's language.

For each language L_ℓ where $\ell \in \{1, 2\}$, the topic-word distribution is denoted by $\beta^{(\ell)} \in \mathbb{R}^{|V_\ell| \times K} = (\beta_1^{(\ell)}, \ldots, \beta_K^{(\ell)})$, where each $\beta_k^{(\ell)} \in \mathbb{R}^{|V_\ell|}$ represents the word distribution for topic k over the vocabulary V_ℓ, satisfying $\sum_{v \in V_\ell} \beta_{v,k}^{(\ell)} = 1$. Each document x_d is associated with a topic proportion vector $\theta_d \in \mathbb{R}^K$ such that $\sum_{k=1}^{K} \theta_{d,k} = 1$.

2.2 VAE-based Topic Model

We adopt a Variational Autoencoder (VAE) backbone. The encoder transforms each document's Bag-of-Words vector x_d into parameters (μ, Σ) of the posterior $q(z|x_d) = \mathcal{N}(z|\mu, \Sigma)$, from which a latent variable z is sampled via reparameterization (Kingma and Welling, 2013) under the Gaussian prior $p(z) = \mathcal{N}(z|\mu_0, \Sigma_0)$. Topic proportions are computed as $\theta_d = \mathrm{softmax}(z)$, and the decoder reconstructs x_d from topic–word distributions $\beta \in \mathbb{R}^{V \times K}$ (optimized or embedding-based (Srivastava and Sutton, 2017; Dieng et al., 2020; Wu et al., 2023b)) using $\mathrm{Multinomial}(\mathrm{softmax}(\beta\theta_d))$. The training objective is:

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

2.3 Maximum Mean Discrepancy (MMD)

Let k be a positive–definite kernel with RKHS $\mathcal{H}_k$. The kernel mean embedding maps a distribution P to $\mu_k(P) = \mathbb{E}_{x \sim P}[k(x, \cdot)] \in \mathcal{H}_k$; with *characteristic* kernels this map is injective, enabling metric comparison of distributions in RKHS (Muandet et al., 2017; Gretton et al., 2012). The squared Maximum Mean Discrepancy (MMD) between P and Q is

$$\begin{aligned} \mathrm{MMD}_k^2(P, Q) = \mathbb{E}_{x,x' \sim P}\, k(x, x') \\ + \mathbb{E}_{y,y' \sim Q}\, k(y, y') \\ - 2\, \mathbb{E}_{x \sim P,\, y \sim Q}\, k(x, y). \end{aligned} \quad (1)$$

which equals $\|\mu_k(P) - \mu_k(Q)\|_{\mathcal{H}_k}^2$ and serves as a principled two-sample discrepancy and alignment objective (Gretton et al., 2012). For weighted empirical samples $\{(x_i, w_i)\}$ and $\{(y_j, u_j)\}$ (weights sum to 1), the quadratic estimator $\sum_{i,i'} w_i w_{i'} k(x_i, x_{i'}) + \sum_{j,j'} u_j u_{j'} k(y_j, y_{j'}) - 2 \sum_{i,j} w_i u_j k(x_i, y_j)$ is differentiable in both sample locations and weights, fitting our topic–word distributions. In embedding spaces we adopt a Gaussian kernel on cosine-induced distances and choose bandwidth via the median heuristic or a small multi-kernel set for scale robustness (Garreau et al., 2017; Muandet et al., 2017). MMD is closely connected to energy distance through distance-induced kernels, lending additional statistical justification (Sejdinovic et al., 2013). We use this loss to align model and LLM-refined topic–word distributions (§3).

3 Methodology

Our proposed framework, LLM-XTM, is a two-phase enhancement applied to a pre-trained cross-lingual topic model. Phase 1 corresponds to the backbone neural topic model (e.g., VAE-based InfoCTM, NMTM, or XTRA), which produces initial topic-word and document-topic distributions. The second phase, which is the focus of this section, applies the LLM-XTM enhancement architecture, illustrated in Figure 1, to refine and align the converged model. This enhancement stage consists of two primary components: (1) using a Large Language Model (LLM) to refine the topic-word distributions (β) for improved coherence, and (2) employing a novel question-answering (QA) inspired mechanism to align the document-topic distributions (θ) for semantic consistency. The following sections will detail each component of this enhancement phase.

3.1 Cross-Lingual Topic Word Refinement

For each topic k, the base neural topic model yields two language-specific top-word lists $w_k^{(\mathrm{en})} \subset V_{\mathrm{en}}$

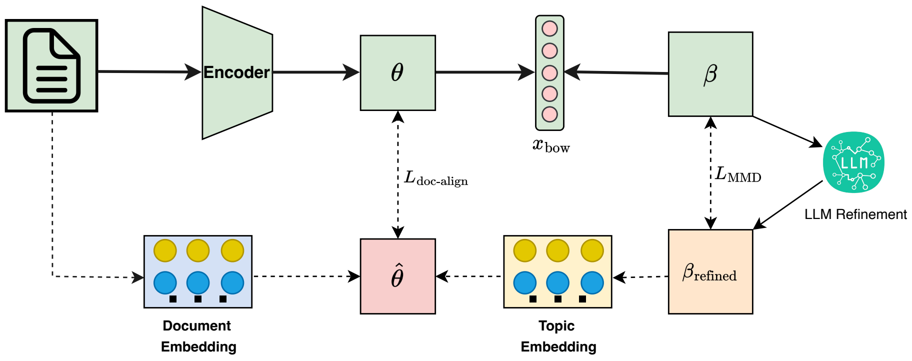


Figure 1: The LLM-XTM architecture enhances a VAE-based topic model using a dual-alignment strategy guided by a Large Language Model (LLM). The LLM first refines the model's topic–word sets, producing semantically cleaner top-word lists ($\beta \rightarrow \beta_{\text{refined}}$). To ensure consistency between the original and refined topic representations, an MMD loss (L_{mmd}) encourages the original top-word distributions to stay close to the refined ones, preserving the latent semantic structure while enhancing coherence. Finally, a document-level alignment loss ($L_{\text{doc-align}}$) matches the model's inferred document–topic distributions (θ) with the LLM-induced semantic targets ($\hat{\theta}$), enforcing cross-lingual consistency across documents and topics.

and $w_k^{(\text{zh})} \subset V_{\text{zh}}$ (top-15 each). We concatenate them into a bilingual candidate pool:

$$C_k = w_k^{(\text{en})} \cup w_k^{(\text{zh})}.$$

This pool is refined with a large language model (LLM)—for example, Gemini accessed via API, though any comparable LLM can be used. The LLM is instructed to remove noisy or irrelevant items, retain the most representative words capturing the shared semantic theme, and add missing but coherent terms in both languages. The output is a fixed-size bilingual set $\bar{w}_k$ that improves interpretability and provides a stable basis for subsequent cross-lingual alignment.

To control the frequency of LLM interactions, we introduce a refinement frequency parameter f. The refinement process is executed every f epochs rather than at each training iteration, allowing the model to alternate between internal optimization and periodic LLM feedback. A smaller f yields more frequent refinements and tighter LLM guidance, whereas a larger f reduces computational overhead but updates less often (see 4.6 for sensitivity analysis).

Implementation details and the exact prompt specification are provided in Appendix E.

3.2 Self-Consistent Refinement

While a single refinement pass with the LLM can improve topic word quality, the outputs remain inherently stochastic and may vary across runs. To enhance stability, we employ a self-consistent refinement strategy. For each topic k, the bilingual candidate pool C_k is submitted to the LLM R times, producing refined sets

$$\tilde{w}_k^{(1)}, \tilde{w}_k^{(2)}, \ldots, \tilde{w}_k^{(R)}.$$

We then aggregate word occurrences across refinement rounds to estimate empirical frequencies:

$$f_k(v) = \frac{1}{R} \sum_{r=1}^{R} \mathbf{1}\{v \in \tilde{w}_k^{(r)}\}, \quad v \in \bigcup_{r=1}^{R} \tilde{w}_k^{(r)}.$$

High-confidence words are selected as those with the largest $f_k(v)$, yielding the final refined bilingual set

$$\bar{w}_k = \text{Top}_M\big(\{(v, f_k(v))\}\big).$$

This self-consistency procedure filters out noisy candidates and preserves terms that consistently appear across multiple refinements, leading to a more reliable and interpretable bilingual representation of each topic. The number of refinement rounds R determines how many independent LLM refinements are aggregated, balancing stability and cost; its sensitivity is analyzed in Section 4.6.

3.3 MMD-based Refinement Loss

To align the model's topic-word distributions with the LLM-refined sets, a loss based on the Maximum Mean Discrepancy (MMD) is used (Gretton et al., 2012; Muandet et al., 2017). This approach requires creating two distinct distributions for each topic k: the model's raw distribution and the LLM's refined target distribution.

The model's raw distribution, $\beta_k^{(raw)}$, is constructed from the top-N words generated by the decoder. The probabilities for the top words in each language are combined into a language-balanced mixture, which is then normalized as follows:

$$\tilde{\beta}_k^{(raw)}(v) = \begin{cases} \beta_k^{(l1)}(v), & v \in W_k^{(l1)} \\ \beta_k^{(l2)}(v), & v \in W_k^{(l2)} \\ 0, & \text{otherwise} \end{cases}$$

$$\beta_k^{(raw)}(v) = \frac{\tilde{\beta}_k^{(raw)}(v)}{\sum_{u \in U_k} \tilde{\beta}_k^{(raw)}(u)}$$

where $W_k^{(l)}$ is the top-word list for language l, and U_k is the union of these lists.

The LLM-refined target distribution, $\beta_k^{(refined)}$, is derived from the self-consistent refinement process where word counts across multiple refinement rounds are tallied. These counts are similarly formed into a language-balanced mixture and normalized to create the target:

$$\tilde{c}_k(v) = \begin{cases} c_k^{(l1)}(v), & v \in \overline{W}_k^{(l1)} \\ c_k^{(l2)}(v), & v \in \overline{W}_k^{(l2)} \\ 0, & \text{otherwise} \end{cases}$$

$$\beta_k^{(refined)}(v) = \frac{\tilde{c}_k(v)}{\sum_{u \in \bar{U}_k} \tilde{c}_k(u)}$$

where $\overline{W}_k^{(l)}$ is the refined word set for language l, and $\overline{U}_k$ is the union of these sets.

The divergence between these two distributions is measured using the squared MMD, calculated with a Gaussian kernel on the cosine distances between word embeddings (Garreau et al., 2017; Sejdinovic et al., 2013). The final training objective is the average MMD loss over all K topics. This loss shifts the model's topic-word distributions toward the more coherent LLM-refined targets.

$$\mathcal{L}_{\text{MMD}} = \frac{1}{K} \sum_{k=1}^{K} \text{MMD}^2(\beta_k^{(\text{raw})}, \beta_k^{(\text{refined})})$$

3.4 Document–Topic Alignment via QA Mechanism

While refining topic words enhances coherence, a core challenge remains: ensuring that documents with the same meaning in different languages are assigned similar topic distributions. To solve this, we introduce a pioneering document-level alignment mechanism that reframes the problem through the intuitive lens of Question Answering (QA).

In this elegant paradigm, each document is treated as a "question" asking, "What is my core semantic theme?". The LLM-refined topics, in turn, act as a pool of high-quality "candidate answers". The goal is to teach the model how to match each question to its most relevant answer, irrespective of the document's language.

The mechanism works by first transforming each LLM-refined bilingual topic word list ($\overline{w}_k$) into a rich semantic vector using a powerful multilingual encoder like BGE-M3 (Chen et al., 2024). This creates a single, potent embedding for each topic ($t_k = \text{Enc}(\overline{w}_k)$), representing the core meaning of the "answer". In parallel, every document (x_d) is mapped into the same semantic space, producing a corresponding "question" vector ($h_d = \text{Enc}(x_d)$). From here, we calculate the cosine similarity between each document "question" (h_d) and all topic "answers" ($\{t_k\}_{k=1}^{K}$):

$$s_{d,k} = \frac{h_d^\top t_k}{\|h_d\|_2 \, \|t_k\|_2}, \qquad k = 1, \ldots, K$$

These similarity scores are normalized via a softmax to obtain a new topic distribution that purely reflects semantic relevance:

$$\hat{\theta}_{d,k} = \frac{\exp(s_{d,k}/\tau)}{\sum_{j=1}^{K} \exp(s_{d,j}/\tau)}.$$

Finally, we align the model's internal topic posterior (θ_d) with this powerful external semantic signal by minimizing the KL-divergence between the two distributions:

$$L_{\text{doc-align}} = \sum_{d=1}^{D} KL(\theta_d || \hat{\theta}_d).$$

This objective forces the model to learn cross-lingually consistent document representations, ensuring that semantic meaning, not just vocabulary, drives the topic assignments.

3.5 Overall Objective

Our framework operates in two phases. **Phase 1** trains a base neural topic model (e.g., NMTM, InfoCTM), optimized with its original objective, which may consist of $\mathcal{L}_{\text{TM}}$ and additional backbone-specific losses (e.g., contrastive or clustering terms). **Phase 2** then enhances the converged model by incorporating our refinement and alignment components. The composite objective in

Model	EC News			Amazon Review			Rakuten Amazon		
	CNPMI	TU	TQ	CNPMI	TU	TQ	CNPMI	TU	TQ
MCTA	0.025	0.489	0.012	0.028	0.319	0.009	0.021	0.272	0.006
MTAnchor	−0.013	0.192	0.000	0.028	0.323	0.009	−0.001	0.214	0.000
u-SVD	0.082	0.830	0.068	0.055	0.634	0.035	0.027	0.571	0.015
SVD-LR	0.083	0.820	0.068	0.053	0.627	0.033	0.026	0.558	0.015
XTRA	0.078	0.978	0.076	0.053	0.979	0.052	0.034	0.966	0.033
+ LLM-XTM	0.088	0.954	0.084	0.072	0.959	0.069	0.037	0.945	0.035
	↑12.8%	↓2.5%	↑10.5%	↑35.8%	↓2.0%	↑32.7%	↑8.8%	↓2.2%	↑6.1%
InfoCTM	0.041	0.943	0.039	0.037	0.930	0.034	0.032	0.870	0.028
+ LLM-XTM	0.062	0.898	0.056	0.050	0.933	0.047	0.040	0.870	0.035
	↑51.2%	↓4.8%	↑43.6%	↑35.1%	↑0.3%	↑38.2%	↑25.0%	↑0.0%	↑25.0%
NMTM	0.034	0.818	0.028	0.043	0.610	0.026	0.012	0.633	0.008
+ LLM-XTM	0.039	0.821	0.032	0.056	0.627	0.035	0.016	0.666	0.011
	↑15.9%	↑0.4%	↑14.3%	↑30.2%	↑2.8%	↑34.6%	↑33.3%	↑5.2%	↑37.5%

Table 2: Topic coherence (CNPMI), topic uniqueness (TU), and topic quality (TQ) where TQ = max(0, CNPMI) × TU. Results for MCTA and MTAnchor are taken directly from (Wu et al., 2023a), while all other baselines and our model are tuned and averaged over five random seeds, with the mean values reported.

Phase 2 over encoder parameters ϕ and decoder parameters ψ is:

$$\mathcal{J}(\phi,\psi) = \mathcal{L}_{\text{Phase 1}} + \lambda_{\text{mmd}}\,\mathcal{L}_{\text{MMD}} \\ + \lambda_{\text{qa}}\,\mathcal{L}_{\text{doc-align}}.$$

where $\mathcal{L}_{\text{Phase1}}$ denotes the loss function used by the chosen backbone model in Phase 1, while $\mathcal{L}_{\text{MMD}}$ (Section 3.3) and $\mathcal{L}_{\text{doc-align}}$ (Section 3.4) are the additional enhancement losses applied in Phase 2.

4 Experiments and Results

Datasets

We evaluated on three public benchmarks: **EC News** (Wu et al., 2020), a bilingual English–Chinese news corpus spanning six domains; **Amazon Review** (Yuan et al., 2018), English–Chinese product reviews converted to a binary task (five-star = 1, others = 0); and **Rakuten Amazon** (Yuan et al., 2018), Japanese–English reviews similarly framed as binary classification by rating. Detailed dataset statistics are provided in C

Baseline Models

We compared our method with a range of representative cross-lingual topic modeling baselines. **MCTA** (Shi et al., 2016) is a probabilistic CLTM framework that models cultural variation across languages. **MTAnchor** (Yuan et al., 2018) aligns topics through multilingual anchor words. **NMTM** (Wu et al., 2020) introduces a neural approach that embeds multilingual topics in a shared latent space. **InfoCTM** (Wu et al., 2023a) enhances topic representations using mutual information maximization to reduce redundancy. **XTRA** (Nguyen et al., 2025b) extends this line by applying contrastive learning jointly on document–topic (θ) and topic–word (β) spaces to achieve dual alignment. Finally, we include two clustering-based refinement methods—**u-SVD** and **SVD-LR** (Chang et al., 2024)—as modern baselines for topic post-processing and coherence enhancement.

Evaluation Metrics

We assess topic quality and utility through both intrinsic and extrinsic evaluations. Intrinsically, **CNPMI** (Hao et al., 2018) measures cross-lingual coherence, **TU** (Nan et al., 2019) quantifies topic diversity using the top 15 words, and **TQ** (Chang et al., 2024) combines both with negative coherence clipped to zero. We further include **LLM-based ratings** (Stammbach et al., 2023) (1–3 scale) to evaluate intra-lingual coherence and cross-lingual alignment.

4.1 Topic Quality Analysis

Table 2 presents intrinsic metrics for topic coherence (CNPMI), diversity (TU), and composite topic quality (TQ) across three benchmark datasets. Integrating **LLM-XTM** with **XTRA** consistently improves topic quality: on EC News, CNPMI increases by 12.8% and TQ by 10.5%; on Amazon Review, coherence improves by 35.8% and TQ

Model	Topic Quality		Classification			
	CNPMI	TU	EN-I	JA-I	EN-C	JA-C
NMTM†	0.012	0.633	0.796	0.826	0.610	0.681
NMTM + LLM-XTM†	0.016	0.666	0.792	0.833	0.621	0.728
NMTM + LLM-XTM (w/o $L_{\text{doc-align}}$)	0.012	0.679	0.791	0.832	0.611	0.723
NMTM + LLM-XTM (w/o L_{MMD})	0.012	0.641	0.795	0.833	0.621	0.723
NMTM + LLM-XTM (w/o self-consistency)	0.011	0.654	0.792	0.832	0.619	0.720
NMTM + LLM-XTM (w/ L_{OT})	0.013	0.664	0.795	0.829	0.620	0.720

Table 3: Ablation study on Rakuten Amazon (50 topics). We report Topic Quality (CNPMI, TU) and Classification Accuracy. Baselines (†) are reported from (Wu et al., 2023a). Ablations are applied to the LLM-XTM framework on the *NMTM* backbone.

by 32.7%. The slight reduction in TU ($\approx$ 2%) suggests that redundant or peripheral words are pruned, yielding more semantically focused topics. On Rakuten Amazon, CNPMI rises by 8.8% and TQ by 6.1%, indicating that LLM-based refinement enhances cross-lingual consistency while maintaining reasonable topic diversity.

Across other backbones, **LLM-XTM** shows even larger impacts. For **InfoCTM**, CNPMI jumps 51.2 % on EC News and 35.1 % on Amazon Review, with corresponding TQ gains of 43.6 % and 38.2 %; TU decreases are small (under 5 %). On the lighter **NMTM** model, LLM-XTM consistently boosts coherence and quality, with CNPMI gains in the range 15.9–30.2 % and TQ improvements up to 37.5 %. These findings confirm that LLM-XTM acts as a general enhancement layer, reliably increasing topic coherence and cross-lingual interpretability while exerting only minimal negative impact on topic diversity.

On the slight TU reduction. Note that TU (topic uniqueness) is usually defined as the average reciprocal of the number of occurrences of each top word across all topics (i.e. if a word appears in multiple topics, its contribution to TU is reduced). Hence, even small overlaps in high-importance words across topics decrease TU, regardless of whether those overlaps are semantically justified. In our setting, LLM-XTM favors core semantic terms and removes noisy or weakly related terms, so multiple topics may retain some shared "meaningful" words. This overlap leads to a modest TU drop despite improvements in coherence. The TU decline thus should be interpreted not purely as loss of diversity, but as a side-effect of focusing topics toward more semantically consistent vocabularies.

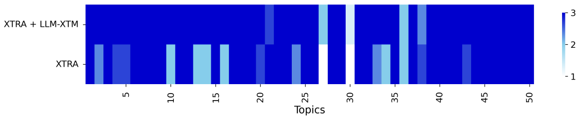
4.2 Evaluation of Topic Distributions via Classification

4.3 Ablation Study

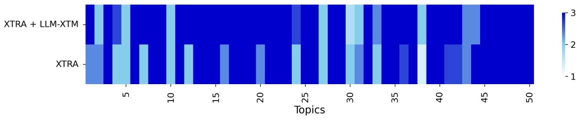
Table 3 reports an ablation on Rakuten Amazon with 50 topics to analyze each component's effect. The full **LLM-XTM** improves topic quality over NMTM†, reaching CNPMI 0.016 and TU 0.666 (vs. 0.012/0.633), and also enhances classification performance (EN-C 0.621, JA-C 0.728). Removing $L_{\text{doc-align}}$ notably weakens cross-lingual alignment, reducing CNPMI to 0.012 and EN-C to 0.611. Excluding L_{MMD} impairs topic–word coherence, dropping TU to 0.641 and JA-C to 0.723, while omitting self-consistency further degrades both CNPMI (0.011) and TU (0.654) with minor accuracy losses. Finally, comparing discrepancy measures, MMD surpasses Optimal Transport (CNPMI 0.016 vs. 0.013), as its kernel-based design more effectively captures semantic relations in shared multilingual embedding space, making it better suited for cross-lingual topic modeling.

4.4 LLM-based Topic Quality Evaluation

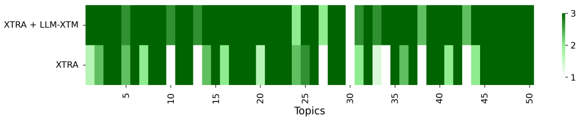
We conducted an automated evaluation using a Large Language Model (LLM) to quantitatively assess the impact of **LLM-XTM** on the strong **XTRA** baseline, following recent topic-model evaluation practices (Stammbach et al., 2023). As shown in Figure 2, the LLM rated intra-lingual coherence and cross-lingual alignment on a 1–3 scale. Results show consistent improvements in both aspects, with notably stronger cross-lingual alignment where the baseline was unstable. Overall, **LLM-XTM** markedly enhances semantic consistency and alignment across languages, confirming its effectiveness as a refinement framework. The full evaluation script is provided in Appendix E, and additional results are detailed in Appendix J.



(a) English intra-lingual semantic similarity



(b) Chinese intra-lingual semantic similarity



(c) Cross-lingual semantic similarity on Amazon Review dataset.

Figure 2: LLM-based evaluations of inner and cross-lingual semantic similarity on dataset Amazon Review. Darker shades indicate higher scores (on a 1–3 scale). Reported values are rounded from the mean scores over four independent LLM runs.

Topic: Music & Singing (NMTM — Amazon Review, Topic 26)					
NMTM, Topic 26 (Poor)					
EN:	song	albums	lyrics	riffs	catchy
ZH:	思念	挚爱	誓言	嫁	信件
Translations:	longing	true love	vow	marry	letter
NMTM + LLM-XTM, Topic 26 (Good)					
EN:	vocals	singer	lyrics	album	drums
ZH:	音乐	专辑	歌手	歌迷	浪漫
Translations:	music	album	singer	fan	romantic
Topic: Fashion & Magazines (InfoCTM — ECNews, Topic 45)					
InfoCTM, Topic 45 (Poor)					
EN:	cover	magazine	campaign	bundchen	fail
ZH:	外壳	上网	采访	付费	封面
Translations:	shell/casing	go online	interview	pay	cover
InfoCTM + LLM-XTM, Topic 45 (Good)					
EN:	vogue	cover	supermodel	magazine	shoot
ZH:	封面	模特	摄影	杂志	时尚
Translations:	cover	model	photography	magazine	fashion
Topic: Personal Care / Cosmetics (XTRA — Rakuten Amazon, Topic 40)					
XTRA, Topic 40 (Poor)					
EN:	makeup	festival	hair	lipstick	camera
JA:	美容	旅行	皮肤	化	保养
Translations:	beauty makeup	travel	skin	cosmetics	skincare
XTRA + LLM-XTM, Topic 40 (Good)					
EN:	skin	shampoo	fragrance	sensitive	lotion
JA:	化	保湿	洗	皮	敏感
Translations:	makeup	moisturizing	face wash	scalp	sensitive

Table 4: Cross-lingual topic examples. Upper blocks show noisy/misaligned sets; lower blocks show coherent sets. Red marks tokens that are semantically off-topic or misaligned.

4.5 Qualitative Analysis: Discovered Topic Word Examples

Table 4 illustrates cross-lingual topic examples in *Music & Singing*, *Fashion & Magazines*, and *Personal Care / Cosmetics*, comparing baseline models with their LLM-enhanced versions (NMTM + LLM-XTM, InfoCTM + LLM-XTM, and XTRA + LLM-XTM). Upper blocks show noisy or misaligned keywords, while lower blocks present refined, coherent sets. Baseline models often include off-topic terms such as "嫁" (marry), "上网" (go online), or "旅行" (travel), which blur topical focus and weaken bilingual alignment. In contrast, LLM-XTM removes noise and strengthens cross-lingual coherence—for instance, in *Cosmetics*, (*skin*, *fragrance*) ↔ (保湿, 洗) form clean, semantically aligned pairs—demonstrating clear improvements in both coherence and semantic consistency across languages.

4.6 Sensitivity Analysis of Refinement Parameters

We study two hyperparameters of our LLM-based refinement: refinement rounds (R) and refinement frequency (f, epochs between LLM calls). Experiments use the Amazon Review dataset with NMTM, holding other settings fixed. We test $R \in \{1, 3, 5, 7, 10, 13\}$ and $f \in \{5, 8, 10, 13\}$.

Refinement rounds (R). Aggregating R independent LLM calls improves CNPMI while TU shows mixed behavior. At $f = 8$, CNPMI increases from 0.0482→0.0562 (+16.6%) as R rises from 1 to 5, but TU rises only marginally from 0.600→0.627 (+4.2%). Beyond $R = 5$, gains diminish: from $R = 5$ to $R = 13$, CNPMI improves just 1.4% (0.0562→0.0570) while TU increases 1.0% (0.625→0.631). Since cost scales linearly with R, settings around R = 5–7 capture most benefits efficiently.

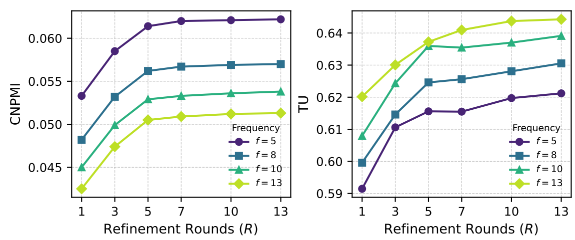


Figure 3: Sensitivity to rounds (R) and frequency (f) in CNPMI (left) and TU (right) on Amazon Review.

Refinement frequency (f). Lower f (more frequent calls) boosts CNPMI but reduces TU. At $R = 5$, CNPMI drops 17.8% (0.614→0.505) while TU increases 3.5% (0.616→0.637) when f increases from 5 to 13, revealing a clear trade-off between coherence and uniqueness.

5 Conclusion

We introduced **LLM-XTM**, a two-stage framework that applies a post-hoc enhancement to cross-lingual topic models: (i) we refine topic–word distributions with an LLM using a self-consistency–based filter and an MMD alignment loss, and (ii) we align document–topic posteriors to semantic targets via a QA-style objective. It improves cross-lingual coherence and transfer while preserving diversity and serves as a plug-in for diverse topic model backbones.

Limitation

The efficacy of LLM-XTM is fundamentally tethered to the quality of the underlying topic model it enhances. While our framework significantly improves coherence and alignment, its performance ceiling is constrained by the initial topics provided by the backbone model; it can refine, but not create, sensible topics from a completely failed initialization. Moreover, the reliance on multiple API calls to external LLMs for self-consistent refinement introduces considerations of computational latency, making the enhancement phase less suitable for real-time or highly resource-constrained environments.

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

This project was supported by the Air Force Office of Scientific Research under award number FA9550-23-S-0001.

References

Dimo Angelov. 2020. Top2vec: Distributed representations of topics. *CoRR*, abs/2008.09470.

Mikel Artetxe and Holger Schwenk. 2019. Massively multilingual sentence embeddings for zero-shot cross-lingual transfer and beyond. *Trans. Assoc. Comput. Linguistics*, 7:597–610.

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, Andrew Y Ng, and Michael I Jordan. 2003. Latent dirichlet allocation. *Journal of Machine Learning Research*, pages 993–1022.

Jordan L. Boyd-Graber and David M. Blei. 2012. Multilingual topic models for unaligned text. *CoRR*, abs/1205.2657.

Chia-Hsuan Chang, Tien-Yuan Huang, Yi-Hang Tsai, Chia-Ming Chang, and San-Yih Hwang. 2024. Refining dimensions for improving clustering-based cross-lingual topic models. *CoRR*, abs/2412.12433.

Chia-Hsuan Chang and San-Yih Hwang. 2021. A word embedding-based approach to cross-lingual topic modeling. *Knowl. Inf. Syst.*, 63(6):1529–1555.

Jianlv Chen, Shitao Xiao, Peitian Zhang, Kun Luo, Defu Lian, and Zheng Liu. 2024. BGE m3-embedding: Multi-lingual, multi-functionality, multi-granularity text embeddings through self-knowledge distillation. *CoRR*, abs/2402.03216.

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.

Damien Garreau, Wittawat Jitkrittum, and Motonobu Kanagawa. 2017. Large sample analysis of the median heuristic. *arXiv preprint arXiv:1707.07269*.

Aaron Grattafiori, Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, and 1 others. 2024. The llama 3 herd of models. *CoRR*, abs/2407.21783.

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. *Preprint*, arXiv:2203.05794.

Shudong Hao, Jordan L. Boyd-Graber, and Michael J. Paul. 2018. Lessons from the bible on modern topics: Low-resource multilingual topic model evaluation. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, NAACL-HLT 2018, New Orleans, Louisiana, USA, June 1-6, 2018, Volume 1 (Long Papers)*, pages 1090–1100. Association for Computational Linguistics.

Thomas Hofmann. 1999. Probabilistic latent semantic indexing. In *Proceedings of the 22nd annual international ACM SIGIR conference on Research and development in information retrieval*, pages 50–57.

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.

Jagadeesh Jagarlamudi and Hal Daumé III. 2010. Extracting multilingual topics from unaligned comparable corpora. In *Advances in Information Retrieval, 32nd European Conference on IR Research, ECIR 2010, Milton Keynes, UK, March 28-31, 2010. Proceedings*, volume 5993 of *Lecture Notes in Computer Science*, pages 444–456. Springer.

Ziwei Ji, Nayeon Lee, Rita Frieske, Tiezheng Yu, Dan Su, Yan Xu, Etsuko Ishii, Yejin Bang, Andrea Madotto, and Pascale Fung. 2023. Survey of hallucination in natural language generation. *ACM Comput. Surv.*, 55(12):248:1–248:38.

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*.

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.

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.

Potsawee Manakul, Adian Liusie, and Mark J. F. Gales. 2023. Selfcheckgpt: Zero-resource black-box hallucination detection for generative large language models. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing, EMNLP 2023, Singapore, December 6-10, 2023*, pages 9004–9017. Association for Computational Linguistics.

David M. Mimno, Hanna M. Wallach, Jason Naradowsky, David A. Smith, and Andrew McCallum. 2009. Polylingual topic models. In *Proceedings of the 2009 Conference on Empirical Methods in Natural Language Processing, EMNLP 2009, 6-7 August 2009, Singapore, A meeting of SIGDAT, a Special Interest Group of the ACL*, pages 880–889. ACL.

Mistral AI. 2025. Mistral small 3. `https://mistral.ai/news/mistral-small-3`. Accessed: April 2026.

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.

Aaron Mueller and Mark Dredze. 2021. Fine-tuning encoders for improved monolingual and zero-shot polylingual neural topic modeling. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, NAACL-HLT 2021, Online, June 6-11, 2021*, pages 3054–3068. Association for Computational Linguistics.

Feng Nan, Ran Ding, Ramesh Nallapati, and Bing Xiang. 2019. Topic modeling with wasserstein autoencoders. In *Proceedings of the 57th Conference of the Association for Computational Linguistics, ACL 2019, Florence, Italy, July 28- August 2, 2019, Volume 1: Long Papers*, pages 6345–6381. Association for Computational Linguistics.

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, Albuquerque, New Mexico, USA, April 29 - May 4, 2025*, 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.

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 Van Ngo, Duc Anh Nguyen, Sang Dinh, and Trung Le. 2025b. Xtra: Cross-lingual topic modeling with topic and representation alignments. *Preprint*, arXiv:2510.02788.

Truong Nguyen, Phi Van Dat, Ngan Nguyen, Linh Ngo Van, Trung Le, and Thanh Hong Nguyen. 2026b. CTPD: cross tokenizer preference distillation. In *Fortieth AAAI Conference on Artificial Intelligence, Thirty-Eighth Conference on Innovative Applications of Artificial Intelligence, Sixteenth Symposium on Educational Advances in Artificial Intelligence, AAAI*, pages 37783–37790. AAAI Press.

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, Albuquerque, New Mexico, USA, April 29 - May 4, 2025*, 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. *Artif. Intell.*, 353:104502.

Xiaochuan Ni, Jian-Tao Sun, Jian Hu, and Zheng Chen. 2009. Mining multilingual topics from wikipedia. In *Proceedings of the 18th International Conference on World Wide Web, WWW 2009, Madrid, Spain, April 20-24, 2009*, pages 1155–1156. ACM.

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*.

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, Thirty-Eighth Conference on Innovative Applications of Artificial Intelligence, Sixteenth Symposium on Educational Advances in Artificial Intelligence, AAAI*, pages 32710–32718. AAAI Press.

Qwen Team. 2025. Qwen3-Coder: Agentic coding in the world. `https://qwenlm.github.io/blog/qwen3-coder/`. Accessed: April 2026.

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.

Sebastian Ruder, Ivan Vulic, and Anders Søgaard. 2019. A survey of cross-lingual word embedding models. *J. Artif. Intell. Res.*, 65:569–631.

Dino Sejdinovic, Bharath K. Sriperumbudur, Arthur Gretton, and Kenji Fukumizu. 2013. Equivalence of distance-based and rkhs-based statistics in hypothesis testing. *Annals of Statistics*, 41(5):2263–2291.

Bei Shi, Wai Lam, Lidong Bing, and Yinqing Xu. 2016. Detecting common discussion topics across culture from news reader comments. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics, ACL 2016, August 7-12, 2016, Berlin, Germany, Volume 1: Long Papers*. The Association for Computer Linguistics.

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.

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.

Hai An Vu, Minh-Phuc Truong, Tu Vu, and Linh Ngo. 2026. Mol: Mixture of layers in cross-tokenizer embedding model distillation. *Knowledge-Based Systems*, 343:116001.

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.

Ivan Vulic and Marie-Francine Moens. 2015. Bilingual word embeddings from non-parallel document-aligned data applied to bilingual lexicon induction. In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing of the Asian Federation of Natural Language Processing, ACL 2015, July 26-31, 2015, Beijing, China, Volume 2: Short Papers*, pages 719–725. The Association for Computer 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. 2023. 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.

Xiaobao Wu, Xinshuai Dong, Thong Nguyen, Chaoqun Liu, Liangming Pan, and Anh Tuan Luu. 2023a. Infoctm: A mutual information maximization perspective of cross-lingual topic modeling. In *Thirty-Seventh AAAI Conference on Artificial Intelligence, AAAI 2023, Thirty-Fifth Conference on Innovative Applications of Artificial Intelligence, IAAI 2023, Thirteenth Symposium on Educational Advances in Artificial Intelligence, EAAI 2023, Washington, DC, USA, February 7-14, 2023*, pages 13763–13771. AAAI Press.

Xiaobao Wu, Xinshuai Dong, Thong Thanh Nguyen, and Anh Tuan Luu. 2023b. 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, Chunping Li, Yan Zhu, and Yishu Miao. 2020. Learning multilingual topics with neural variational inference. In *Natural Language Processing and Chinese Computing - 9th CCF International Conference, NLPCC 2020, Zhengzhou, China, October 14-18, 2020, Proceedings, Part I*, volume 12430 of *Lecture Notes in Computer Science*, pages 840–851. Springer.

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.

An Yang, Anfeng Li, Baosong Yang, and 1 others. 2025a. Qwen3 technical report. *CoRR*, abs/2505.09388.

Weiwei Yang, Jordan L. Boyd-Graber, and Philip Resnik. 2019. A multilingual topic model for learning weighted topic links across corpora with low comparability. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing, EMNLP-IJCNLP 2019, Hong Kong, China, November 3-7, 2019*, pages 1243–1248. Association for Computational Linguistics.

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.

Michelle Yuan, Benjamin Van Durme, and Jordan L. Ying. 2018. Multilingual anchoring: Interactive topic modeling and alignment across languages. In *Advances in Neural Information Processing Systems*

31: Annual Conference on Neural Information Processing Systems 2018, NeurIPS 2018, December 3-8, 2018, Montréal, Canada, pages 8667–8677.

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 Models and Cross-Lingual Extensions. Topic modeling seeks to identify hidden semantic patterns within large text collections. The foundational Latent Dirichlet Allocation (LDA) (Blei et al., 2003) conceptualizes documents as mixtures of latent topics, establishing the basis for later developments. Neural Topic Models (NTMs) build upon this idea through deep generative formulations such as variational autoencoder (VAE) frameworks, including NVDM and ProdLDA (Srivastava and Sutton, 2017), as well as embedding-oriented approaches like ETM (Dieng et al., 2020). Subsequent research integrates pre-trained language models (e.g., BERT) to better capture contextual and semantic nuances (Bianchi et al., 2021a,b; Hoyle et al., 2020). Other lines of work leverage optimal transport techniques (Zhao et al., 2021; Wu et al., 2023b) or contrastive learning objectives (Nguyen and Luu, 2021; Nguyen et al., 2024), with more recent approaches combining the two paradigms (Vuong et al., 2025; Khanh et al., 2026b), as well as mutual-information-based regularizers (Pham et al., 2024b), to improve topic coherence and representation separation. Complementary efforts target the training procedure itself, ranging from sharpness-aware minimization (Nguyen et al., 2025d) to multi-objective surrogate formulations (Khanh et al., 2026a; Le et al., 2025). In parallel, clustering-based approaches such as Top2Vec (Angelov, 2020) and BERTopic (Grootendorst, 2022) derive topics directly from dense embeddings, enhancing interpretability and flexibility across datasets, while a growing body of work tackles short-text, subdocument, or temporal settings where BoW signals are sparse or drift over time (Nguyen et al., 2025a; Vu et al., 2025; Nguyen et al., 2026c, 2025c, 2026a). Overall, research in this area has shifted from traditional count-based probabilistic models toward representation-driven frameworks that emphasize semantic alignment and contextual understanding.

Cross-lingual topic modeling (CLTM) extends topic modeling to multilingual settings to discover aligned themes across languages. Early work such as Mimno et al. (2009) relied on parallel corpora, limiting broader applicability. Later methods leveraged bilingual dictionaries to align vocabularies (Jagarlamudi and III, 2010; Boyd-Graber and Blei, 2012), with translation-based refinements proposed by Shi et al. (2016), Yuan et al. (2018), Yang et al. (2019), Wu et al. (2020), and Wu et al. (2023a). A more recent line, XTRA (Nguyen et al., 2025b), applies contrastive learning simultaneously on θ (document–topic distributions) and β (topic–word projections) to achieve dual semantic alignment; concurrently, GloCTM (Phat et al., 2026) pursues cross-lingual alignment through a shared global context space rather than relying on bilingual lexicons. Another strand employs multilingual embeddings (Chang and Hwang, 2021), but these often assume isomorphic embedding spaces that may not hold for distant languages. Transformer-based and zero-shot models (e.g., (Bianchi et al., 2021b; Mueller and Dredze, 2021)) reduce dependence on parallel data yet still struggle with consistent cross-lingual alignment. In parallel, refinement methods for clustering-based CLTM, such as u-SVD and SVD-LR (Chang et al., 2024), remove language-dependent dimensions from embeddings before clustering, improving alignment robustness.

Furthermore, as cross-lingual topic modeling increasingly relies on dense representations, embedding alignment and distillation have emerged as critical areas of research. In our QA-inspired alignment mechanism, we rely on powerful pre-trained multilingual encoders like BGE-M3 (Chen et al., 2024) to establish a shared semantic space. However, recent advances in embedding model distillation and cross-tokenizer preference alignment (Truong et al., 2025; Vu et al., 2026; Nguyen et al., 2026b) have demonstrated highly effective ways to compress and align these representation spaces. These techniques offer promising avenues for making cross-lingual topic alignment more computationally efficient and structurally robust in future iterations.

MMD and Distribution Alignment. Maximum Mean Discrepancy (MMD) formalizes distributional comparison by embedding probability measures into an RKHS and measuring distances between their kernel mean embeddings, a perspective that underlies modern two-sample and independence testing and reveals an equivalence to energy-distance statistics via negative-type semimetrics (Gretton et al., 2012; Muandet et al., 2017; Sejdinovic et al., 2013). In practice, sensitivity to kernel scale makes bandwidth selection pivotal: the median heuristic is a strong default whose asymptotic behavior in testing settings has been clarified, while multi-kernel MMD (MK-MMD) mitigates scale mismatch by aggregating bandwidths and is widely adopted in deep representation learning (Garreau et al., 2017; Long et al., 2015). For unsupervised domain adaptation, minimizing MMD between source and target features—popularized by Deep Adaptation Networks—yields effective feature alignment, and *weighted* MMD further corrects for class-prior shift when label proportions differ across domains (Long et al., 2015; Yan et al., 2017). Beyond testing and adaptation, MMD serves directly as a learning objective for generative modeling: Generative Moment Matching Networks match data and model distributions by minimizing MMD, and MMD-GAN variants learn kernels adversarially to improve sample fidelity, stability, and training efficiency (Li et al., 2015, 2017).

B Algorithm

In this section, we present the **LLM-XTM** training procedure:

C Dataset Statistics

- **EC News.** This corpus comprises parallel English–Chinese news articles covering six domains: business, education, entertainment, sports, technology, and fashion. Each language maintains a capped vocabulary of 5,000 tokens. The data are divided into 77,480 training samples (37,480 English, 40,000 Chinese) and 19,370 test samples (9,370 English, 10,000 Chinese).

- **Amazon Review.** This dataset contains bilingual product reviews in English and Chinese. We follow a binary sentiment setup, where reviews with a 5-star rating are labeled as positive and all others as negative. Both vocabularies are restricted to 5,000 words. The split includes 40,000 documents for training (20,000 per language) and 10,000 for evaluation (5,000 per language).

- **Rakuten Amazon.** This collection combines Japanese reviews from Rakuten with English reviews from Amazon, also used for binary sentiment prediction. Each language keeps a 5,000-word vocabulary limit, and the data are partitioned into 40,000 training and 10,000 test documents, balanced across languages.

For consistency with previous studies, we reproduce the experimental configuration of InfoCTM (Wu et al., 2023a) and employ its publicly

Algorithm 1 LLM-XTM training procedure

Require: Input corpus $\mathbf{X} = \mathbf{X}^{(1)} \cup \mathbf{X}^{(2)}$, topic number K, top-N words, refinement rounds R, refinement frequency F, temperature τ, loss weights $\lambda_{\mathrm{mmd}}, \lambda_{\mathrm{qa}}$.

Ensure: Optimized parameters $\Theta^* = \{\text{Encoders}^*, \beta^{(1)*}, \beta^{(2)*}\}$.

Initialize pretrained backbone cross-lingual topic model parameters Θ and optimizer.
for epoch $e = 1$ to N **do**
    Train backbone to compute θ_d and $\beta^{(1)}, \beta^{(2)}$ for mini-batches.
    if $e \bmod F = 0$ **then** ▷ Refinement step
        for each topic $k = 1$ to K **do**
            Extract top-N words from $\beta^{(1)}, \beta^{(2)}$.
            Run R LLM refinements and vote to obtain refined set $\bar{w}_k$.
        end for
        Build refined topic–word distributions $\beta^{(\text{refined})}$.
    end if
    Construct raw distributions $\beta^{(\text{raw})}$.
    Compute L_{MMD} between $\beta^{(\text{raw})}$ and $\beta^{(\text{refined})}$.
    Encode documents h_d and topics t_k with a multilingual encoder.
    Derive target $\hat{\theta}_d = \mathrm{softmax}(\cos(h_d, t_k)/\tau)$.
    Compute $L_{\mathrm{doc}} = \sum_d \mathrm{KL}(\theta_d \parallel \hat{\theta}_d)$.
    Update Θ by minimizing $\mathcal{J} = L_{\text{Phase1}} + \lambda_{\mathrm{mmd}} L_{\mathrm{MMD}} + \lambda_{\mathrm{qa}} L_{\mathrm{doc}}$.
end for

released preprocessed datasets: *EC News*, *Amazon Review*, and *Rakuten Amazon*.

Unlike the original setup, both word- and document-level embeddings in our framework are derived from the `BAAI/bge-m3` model (Chen et al., 2024), ensuring a unified multilingual semantic space for cross-lingual topic alignment.

D Classification Performance Within and Across Languages

Following standard practice, we use document–topic representations as input features for linear SVMs and evaluate both intra-lingual (-I) and cross-lingual (-C) classification accuracy on *EC News*, *Amazon Review*, and *Rakuten Amazon* (Table 5). Models without document–topic posteriors are excluded from comparison.

Across all datasets, integrating **LLM-XTM** consistently improves cross-lingual transfer while maintaining strong in-language performance. The largest gains appear on *Rakuten Amazon* (–C: 0.734 → 0.788 for InfoCTM; 0.682 → 0.728 for NMTM), with smaller but stable improvements on *EC News* and *Amazon Review*. In-language accuracy remains comparable or slightly higher (e.g., InfoCTM EN-I: 0.789 → 0.800).

These results show that **LLM-XTM** enhances cross-lingual generalization without compromising monolingual accuracy, consistent with its design objectives of MMD-based refinement and QA-driven alignment.

E Prompt for Cross-Lingual Topic Refinement

We provide the exact prompt used to query the Gemini API for topic refinement. The same template is applied to all topics, with the top-15 words from each language filled in.

```
Given the following cross-lingual topic words
from English and Chinese for N topics, refine
each topic:
1) Identify the main theme shared across both
languages.
2) Remove irrelevant/noisy words that do not
fit the theme.
3) Add relevant words that strengthen
coherence and cross-lingual coverage.
4) Use only SINGLE WORDS (no phrases, no
underscores, no hyphenated expressions).
5) Return exactly 15 words per language for
each topic.
Output format for all topics:
Topic <id>: <brief theme>
EN: word1 - word2 - ... - word15
CN: word1 - word2 - ... - word15
Rules:
- Exactly 15 words after EN: and CN:.
- Separate words with " - ".
- List topics in order from 0 to N−1.
```

Figure 4: Prompt used for cross-lingual topic refinement

F Detailed Prompts for LLM Evaluation

This appendix provides the detailed system prompts used for the LLM-based evaluation tasks described in the main text. Tables 6 and 7 shows the prompts side-by-side for intra-lingual coherence and cross-lingual similarity assessment across the different datasets.

Model	EC News				Amazon Review				Rakuten Amazon			
	EN-I	EN-C	ZH-I	ZH-C	EN-I	EN-C	ZH-I	ZH-C	EN-I	EN-C	JA-I	JA-C
InfoCTM	0.7813	0.5242	0.7652	0.5592	0.7894	0.6862	0.7262	0.5984	0.7940	0.6910	0.8256	0.7342
+ LLM-XTM	0.7849	0.5445	0.7682	0.5875	0.8003	0.6863	0.7321	0.6043	0.7984	0.7314	0.8296	0.7875
	↑0.5%	↑3.9%	↑0.4%	↑5.1%	↑1.4%	↑0.0%	↑0.8%	↑1.0%	↑0.6%	↑5.8%	↑0.5%	↑7.3%
NMTM	0.7917	0.4941	0.7742	0.5139	0.7859	0.5920	0.7210	0.5750	0.7966	0.6104	0.8264	0.6816
+ LLM-XTM	0.7975	0.5173	0.7785	0.5333	0.7859	0.6513	0.7071	0.6085	0.7923	0.6211	0.8334	0.7282
	↑0.7%	↑4.7%	↑0.6%	↑3.8%	↑0.0%	↑10.0%	↓1.9%	↑5.8%	↓0.5%	↑1.8%	↑0.9%	↑6.8%

Table 5: Document classification accuracy using document–topic vectors (θ) as features for linear SVMs. intra-lingual (-I) and cross-lingual (-C) accuracy across *EC News*, *Amazon Review*, and *Rakuten Amazon*.

Dataset	Prompt
EC News	You are a helpful assistant evaluating the top words of a topic model output for a given topic. The dataset is EC News, a collection of English and Chinese news with 6 categories: business, education, entertainment, sports, tech, and fashion. Please rate how related the following words are to each other 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 appropriateness of the topic.
Amazon Review	You are a helpful assistant evaluating the top words of a topic model output for a given topic. The dataset is Amazon Review, which includes English and Chinese reviews from the Amazon website. Please rate how related the following words are to each other 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 appropriateness of the topic.
Rakuten Amazon	You are a helpful assistant evaluating the top words of a topic model output for a given topic. The dataset is Rakuten Amazon, which contains Japanese reviews from Rakuten, and English reviews from Amazon. Please rate how related the following words are to each other 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 appropriateness of the topic.

Table 6: intra-lingual Coherence Prompts for LLM-based Evaluation

Dataset	Prompt
EC News	You are a helpful assistant evaluating the similarity of topics derived from topic modeling on parallel news corpora. The dataset is EC News, with English and Chinese news. You will be given two sets of top words, one for an English topic (Language 1) and one for a Chinese topic (Language 2). Please rate how similar the underlying topics represented by these two sets of words are, on a scale from 1 to 3 ("1"=not very similar, "2"=moderately similar, "3"=very similar). Reply with a single number.
Amazon Review	You are a helpful assistant evaluating the similarity of topics derived from topic modeling on parallel review corpora. The dataset is Amazon Review, with English and Chinese reviews. You will be given two sets of top words, one for an English topic (Language 1) and one for a Chinese topic (Language 2). Please rate how similar the underlying topics represented by these two sets of words are, on a scale from 1 to 3 ("1"=not very similar, "2"=moderately similar, "3"=very similar). Reply with a single number.
Rakuten Amazon	You are a helpful assistant evaluating the similarity of topics derived from topic modeling on parallel review corpora. The dataset is Rakuten Amazon, with Japanese reviews (Rakuten - Language 2) and English reviews (Amazon - Language 1). You will be given two sets of top words, one for an English topic and one for a Japanese topic. Please rate how similar the underlying topics represented by these two sets of words are, on a scale from 1 to 3 ("1"=not very similar, "2"=moderately similar, "3"=very similar). Reply with a single number.

Table 7: cross-lingual Similarity Prompts for LLM-based Evaluation

Model	Configuration	CNPMI	TU	TQ
u-SVD	Baseline	0.0281	0.5893	0.0165
SVD-LR	Baseline	0.0312	0.5940	0.0185
InfoCTM	Base	0.0279	0.7953	0.0222
	+ LLM-XTM	0.0497	0.7213	0.0358
	Improvement	+78.1%	-9.3%	+61.3%
XTRA	Base	0.0106	0.7087	0.0075
	+ LLM-XTM	0.0207	**0.8027**	0.0166
	Improvement	+95.3%	+13.3%	+121.3%
NMTM	Base	0.0417	0.6213	0.0259
	+ LLM-XTM	**0.0531**	0.6987	**0.0371**
	Improvement	+27.3%	+12.6%	+43.2%

Table 8: Results on the Airiti Thesis long-document benchmark (Chang et al., 2024). CNPMI is computed using the Wikipedia reference corpus. Bold indicates the best value in each metric column among the reported systems.

G Implementation Details

We run experiments on a single NVIDIA P100 GPU (Kaggle). Training is divided into two phases: a **base training** phase , followed by **LLM-based refinement** phase of 30 epochs.

Refinement is performed over 5 rounds using these settings:

- Refinement frequency: {8, 10} steps
- MMD loss weight: 20 000
- Document alignment loss weight: {100, 200, 300}

H Generalization to Long Documents (Airiti Thesis)

To test generalization beyond short reviews and news, we evaluate LLM-XTM on the Airiti Thesis dataset introduced by Chang et al. (Chang et al., 2024), which contains longer academic thesis abstracts in Chinese and English. Following Chang et al. (2024), we use the same data source and preprocessing setting; all CNPMI scores in Table 8 are computed against the Wikipedia reference corpus, so coherence is measured independently of the training text.

LLM-XTM yields substantial gains across all three neural backbones on this long-document benchmark. XTRA more than doubles Topic Quality (+121.3%), InfoCTM improves by +61.3%, and NMTM improves by +43.2%. The strongest overall TQ is achieved by **NMTM + LLM-XTM** (0.0371), which is more than 2× the best clustering baseline SVD-LR (0.0185), while Topic Uniqueness remains robust and even increases for both XTRA and NMTM.

I Model-Agnostic Evaluation with Open-Weight LLMs

To test whether the enhancement depends on a proprietary API, we replace Gemini 2.5 Flash with several open-weight alternatives on Amazon Review (NMTM backbone): Llama-3.3-70B (Grattafiori et al., 2024), Qwen3-Coder-480B-A35B-Instruct (Qwen Team, 2025; Yang et al., 2025a), and Mistral-Small-24B (Mistral AI, 2025). We additionally include a compact 7B Mistral variant to probe the low-cost end of the trade-off. The quantitative comparison is reported in Table 9.

The results confirm that LLM-XTM is model-agnostic rather than tied to a single proprietary engine. As shown in Table 9, **Llama-3.3-70B** matches Gemini 2.5 Flash at TQ 0.0351, showing that an open-weight model can reach the same overall quality. **Qwen3-Coder-480B-A35B-Instruct** and **Mistral-Small-24B** still deliver strong gains (+27.5% and +24.0%), while even the 7B Mistral variant improves TQ by +13.4%. The overall pattern suggests a predictable trade-off: smaller models preserve slightly higher uniqueness, whereas larger models maximize CNPMI and TQ.

J Additional LLM-based Evaluation Results

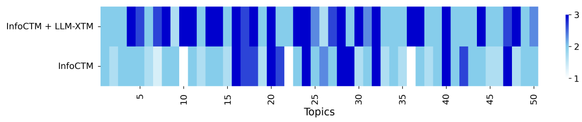


Figure 5: English intra-lingual semantic similarity (Amazon Review, InfoCTM).

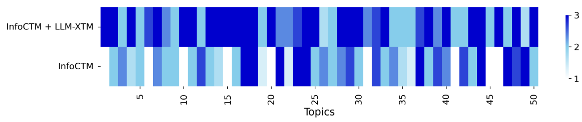


Figure 6: Chinese intra-lingual semantic similarity (Amazon Review, InfoCTM).

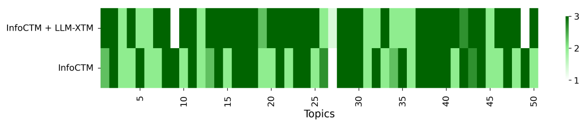


Figure 7: Cross-lingual semantic similarity on Amazon Review (InfoCTM).

Refinement Engine	Model Size	CNPMI	TU	TQ	vs. Baseline
Base Model (NMTM)	–	0.0430	0.6100	0.0262	–
Gemini 2.5 Flash	–	**0.0560**	0.6270	**0.0351**	+34.0%
Llama-3.3-70B	70B	0.0544	0.6453	**0.0351**	+34.0%
Qwen3-Coder-480B-A35B-Instruct	480B (35B active)	0.0508	0.6567	0.0334	+27.5%
Mistral-Small-24B	24B	0.0506	0.6413	0.0325	+24.0%
Mistral-7B-v0.3	7B	0.0445	**0.6673**	0.0297	+13.4%

Table 9: Open-weight LLM comparison on Amazon Review with the NMTM backbone. Bold indicates the best value in each metric column.

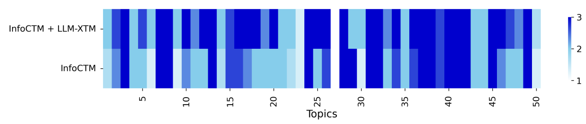


Figure 8: English intra-lingual semantic similarity (EC-News, InfoCTM).

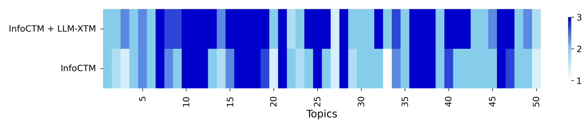


Figure 9: Chinese intra-lingual semantic similarity (EC-News, InfoCTM).

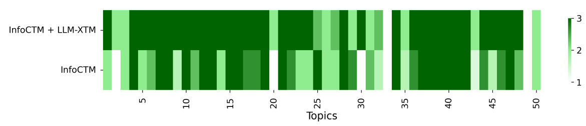


Figure 10: Cross-lingual semantic similarity on EC-News (InfoCTM).

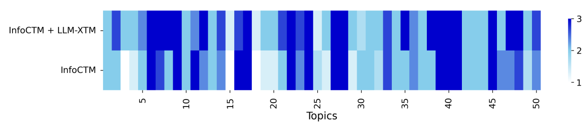


Figure 11: English intra-lingual semantic similarity (Rakuten_Amazon, InfoCTM).

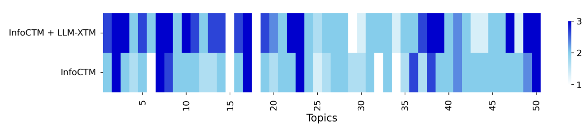


Figure 12: Japanese intra-lingual semantic similarity (Rakuten_Amazon, InfoCTM).

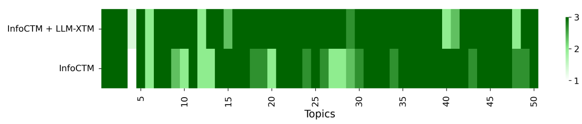


Figure 13: Cross-lingual semantic similarity on Rakuten_Amazon (InfoCTM).

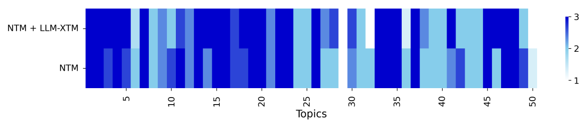


Figure 14: English intra-lingual semantic similarity (Amazon Review, NMTM).

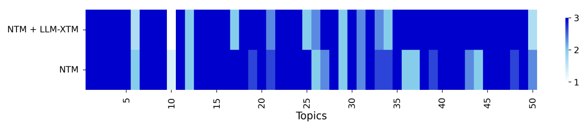


Figure 15: Chinese intra-lingual semantic similarity (Amazon Review, NMTM).

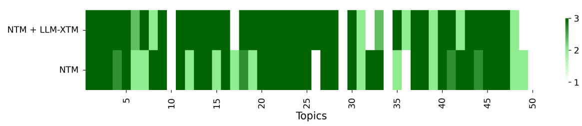


Figure 16: Cross-lingual semantic similarity on Amazon Review (NMTM).

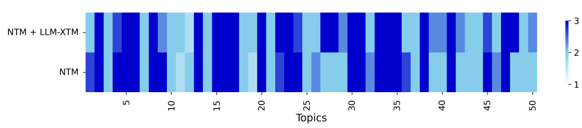


Figure 17: English intra-lingual semantic similarity (EC-News, NMTM).

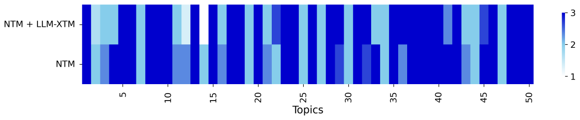


Figure 18: Chinese intra-lingual semantic similarity (ECNews, NMTM).

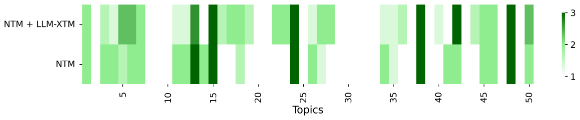


Figure 19: Cross-lingual semantic similarity on EC-News (NMTM).

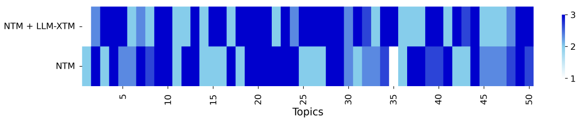


Figure 20: English intra-lingual semantic similarity (Rakuten_Amazon, NMTM).

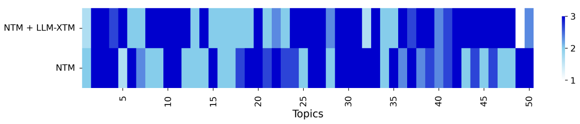


Figure 21: Japanese intra-lingual semantic similarity (Rakuten_Amazon, NMTM).

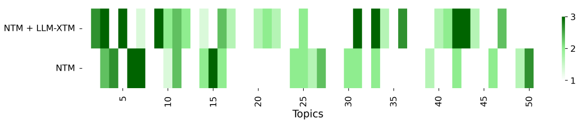


Figure 22: Cross-lingual semantic similarity on Rakuten_Amazon (NMTM).

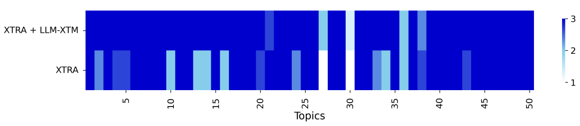


Figure 23: English intra-lingual semantic similarity (Amazon Review, XTRA).

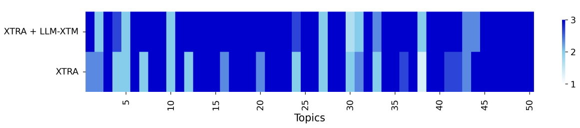


Figure 24: Chinese intra-lingual semantic similarity (Amazon Review, XTRA).

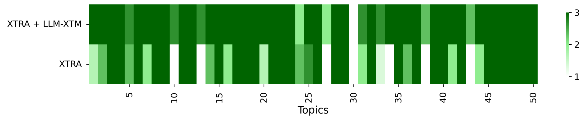


Figure 25: Cross-lingual semantic similarity on Amazon Review (XTRA).

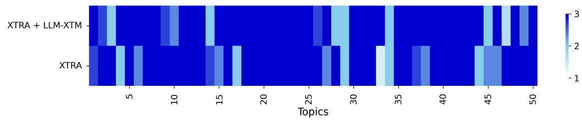


Figure 26: English intra-lingual semantic similarity (EC-News, XTRA).

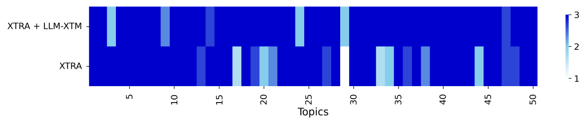


Figure 27: Chinese intra-lingual semantic similarity (ECNews, XTRA).

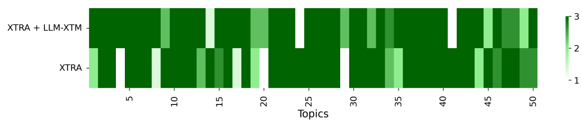


Figure 28: Cross-lingual semantic similarity on EC-News (XTRA).

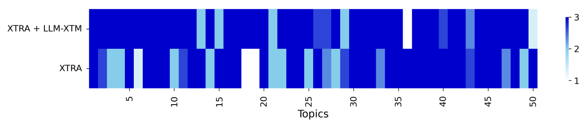


Figure 29: English intra-lingual semantic similarity (Rakuten_Amazon, XTRA).

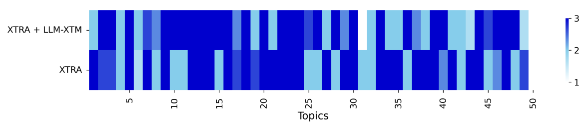


Figure 30: Japanese intra-lingual semantic similarity (Rakuten_Amazon, XTRA).

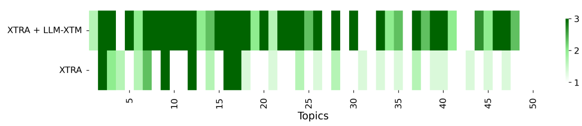


Figure 31: Cross-lingual semantic similarity on Rakuten_Amazon (XTRA).