A Large Language Model Guided Topic Refinement Mechanism for Short Text Modeling

Shuyu Chang[1,*], Rui Wang[1,*], Peng Ren[1], Qi Wang[1], and Haiping Huang[1,2](✉)

[1] Nanjing University of Posts and Telecommunications, Nanjing, Jiangsu, China
[2] Jiangsu High Technology Research Key Laboratory for Wireless Sensor Networks, Nanjing, Jiangsu, China
{shuyu_chang,rui_wang,1022041122,1223045440,hhp}@njupt.edu.cn

Abstract. Effectively modeling topics in short texts, such as tweets and news snippets, is crucial to understanding rapidly evolving social trends. However, existing topic models struggle to capture the underlying semantic patterns of short texts due to their inherent data sparsity. This characteristic leads to insufficient co-occurrence information, impairing the coherence and granularity of extracted topics. To address these challenges, we introduce Topic Refinement, a novel model-agnostic mechanism for short-text topic modeling that leverages the advanced text comprehension capabilities of Large Language Models (LLMs). Unlike traditional approaches, this post-processing mechanism employs prompt engineering to refine topics across various modeling methods. We guide LLMs to identify semantically intruder words within the extracted topics and suggest coherent alternatives to replace them. This process emulates human-like identification, evaluation, and refinement of the discovered topics. Extensive experiments on four diverse datasets demonstrate that Topic Refinement improves the topic quality and performance in topic-related text classification tasks.

Keywords: Topic modeling · Topic refinement · Large language models · Short texts · Prompt engineering.

1 Introduction

Short texts, such as online comments and news headlines, are important in mirroring public opinion [15]. Topic modeling emerges as a valuable tool to extract topics from vast data [13, 16]. Nevertheless, the sparsity of context in short texts poses challenges for conventional topic models [1, 20]. This brevity leads to low information density [10], making it difficult to capture and represent the semantics of discussed topics.

Recent advancements have alleviated these challenges by enhancing word co-occurrence patterns [19, 3] and incorporating external knowledge [6] to strengthen semantic relationships. Furthermore, approaches like quantified distribution analysis [18] and contrastive learning techniques [8] have been employed to improve

[*] These authors contributed equally to this work.

the learning signals derived from short texts. Despite these efforts, existing methods frequently yield suboptimal semantic coherence and granularity (i.e., clarity and diversity from a user perspective) [7] of topics in practical applications.

In light of these limitations, we shift to the widely concerned Large Language Models (LLMs) such as PaLM [4], GPT [2], and others. With their unprecedented semantic comprehension capabilities trained across extensive datasets, LLMs are expected to further improve topic modeling quality for short texts. Existing LLM-based topic modeling frameworks like PromptTopic [12] and TopicGPT [9] generate topics by inputting documents to produce topic categorizations or descriptions. However, they deviate from the classic bag-of-words format, complicating evaluation and comparison with traditional topic modeling outputs. These models also require whole documents as inputs, leading to high token consumption and computational costs.

To bridge this gap, we introduce Topic Refinement, a novel model-agnostic mechanism that leverages LLMs via prompt engineering to refine topics initially mined by base topic models. This mechanism iterates over each topic and sequentially assumes a word as the potential intruder word while prompting LLMs to identify the topic and assess word alignment. If alignment is confirmed, the word is retained; otherwise, coherent words are generated as candidates to replace the intruder word. Topic Refinement emulates the human process of identifying, evaluating, and refining extracted topics to make them more coherent and accessible.

The main contributions of our work are summarized as follows:

- We introduce Topic Refinement for short text modeling, pioneering a new pathway to improve topic quality after initial mining.
- Topic Refinement integrates LLMs with diverse base topic models to precisely evaluate and refine the representative words for each extracted topic.
- Extensive experiments on four datasets demonstrate improvements over existing base models across various metrics.

2 Methodology

2.1 Problem Definition

Topic Refinement aims to enhance the coherence and granularity of topics generated from base topic models for short texts. Formally, given a set of topics $T = \{t_1, t_2, \cdots, t_K\}$ discovered by a base topic model, each topic consists of N representative words, denoted by $t_i = \{w_{i1}, w_{i2}, \cdots w_{iN}\}$, where $1 \leq i \leq K$. In this context, an intruder word is defined as a word that is not semantically coherent with other words in the topic. The Topic Refinement transforms each base topic t_i into a refined topic t'_i by determining whether w_{ij} is an intruder word in t_i and correcting this intruder word. This transformation is guided by a refinement function $\mathcal{R}$:

$$\mathcal{R} : (t_i \setminus \{w_{ij}\}, w_{ij}) \rightarrow w'_{ij}, \tag{1}$$

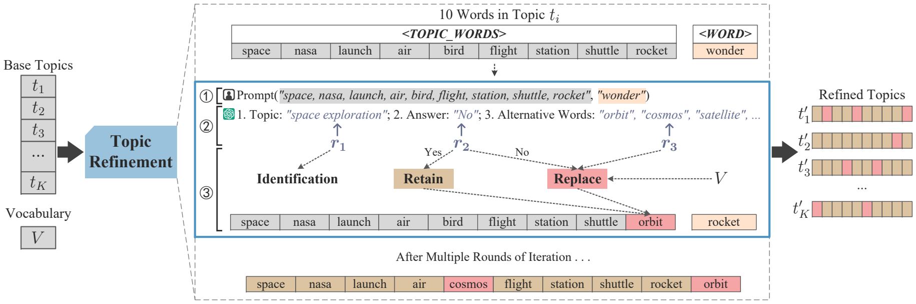
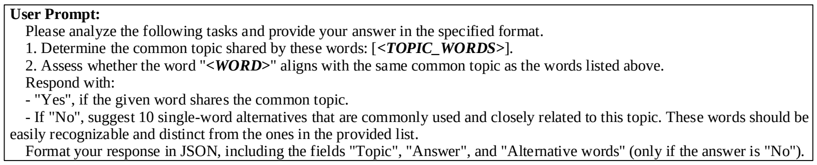

Base Topics
t_1
t_2
t_3
...
t_K
Vocabulary
V
Topic Refinement
10 Words in Topic t_i
<TOPIC_WORDS>
space nasa launch air bird flight station shuttle rocket
<WORD>
wonder
①
Prompt("space, nasa, launch, air, bird, flight, station, shuttle, rocket", "wonder")
②
1. Topic: "space exploration"; 2. Answer: "No"; 3. Alternative Words: "orbit", "cosmos", "satellite", ...
r_1
r_2
r_3
Yes
No
③
Identification
Retain
Replace
V
space nasa launch air bird flight station shuttle orbit
rocket
After Multiple Rounds of Iteration . . .
space nasa launch air cosmos flight station shuttle rocket orbit
Refined Topics
t'_1
t'_2
t'_3
...
t'_K

Fig. 1. Overview of Topic Refinement ($N = 10$).


User Prompt:

Please analyze the following tasks and provide your answer in the specified format.

1. Determine the common topic shared by these words: [<***TOPIC_WORDS***>].

2. Assess whether the word "<***WORD***>" aligns with the same common topic as the words listed above.

Respond with:

- "Yes", if the given word shares the common topic.
- If "No", suggest 10 single-word alternatives that are commonly used and closely related to this topic. These words should be easily recognizable and distinct from the ones in the provided list.

Format your response in JSON, including the fields "Topic", "Answer", and "Alternative words" (only if the answer is "No").

Fig. 2. LLM prompt template for Topic Refinement.

where w'_{ij} is the alternative word for topic t_i. We propose to model the refinement function via LLMs, leveraging their advanced capabilities in text understanding.

2.2 Topic Refinement

Figure 1 depicts the overall architecture of our Topic Refinement mechanism. We first extract topics $T = \{t_1, t_2, \cdots, t_K\}$ and vocabulary V from a short text dataset through a base topic model (e.g., probabilistic or neural topic models). Our mechanism then unfolds in three phases: ① Prompt Construction, ② LLM Request and Response, and ③ Iterative Refinement.

Prompt Construction To model the refinement function $\mathcal{R}$, our prompt template guides LLMs to identify intruder words in each topic and generate alternative words related to this topic. Given topic t_i with words $\{w_{i1}, w_{i2}, \cdots w_{iN}\}$, we structure the prompt into two tasks: (1) LLM $\mathcal{M}$ identifies the common topic shared by words $t_i \setminus \{w_{ij}\}$, represented by $<$*TOPIC_ WORDS*$>$; (2) $\mathcal{M}$ assesses whether word w_{ij}, denoted as $<$*WORD*$>$, is semantically consistent with this topic, and if not, suggests ten alternative words closely related to that topic as candidate alternatives. For the clarity and diversity of topics, these candidate words are required to be easily recognizable and distinctly different from those in the provided list $t_i \setminus \{w_{ij}\}$. The prompt template, defined as $p = \text{Prompt}(<$*TOPIC_ WORDS*$>, <$*WORD*$>)$, also enforces JSON output format for easier processing. The detailed template is shown in Figure 2.

Table 1. Statistics of datasets after preprocessing.

Datasets	#Docs	#Labels	#Vocab.	Avg. Length
Tweet	2,133	89	1,127	5.550
AGNews	14,845	4	3,290	4.268
TagMyNews	27,369	7	4,325	4.483
YahooAnswer	12,258	10	3,423	4.151

LLM Request and Response After constructing prompt $p = \mathrm{Prompt}(t_i \setminus \{w_{ij}\}, w_{ij})$, we query LLM $\mathcal{M}$ and process its response. Essentially, $\mathcal{M}$ models the refinement function by estimating the conditional probability of generated text: $\mathrm{P}_{\mathrm{LLM}}(s_1, s_2, \cdots, s_n \mid p)$, where $(s_1, s_2, \cdots, s_n)$ is the generated word sequence with variable lengths. The LLM output is formatted in JSON and mapped to a tuple (r_1, r_2, r_3). Here, r_1 is the identified topic, r_2 is a binary judgment ("*Yes*" or "*No*") of whether w_{ij} aligns with the topic, and r_3 contains ten replacement candidates if r_2 is "*No*", empty otherwise. This process above is formalized as:

$$(r_1, r_2, r_3) = \mathcal{M}(\mathrm{Prompt}(t_i \setminus \{w_{ij}\}, w_{ij})). \tag{2}$$

Iterative Refinement This phase, informed by the outputs from LLM $\mathcal{M}$, selects a more coherent alternative to replace the identified intruder word. When r_2 indicates w_{ij} lies outside the semantic boundary of the topic t_i, we first attempt to select w'_{ij} from the alternative words in r_3 based on vocabulary V. If no word from r_3 exists in V, we compute their embedding centroid and select the closest word from V that is not already in t_i. This ensures that the alternative word w'_{ij} is not only semantically coherent but also retains the lexical integrity of the dataset. As words are generally arranged according to topic relevance, the process iterates over each word in reverse order to yield refined topics t'_i.

3 Experiments

3.1 Experimental Setup

Datasets We evaluate on four short text datasets: **Tweet** [21] from the 2011 and 2012 TREC microblog tracks, **AGNews** [22] with randomly sampled 5,000 news titles per category gathered by ComeToMyHead, **TagMyNews** [11] with news titles from various sources, and **YahooAnswer** [22] with randomly sampled 2,500 question titles per category from Yahoo Answers. Table 1 presents their statistics after preprocessing [18].

Base Models We evaluate Topic Refinement with nine representative base topic models: (1) **LDA** [1] that uses Gibbs sampling for parameter estimation, with GibbsLDA++ implementation. (2) **BTM** [3] that enriches word co-occurrence within short texts. (3) **Gaussian-BAT** [14] that incorporates word semantics by Gaussian modeling. (4) **CNTM** [8] that leverages contrastive learning and

Table 2. Topic coherence results on four datasets.

Datasets	**Base Models**	C_A	Δ_{C_A}	C_P	Δ_{C_P}	C_V	Δ_{C_V}	**UCI**	Δ_{UCI}	**NPMI**	Δ_{NPMI}
Tweet	LDA	0.175	+0.045	0.142	+0.223	0.375	+0.022	-0.675	+1.452	0.001	+0.086
	BTM	0.191	+0.022	0.152	+0.122	0.377	+0.017	-0.682	+0.705	0.002	+0.040
	G-BAT	0.160	+0.035	-0.005	+0.187	0.419	−0.012	-1.992	+1.348	-0.052	+0.065
	CNTM	0.183	+0.032	0.153	+0.120	0.382	+0.016	-0.744	+0.675	0.001	+0.040
	TSCTM	0.182	+0.027	0.134	+0.158	0.392	+0.014	-0.993	+0.900	-0.008	+0.050
	BERTopic	0.205	+0.016	0.203	+0.089	0.394	+0.011	-0.644	+0.588	0.012	+0.032
	ECRTM	0.191	+0.026	0.109	+0.152	0.421	+0.002	-1.920	+1.182	-0.039	+0.058
	CWTM	0.211	+0.021	0.191	+0.157	0.415	+0.012	-1.342	+1.322	-0.013	+0.065
	LLM-TM	0.154	+0.035	-0.026	+0.206	0.392	−0.005	-1.964	+1.443	-0.059	+0.072
AGNews	LDA	0.174	+0.026	0.220	+0.110	0.373	+0.021	0.062	+0.410	0.036	+0.028
	BTM	0.166	+0.039	0.153	+0.196	0.367	+0.028	-0.034	+0.736	0.023	+0.051
	G-BAT	0.181	+0.023	0.192	+0.145	0.398	+0.005	-0.622	+0.837	0.012	+0.044
	CNTM	0.266	+0.013	0.469	+0.059	0.448	+0.010	0.744	+0.275	0.101	+0.017
	TSCTM	0.205	+0.033	0.155	+0.224	0.437	+0.005	-1.705	+1.704	-0.022	+0.082
	BERTopic	0.228	+0.029	0.295	+0.138	0.442	+0.009	-0.458	+0.859	0.038	+0.045
	ECRTM	0.203	+0.051	0.125	+0.339	0.443	+0.004	-2.387	+2.630	-0.053	+0.125
	CWTM	0.173	+0.017	0.206	+0.094	0.387	+0.006	-0.678	+0.606	0.005	+0.032
	LLM-TM	0.166	+0.042	0.081	+0.206	0.362	+0.030	-1.296	+1.262	-0.033	+0.072
TagMyNews	LDA	0.189	+0.027	0.237	+0.147	0.391	+0.018	0.021	+0.594	0.038	+0.039
	BTM	0.177	+0.039	0.142	+0.215	0.389	+0.020	-0.594	+1.039	0.005	+0.061
	G-BAT	0.216	+0.021	0.332	+0.082	0.407	+0.017	0.302	+0.417	0.061	+0.026
	CNTM	0.293	+0.010	0.515	+0.060	0.475	+0.009	0.608	+0.421	0.103	+0.021
	TSCTM	0.209	+0.033	0.153	+0.179	0.465	+0.003	-1.911	+1.343	-0.027	+0.066
	BERTopic	0.217	+0.018	0.287	+0.129	0.459	+0.010	-1.090	+0.833	0.012	+0.042
	ECRTM	0.190	+0.045	0.086	+0.261	0.451	−0.008	-2.623	+2.114	-0.064	+0.100
	CWTM	0.198	+0.016	0.244	+0.079	0.407	+0.009	-0.542	+0.475	0.019	+0.027
	LLM-TM	0.158	+0.046	0.047	+0.254	0.382	+0.024	-1.623	+1.581	-0.045	+0.088
YahooAnswer	LDA	0.190	+0.030	0.278	+0.086	0.378	+0.018	0.510	+0.266	0.063	+0.023
	BTM	0.172	+0.032	0.185	+0.123	0.365	+0.024	0.219	+0.308	0.036	+0.029
	G-BAT	0.213	+0.022	0.265	+0.121	0.399	+0.016	0.304	+0.426	0.062	+0.028
	CNTM	0.294	+0.007	0.534	+0.020	0.455	+0.006	1.311	+0.109	0.141	+0.006
	TSCTM	0.215	+0.024	0.234	+0.160	0.437	+0.005	-0.919	+1.092	0.016	+0.056
	BERTopic	0.252	+0.006	0.421	+0.036	0.448	+0.003	0.533	+0.197	0.094	+0.011
	ECRTM	0.247	+0.045	0.315	+0.181	0.454	+0.016	-0.829	+1.419	0.027	+0.072
	CWTM	0.170	+0.024	0.220	+0.098	0.375	+0.009	0.014	+0.504	0.032	+0.030
	LLM-TM	0.152	+0.042	0.044	+0.216	0.381	+0.012	-1.251	+1.199	-0.031	+0.067

word embedding to capture semantic patterns. (5) **TSCTM** [18] that is based on contrastive learning with quantified sampling. (6) **BERTopic** [6] that utilizes pre-trained language models to extract topics. (7) **ECRTM** [17] that forces each topic embedding to be the center of a cluster in semantic space via optimal transport. (8) **CWTM** [5] that integrates contextualized word embeddings from BERT to learn topic vectors. (9) **LLM-TM** [7] that prompts LLMs to generate topics from a given set of documents. We adapt the prompts to generate words per topic for a standardized evaluation. All models are executed with their recommended hyperparameters.

Evaluation Metrics Following previous mainstream work, we evaluate the topic quality by topic coherence metrics. We utilize Palmetto[3] to calculate these coherence metrics C_A, C_P, C_V, UCI, and NPMI, with higher values indicating

[3] https://github.com/dice-group/Palmetto

Table 3. Topic granularity results on four datasets.

Base Models	*Tweet*				*AGNews*				*TagMyNews*				*YahooAnswer*			
	$\boldsymbol{\mathcal{S}}$	$\boldsymbol{\Delta_{\mathcal{S}}}$	$\boldsymbol{\mathcal{D}}$	$\boldsymbol{\Delta_{\mathcal{D}}}$	$\boldsymbol{\mathcal{S}}$	$\boldsymbol{\Delta_{\mathcal{S}}}$	$\boldsymbol{\mathcal{D}}$	$\boldsymbol{\Delta_{\mathcal{D}}}$	$\boldsymbol{\mathcal{S}}$	$\boldsymbol{\Delta_{\mathcal{S}}}$	$\boldsymbol{\mathcal{D}}$	$\boldsymbol{\Delta_{\mathcal{D}}}$	$\boldsymbol{\mathcal{S}}$	$\boldsymbol{\Delta_{\mathcal{S}}}$	$\boldsymbol{\mathcal{D}}$	$\boldsymbol{\Delta_{\mathcal{D}}}$
LDA	0.254	+0.048	13.12	+4.02	0.279	+0.045	12.25	+4.13	0.280	+0.053	14.17	+4.47	0.368	+0.034	15.93	+4.01
BTM	0.259	+0.048	14.03	+3.86	0.267	+0.067	10.41	+6.13	0.249	+0.083	11.73	+6.64	0.345	+0.037	11.30	+4.96
G-BAT	0.248	+0.058	14.21	+4.06	0.288	+0.056	15.44	+3.79	0.328	+0.046	19.29	+3.43	0.367	+0.039	18.77	+3.24
CNTM	0.253	+0.046	14.35	+3.39	0.327	+0.038	22.31	+2.95	0.341	+0.035	25.88	+2.59	0.446	+0.008	27.97	+0.83
BERTopic	0.274	+0.033	16.18	+2.57	0.237	+0.090	16.45	+6.17	0.255	+0.080	20.62	+5.41	0.318	+0.056	21.42	+3.87
TSCTM	0.252	+0.061	14.46	+4.81	0.288	+0.048	19.42	+3.59	0.279	+0.041	19.78	+4.14	0.381	+0.012	25.22	+0.99
ECRTM	0.245	+0.060	16.46	+4.33	0.218	+0.122	15.18	+7.94	0.243	+0.113	18.24	+7.80	0.333	+0.065	23.46	+4.14
CWTM	0.273	+0.062	15.32	+5.80	0.296	+0.033	14.50	+2.40	0.304	+0.032	16.53	+2.25	0.347	+0.028	15.26	+2.57
LLM-TM	0.218	+0.076	11.17	+5.06	0.231	+0.091	11.68	+5.81	0.234	+0.108	11.89	+7.12	0.263	+0.071	12.31	+5.65

better topic quality. Inspired by Mu *et al.* [7], we also introduce two topic granularity metrics based on word embeddings to evaluate topic clarity and diversity: within-topic similarity $\mathcal{S}$ and between-topic distance $\mathcal{D}$.

$$\mathcal{S} = \frac{2}{KN(N-1)} \sum_{i}^{K} \sum_{j}^{N-1} \sum_{k=j+1}^{N} \frac{e_{w_{ij}} \cdot e_{w_{ik}}}{\left\| e_{w_{ij}} \right\|_2 \left\| e_{w_{ik}} \right\|_2}, \tag{3}$$

$$\mathcal{D} = \frac{2}{K(K-1)} \sum_{i}^{K-1} \sum_{m=i+1}^{K} \sum_{d}^{D} \left| e_{t_i}^{(d)} - e_{t_m}^{(d)} \right|^2, \tag{4}$$

where $e_{w_{ij}}$ is the word embedding, $e_{t_i} = \frac{1}{N} \sum_{j}^{N} e_{w_{ij}}$ is the topic centroid, and D is the embedding dimension. The symbol Δ represents the changes in metric after refinement. Additionally, we evaluate the impact of Topic Refinement on text classification using Accuracy and F1 metrics.

Implementation Details For our Topic Refinement mechanism, we use GPT-3.5-turbo as the LLM $\mathcal{M}$ and set its temperature as 0 to lower the completion randomness. Keeping with prior work, each topic t_i mined by base topic models is represented by N words, where N is typically set to 10. The evaluation metrics based on word embeddings are calculated utilizing the GloVe[4], where $D = 300$.

3.2 Topic Quality Evaluation

Topic Coherence We evaluate base topic models with $K = 20$ and 50 in this part. Results in Table 2 show that Topic Refinement consistently improves the topic coherence across various datasets and base models, especially where base models initially performed poorly. The mechanism effectively mitigates the challenges of data sparsity in short texts and enhances those with initially high coherence, such as CNTM and BERTopic. While LLM-TM shows promise, it struggles with bag-of-words format generation, particularly for short texts. Even when the correct format is achieved, maintaining topic coherence can be problematic, highlighting the necessity for further refinement.

[4] https://nlp.stanford.edu/data/glove.840B.300d.zip

Table 4. Average token costs for each topic.

Methods	*Tweet*		*AGNews*		*TagMyNews*		*YahooAnswer*	
	#Input	**#Output**	**#Input**	**#Output**	**#Input**	**#Output**	**#Input**	**#Output**
Ours	1,591.01	274.90	1,588.60	264.22	1,593.70	275.83	1,578.43	231.50
LLM-TM	4,958.73	1,214.07	33,382.61	7,541.41	61,872.93	14,448.79	27,358.56	5,913.36

Table 5. Performance results of text classification on four datasets. *Rate*↑ means the growth rate of Topic Refinement.

Base Models	*Tweet*				*AGNews*				*TagMyNews*				*YahooAnswer*			
	Acc	***Rate*** ↑	**F1**	***Rate*** ↑	**Acc**	***Rate*** ↑	**F1**	***Rate*** ↑	**Acc**	***Rate*** ↑	**F1**	***Rate*** ↑	**Acc**	***Rate*** ↑	**F1**	***Rate*** ↑
LDA	0.913	16.98%	0.720	18.19%	0.705	12.34%	0.702	10.26%	0.705	9.22%	0.676	9.76%	0.621	0.64%	0.614	1.14%
BTM	0.897	17.50%	0.692	9.54%	0.701	5.56%	0.706	5.95%	0.686	11.81%	0.656	12.35%	0.616	0.32%	0.610	0.66%
G-BAT	0.874	19.57%	0.672	26.79%	0.698	3.87%	0.703	3.98%	0.680	1.03%	0.645	0.93%	0.610	1.15%	0.602	1.50%
CNTM	0.881	16.00%	0.629	10.02%	0.726	3.31%	0.730	3.84%	0.701	11.70%	0.672	12.65%	0.628	3.34%	0.621	3.86%
BERTopic	0.897	23.19%	0.683	16.54%	0.704	1.70%	0.707	1.70%	0.690	9.57%	0.655	9.92%	0.619	1.62%	0.612	1.96%
TSCTM	0.895	15.20%	0.688	9.88%	0.718	5.85%	0.715	5.03%	0.693	7.79%	0.660	7.88%	0.626	0.96%	0.618	0.81%
ECRTM	0.897	9.70%	0.675	8.30%	0.587	3.41%	0.593	5.06%	0.650	6.77%	0.617	8.10%	0.598	3.34%	0.592	4.90%
CWTM	0.848	22.88%	0.568	21.13%	0.726	10.61%	0.727	10.45%	0.683	16.54%	0.651	15.21%	0.617	3.57%	0.611	3.93%
LLM-TM	0.883	11.10%	0.658	8.66%	0.729	6.45%	0.730	5.89%	0.686	4.23%	0.650	3.69%	0.619	1.45%	0.610	1.64%

Topic Granularity Table 3 presents average improvements in within-topic similarity $\mathcal{S}$ and between-topic distance $\mathcal{D}$ across all models and datasets. Base models with higher topic granularity typically exhibit higher values of $\mathcal{S}$ and $\mathcal{D}$, indicating that their topics are more distinguishable and easier to understand. Models with lower initial topic granularity see greater increases from Topic Refinement, demonstrating the model-agnostic nature of our mechanism.

Token Cost Table 4 shows the average token costs for each topic across various datasets. The input token costs are relatively consistent across datasets due to the similar prompts. This contrasts with LLM-TM, where token costs scale with dataset size. Our approach, however, links token costs primarily to the number of topics rather than dataset size, which is more efficient and economical. Interestingly, datasets with higher initial topic quality (e.g., YahooAnswer) require fewer output tokens. The implication is that datasets with lower modeling complexity require less intervention from our mechanism.

3.3 Text Classification

We also evaluate the performance in text classification tasks achieved by refining base topics. Our experiment utilizes the SentenceTransformer[5] model to generate embeddings for each document and topic. Document-topic distributions are computed via cosine similarity between their embeddings. We employ an SVM classifier for text classification, with an 8:2 train-test split. The results in Table 5 confirm that Topic Refinement improves the performance of topic-related text classification tasks. We can see that base models that effectively capture high-quality topics generally exhibit better classification performance. Moreover, datasets with higher improvements in topic quality (i.e. Tweet) after

[5] https://huggingface.co/sentence-transformers/all-MiniLM-L12-v2

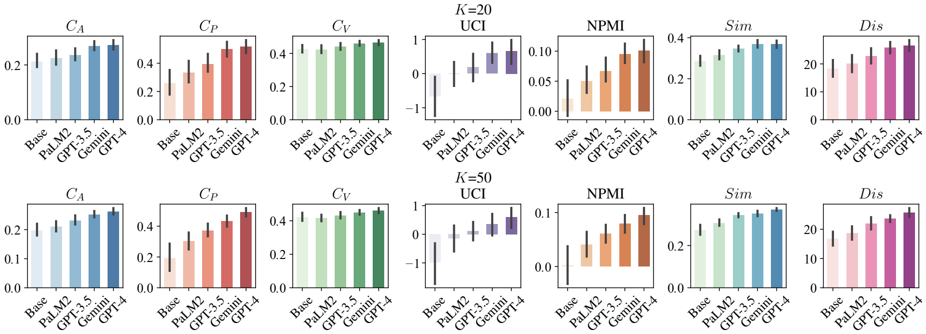


Fig. 3. Ablation study results with various LLMs. The error bars denote a 95% confidence interval for the statistical variability of results across nine base models.

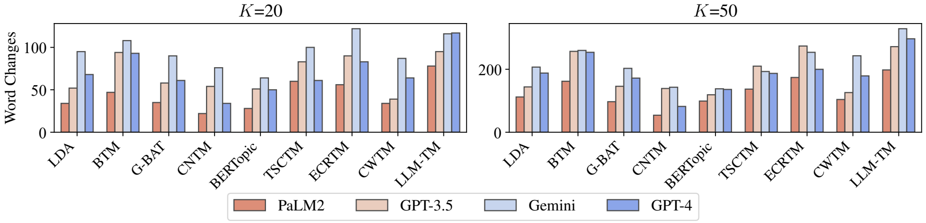


Fig. 4. Number of word changes between base topics and refined topics.

refinement show a more significant increase in classification performance. These findings suggest that Topic Refinement enhances the topic utility, which benefits the SVM classifier in discerning relevant textual features.

3.4 Ablation Study

We carry out ablation studies to evaluate the impact of four LLMs (PaLM2, GPT-3.5, Gemini Pro, and GPT-4) on the TagMyNews dataset with $K = 20$ and 50. Figure 3 compares quality metrics within the topics after Topic Refinement and all tested LLMs improve topic quality. Notably, LLMs with more recent and superior capabilities exhibit better refinement performance. Figure 4 illustrates the number of word changes in K topics after Topic Refinement. Although Gemini Pro and GPT-4 yield comparable results in quality metrics, GPT-4 shows fewer topic alterations, indicating its precision in refinement.

3.5 Visualization and Case Study

To provide an intuitive understanding of how Topic Refinement affects topic quality, we visualize the coherence of the topics from ECRTM and LLM-TM on the TagMyNews with $K = 20$. These two base models are selected because they exhibit the most word changes upon refinement. Figure 5 compares the

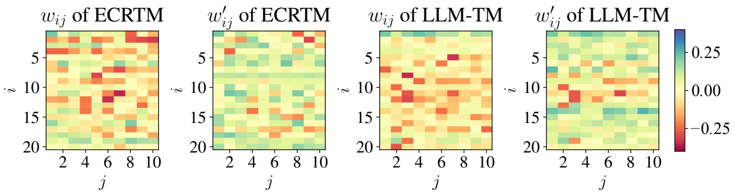


Fig. 5. Visualization of NPMI for word w_{ij} and w'_{ij}.

Table 6. Case study of base and refined topics. The replaced words in base topics are in red, and the alternative words in refined topics are in blue.

Topic 1: Finance
wealth billion fund private repay yuan *lcd mutual* shareholder refund wealth billion fund private repay yuan *budget investment* shareholder refund
Topic 2: Theater
review theater *elizabeth taylor* shakespeare *shore leo love gil dragon* review theater *performance director* shakespeare *play actor drama act stage*
Topic 3: Investment
georgia investor return *legacy* bank *society android editorial math critic* *credit* investor return *wealth* bank *economy stock investment money budget*
Topic 4: Medicine
treatment spread relief stress eye *pollution power* wound *shark tournament* treatment spread relief stress eye *infection illness* wound *doctor medicine*

NPMI values between each word w_{ij} and its topic peers $t_i \setminus \{w_{ij}\}$ before and after refinement, illustrating a visible shift towards higher values. Further, we list some base and refined topics from ECRTM (Topic 1 and 2) and LLM-TM (Topic 3 and 4) to show the changes before and after refinement. Table 6 shows that the Topic Refinement directly replaces the intruder words.

4 Conclusion

This paper explored the potential of using LLMs to enhance the topic modeling quality in short texts. Our proposed mechanism, Topic Refinement, utilizes prompt engineering with LLMs to identify and correct semantically intruder words within extracted topics. This process emulates human-like evaluation and refinement of topics to enhance the topic quality. Extensive experiments across the four datasets and nine base topic models conclusively demonstrated the effectiveness of this mechanism. Future research could involve the dynamic adjustment of the topic modeling process based on LLM feedback.

Acknowledgments. This work was supported by the National Natural Science Foundation of China under Grant No. 92467202 and No. 62102192, the Fellowship of China Postdoctoral Science Foundation under Grant No. 2022M710071, the Open Fund of Anhui Province Key Laboratory of Cyberspace Security Situation Awareness and Evaluation under Grant No. TK224013, and the Postgraduate Research and Practice Innovation Program of Jiangsu Province under Grant No. KYCX23_1077.

References

1. Blei, D.M., Ng, A.Y., Jordan, M.I.: Latent dirichlet allocation. JMLR **3**, 993–1022 (2003)
2. Brown, T.B., Mann, B., Ryder, N., et al.: Language models are few-shot learners. In: Proceedings of NeurIPS. vol. 33, pp. 1877–1901 (2020)
3. Cheng, X., Yan, X., Lan, Y., Guo, J.: BTM: topic modeling over short texts. IEEE TKDE **26**(12), 2928–2941 (2014)
4. Chowdhery, A., Narang, S., Devlin, J., et al.: Palm: Scaling language modeling with pathways. JMLR **24**(240), 1–113 (2023)
5. Fang, Z., He, Y., Procter, R.: CWTM: leveraging contextualized word embeddings from BERT for neural topic modeling. In: Proceedings of LREC-COLING. pp. 4273–4286 (2024)
6. Grootendorst, M.: Bertopic: Neural topic modeling with a class-based TF-IDF procedure (2022)
7. Mu, Y., Dong, C., Bontcheva, K., Song, X.: Large language models offer an alternative to the traditional approach of topic modelling. In: Proceedings of LREC-COLING. pp. 10160–10171 (2024)
8. Nguyen, T., Luu, A.T.: Contrastive learning for neural topic model. In: Proceedings of NeurIPS. vol. 34, pp. 11974–11986 (2021)
9. Pham, C., Hoyle, A., Sun, S., et al.: TopicGPT: A prompt-based topic modeling framework. In: Proceedings of NAACL. pp. 2956–2984 (2024)
10. Qiang, J., Qian, Z., Li, Y., et al.: Short text topic modeling techniques, applications, and performance: A survey. IEEE TKDE **34**(3), 1427–1445 (2022)
11. Vitale, D., Ferragina, P., Scaiella, U.: Classification of short texts by deploying topical annotations. In: Proceedings of ECIR. vol. 7224, pp. 376–387 (2012)
12. Wang, H., Prakash, N., Hoang, N., et al.: Prompting large language models for topic modeling. In: Proceedings of IEEE BigData. pp. 1236–1241 (2023)
13. Wang, Q., Zhu, C., Zhang, Y., et al.: Short text topic learning using heterogeneous information network. IEEE TKDE **35**(5), 5269–5281 (2023)
14. Wang, R., Hu, X., Zhou, D., et al.: Neural topic modeling with bidirectional adversarial training. In: Proceedings of ACL. pp. 340–350 (2020)
15. Wang, R., Ren, P., Liu, X., et al.: DCTM: dual contrastive topic model for identifiable topic extraction. IP&M **61**(5), 103785 (2024)
16. Wang, R., Zhou, D., Huang, H., Zhou, Y.: MIT: Mutual information topic model for diverse topic extraction. IEEE TNNLS **36**(2), 2523–2537 (2025)
17. Wu, X., Dong, X., Nguyen, T.T., Luu, A.T.: Effective neural topic modeling with embedding clustering regularization. In: Proceedings of ICML. vol. 202, pp. 37335–37357 (2023)
18. Wu, X., Luu, A.T., Dong, X.: Mitigating data sparsity for short text topic modeling by topic-semantic contrastive learning. In: Proceedings of EMNLP. pp. 2748–2760 (2022)
19. Yan, X., Guo, J., Lan, Y., Cheng, X.: A biterm topic model for short texts. In: Proceedings of WWW. pp. 1445–1456 (2013)
20. Yang, T., Hu, L., Shi, C., et al.: HGAT: heterogeneous graph attention networks for semi-supervised short text classification. ACM TOIS **39**(3), 32:1–32:29 (2021)
21. Yin, J., Wang, J.: A model-based approach for text clustering with outlier detection. In: Proceedings of ICDE. pp. 625–636 (2016)
22. Zhang, X., Zhao, J.J., LeCun, Y.: Character-level convolutional networks for text classification. In: Proceedings of NeurIPS. vol. 28, pp. 649–657 (2015)