Front. Comput. Sci., 2025, 19(2): 192325
https://doi.org/10.1007/s11704-024-40069-7

LETTER

Foundation models for topic modeling: a case study

Han ZENG[1,2], Jia-Ming SUN[1,2], Chun-Shu LI[1,2], Zhuying LI[1,2], Tong WEI (✉)[1,2]

1 School of Computer Science and Engineering, Southeast University, Nanjing 210096, China
2 Key Laboratory of Computer Network and Information Integration, Southeast University, Nanjing 210096, China

1 Introduction

In Natural Language Processing (NLP), topic modeling is a class of methods used to analyze and explore textual corpora, i.e., to discover the underlying topic structures from text and assign text pieces to different topics. In NLP, a *topic* means a set of relevant words appearing together in a particular pattern, representing some specific information. It is beneficial for tracking social media trends, constructing knowledge graphs, and analyzing writing styles.

Topic modeling has always been an area of extensive research in NLP. Traditional methods like Latent Semantic Analysis (LSA) and Latent Dirichlet Allocation (LDA), based on the "bag of words" (BoW) model, often fail to grasp the semantic nuances of the text, making them less effective in contexts involving polysemy or data noise, especially when the amount of data is small.

Recent advancements in Large Language Models (LLMs) have revolutionized this field. Thanks to the computing power improvements of hardware, various pre-trained foundation models like LLMs have sprung up. LLMs have a solid ability to understand semantic information and can generate fluent and consistent text with human instructional prompts.

With their profound semantic understanding, LLMs can fluently and consistently respond to human-like prompts, enhancing the efficiency and accuracy of topic modeling. Our work is closely related to TopicGPT [1], a recent prompt-based topic modeling framework. TopicGPT completes various topic modeling tasks by combining natural language instructions and input text, feeding them to LLMs, and then parsing the output of LLMs. Relying on the vast number of parameters in LLMs, LLM-based topic modeling methods can directly, efficiently, and accurately extract the text's topic, providing stronger interpretability.

However, after our experiments, we found that TopicGPT still has some imperfections: 1) in the original paper, only OpenAI's ChatGPT was used for experiments, and no other open-source LLM was used; 2) the prompts and datasets in the original paper are all in English, without good support for Chinese texts; 3) the datasets used in the original paper are Wiki and Bills, which contain Wikipedia articles and summaries of bills of the U.S. Congress, but not other types of text, such as *news*.

Our contributions are threefold:

- We used multiple LLMs for comparison, including the Qwen models deployed locally.
- We crawled data from the Internet to build a news dataset and designed Chinese prompts to test TopicGPT in the Chinese environment.
- We expanded the capabilities of TopicGPT and added support for third-level topic generation.

2 Methodology

This section briefly introduces how we use LLMs to accomplish topic modeling tasks. There are two main tasks: topic generation (Section 2.1) and topic assignment (Section 2.3), together with optional topic refinement (Section 2.2). Figure 1 gives an overview of our method.

2.1 Topic generation

The task of topic generation aims to generate topics given a set of documents (See Fig. 1(a)). Given a document d and some seed topics S, we iteratively prompt the LLMs to assign d to a topic in S or generate a new topic s for it and add it to S. The prompt contains the existing topics S, the current document d, and descriptive natural language instructions.

Document d can simultaneously subordinate to multiple existing or new topics in this process. Seed topics are set manually. The instructions in the prompt will require that the output of the large model include, in addition to the label of the topic (such as "trade"), a short description of the topic (such as "refers to the import and export amounts of capital, goods, and services.").

In addition to the direct first-level topic generation mentioned above, we can make the LLMs generate second-level topics subordinate to the first-level topics based on prompts (See Fig. 1(c)). Given a document set D and the first-level topic list S generated by D, the second-level topic generation process is as follows:

1. Select a certain first-level topic s and find its

Received January 13, 2024; accepted May 23, 2024

E-mail: weit@seu.edu.cn

Special Issue—Excellent Young Computer Scientists Vision on Foundation Models

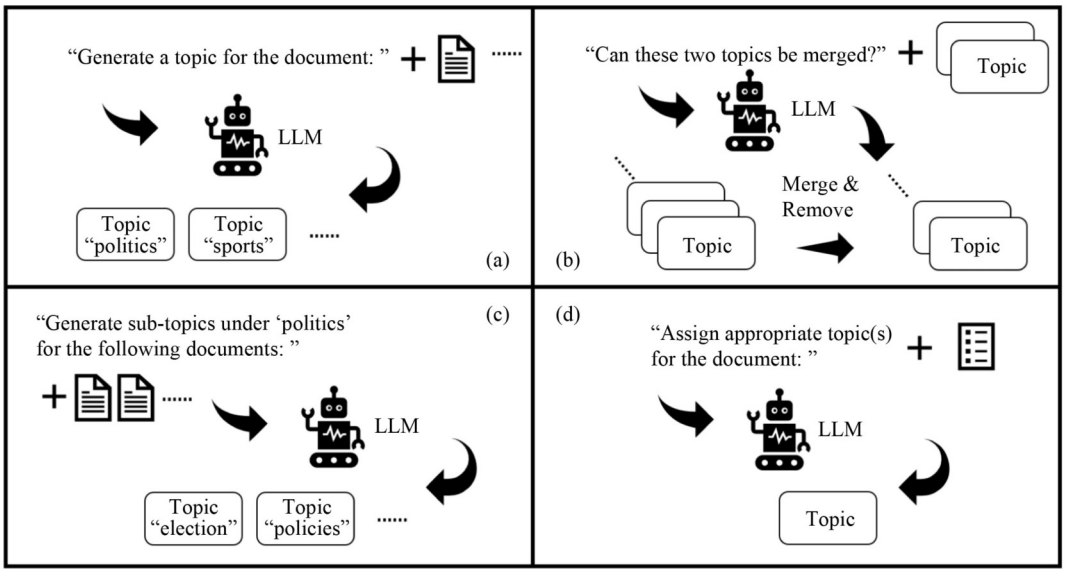

Fig. 1 Overview of our method. (a) First-level topic generation; (b) topic refinement; (c) second-level topic generation; (d) topic assignment

corresponding document subset D' in D.

2. Feed the prompt, D' together with the corresponding first-level topic into the LLMs to generate the second-level topics for the first-level topic.
3. Select another first-level topic from S and repeat the above steps.

In addition to the second-level topic label, the output also has a short description of the label. This methodology can be seamlessly extrapolated to facilitate the generation of third-level or more granular topic levels. This procedure entails treating the generated second-level topics as first-level topics and prompting the LLMs using the identical prompting strategy mentioned previously.

2.2 Topic refinement

Limited by the capabilities of LLMs, some problems may arise in the results of topic generation, e.g., 1) the granularity of generated topics is not uniform; 2) the generated topics contain very few corresponding documents. These problems may be alleviated by topic refinement, i.e., merging similar topics and removing rare topics (See Fig. 1(b)). The process is as follows:

1. The generated topic labels and descriptions are encoded through a pre-trained text embedding model, and the similarity between each topic is calculated.
2. Take the topic pair with the highest similarity and the corresponding prompt to the LLMs and ask whether the pair should be merged; if yes, output the merged topic label and the description, delete the original topic from the topic list and add the new topic. Repeat this step until the similarity between any two topics in the topic list is lower than a certain threshold.
3. Finally, delete topics from the topic list that correspond to too few documents.

Given the differences in LLMs' capabilities and the quality of generated topics, topic refinement is not necessary for well-generated topics.

2.3 Topic assignment

After obtaining the appropriate topic hierarchy from multiple documents, LLMs can be prompted to categorize new documents accordingly. This involves inputting the document, the established topic hierarchy, and instructions for topic assignment into the model (See Fig. 1(d)). The output labels the document with a relevant topic and provides the rationale for its categorization and a related quote from the text, enhancing interpretability.

3 Experiments

To effectively evaluate the performance of the LLMs for topic modeling, we created a dataset of Chinese news. Specifically, we crawled 219 news items of various types from United Daily News and manually labeled their respective topics. The distribution of the topics is shown in Fig. 2. 149 new items were used for topic generation, and 70 were used for topic assignment. As for LLMs, we used two aligned versions of Qwen language models [2] (Qwen-14B-Chat, Qwen-72B-Chat) and GPT-4 for comparison. We rewrote the prompts used in TopicGPT to make them suitable for Chinese contexts and to obtain better outputs. The following parts qualitatively and quantitatively demonstrate the results of the experiments. All detailed results can be found on our GitHub page.

3.1 Results on generation

GPT-4, Qwen-72B, and Qwen-14B generated 16, 41, and 98 first-level topics from 149 news documents. Among them,

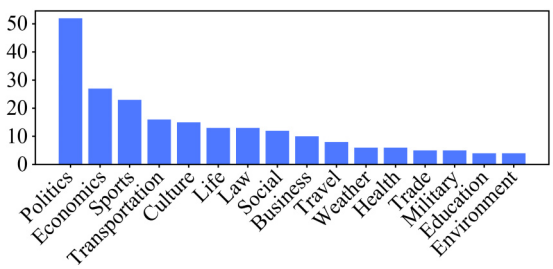

Fig. 2 Statistics over the news dataset

GPT-4 showed distinct, high-quality first-level topics from 149 news documents, demonstrating uniform granularity with minimal overlap. In contrast, the Qwen models, though effective, demonstrated a lower generation quality. To illustrate these findings, we have included a visualization of the topics generated, as depicted in Fig. 3.

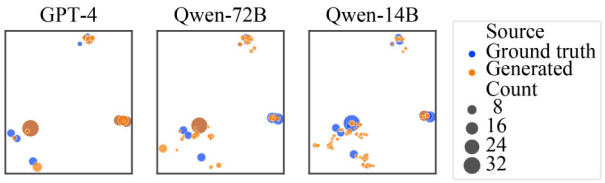

Fig. 3 Visualization of the generated topics of the three models. The size of each dot corresponds to the number of news documents associated with that particular topic

In Fig. 3, each generated topic was initially embedded into a 768-dimensional space and subsequently reduced to two dimensions for visualization. We used the GTE-base [3] model for embedding and the UMAP library [4] for dimensionality reduction. It can be seen that the GPT-4 model generated fewer topics, closely aligning with the ground truths; the Qwen-14B model generated more fragmented topics, scattering notably distant from the ground truths. The Qwen-72B model shows an intermediate performance between them.

In the topic refinement phase, 98 first-level topics were condensed into 18 for Qwen-14B. In contrast, Qwen-72B and GPT-4's output quality negated the need for further refinement. Similarly, Qwen-14B showed inconsistencies in second-level topic generation, while the others maintained better topic granularity control.

We also conduct exploratory experiments on third-level topic generation. Considering the size and quality of the data, we only ran third-level topic generation for news under the first-level "Politics" topic among the generation results by GPT-4. Detailed results can be accessed on our GitHub repository (See github.com/voyageofsean/llm-topic-modeling website). It can be seen that GPT-4 can still generate reasonable results on third-level topics.

3.2 Results on assignment

In Table 1, we present our experimental results on topic assignment, evaluated using three widely-used clustering metrics: Purity (P_1), Adjusted Rand Index (ARI), and Normalized Mutual Information (NMI). Purity quantifies the dominance of the most popular ground truth label within clusters, providing a measure of intra-cluster homogeneity. Yet, it does not cover scenarios where a single ground-truth label spans several clusters. Inverse Purity, defined in the opposite way of Purity, tackles this issue but overlooks the chance of blending all labels in one cluster. P_1 calculates the harmonic mean of Purity and Inverse Purity, ensuring balanced label representation. ARI evaluates the pairwise consistency between two cluster sets, adjusted for chance. Finally, NMI quantifies the similarity between actual and formed clusters using mutual information.

Table 1 Topical alignment between ground-truth labels, predicted assignments of the three models, and traditional topic modeling method LDA. Higher values indicate better performance

Model	P_1 ↑	ARI ↑	NMI ↑
GPT-4	76.20	65.44	83.77
Qwen-72B	72.76	54.83	81.90
Qwen-14B	55.61	19.94	73.41
LDA	52.28	13.55	52.28

It can be seen from Table 1 that GPT-4 assigns topics accurately and shows a good understanding of semantics in all three metrics. However, the results of Qwen-14B show no satisfaction, indicating its inability to understand limited documents and adhere strictly to prompt instructions. Traditional method LDA performs the worst since it treats sentences as BoW and cannot understand the semantics.

3.3 Summary

In summary, GPT-4 outperforms the other two LLMs throughout the entire process, primarily due to the limitations of the models' scale. Meanwhile, Qwen-72B performs considerably better than Qwen-14B and achieves comparable results, albeit slightly inferior, to those of GPT-4. We summarized our experiments in Table 2.

Table 2 Summary of three LLMs on multiple tasks

Phase	GPT-4	Qwen-72B	Qwen-14B
1st-level generation	Excellent, consistent granularity	Good, consistent granularity	Weak, inconsistent granularity
Refinement	(Not needed)	(Not needed)	Inconsistent granularity still exists
2nd-level generation	Excellent, slightly inconsistent granularity	Good, slightly inconsistent granularity	Poor, fragmented topics
Assignment	Excellent, correct assignments	Excellent, correct assignments	Poor, little understanding of prompts

Competing interests The authors declare that they have no competing interests or financial conflicts to disclose.

References

1. Pham C M, Hoyle A, Sun S, Iyyer M. Topicgpt: a prompt-based topic modeling framework. 2023, arXiv preprint arXiv: 2311.01449
2. Bai J, Bai S, Chu Y, et al. Qwen technical report. 2023, arXiv preprint arXiv: 2309.16609
3. Li Z, Zhang X, Zhang Y, Long D, Xie P, Zhang M. Towards general text embeddings with multi-stage contrastive learning. 2023, arXiv preprint arXiv: 2308.03281
4. McInnes L, Healy J, Saul N, Grossberger L. Umap: uniform manifold approximation and projection. The Journal of Open Source Software, 2018, 3(29): 861