Journal of Library and Information Studies 24:1 (June 2026) pp.17-40
https://doi.org/10.6182/jlis.202606_24(1).017

A Study of Automated Topic Labeling Based on Large Language Models

Sung-Chien Lin[1]

Abstract

This study proposes an automated topic labeling method based on large language models (LLMs), capable of generating meaningful term and summary labels for each topic within a topic model. Two generation strategies are introduced: Concise Descriptor Labeling (CDL) and Context-Enhanced Labeling (CEL). In addition to qualitative observation and conventional measures of stability and topical relevance, this study further evaluates coverage and discriminativeness through a topic assignment task. The experimental results show that CDL tends to produce concise and general disciplinary terms, characterized by high stability and close semantic alignment with the topic descriptors. In contrast, CEL often generates specialized technical terminology; although lexical variations may occur across multiple generations, semantic consistency remains high. Though CEL demonstrates slightly better performance in terms of coverage and discriminativeness, both strategies are acceptable and complementary, allowing researchers to select the appropriate approach depending on the interpretive and application context of the topic model. Future research may focus on refining these generation strategies and exploring their integration to enhance the applicability and interpretability of topic models.

Keywords: Topic Models; Large Language Models (LLMs); Automated Topic Labeling

1. Introduction

Latent Dirichlet Allocation (LDA) is a widely applied topic modeling method for analyzing large-scale text collections. Through LDA, researchers can transform textual content into a topic model consisting of two sets of parameters: the probability distributions of words within topics, and the topic proportions within documents. With these data, researchers are able to identify key terms of each topic and determine the topical composition of documents, which can be further applied to downstream Natural Language Processing (NLP) or information retrieval (IR) tasks. However, when LDA is used for text-based topic analysis or research trends exploration, the numerical output of the model often provides little support for users to intuitively and quickly understand the underlying meaning of topics (Alokaili et al., 2020). Assigning meaningful labels and explanations to topics not only helps users interpret topical features, but also facilitates subsequent applications.

Appropriate labels enable users to quickly grasp the meaning of topics (relevance), understand their scope (coverage), and distinguish them effectively from one another (discriminativeness; Mei et al., 2007). In practice,

[1] Department of Information and Communications, Shih-Hsin University, Taipei, Taiwan
E-mail: scl@mail.shu.edu.tw

researchers conducting topic modeling with LDA often rely on high-probability words within topics—known as topic descriptors—as the basis for labeling. For example, Maier et al. (2018) analyzed issues related to food safety and found that the top descriptors of a given topic included food, outbreak, salmonella, illness, report, and people (see Table 1). From this list, users could infer that the topic was related to foodborne diseases. However, using such descriptors directly as labels imposes high cognitive load, as it is difficult to memorize and inconvenient for analysis. Consequently, most LDA-based studies infer the semantic content of descriptors and select representative words or short phrases as topic labels.

This human-dependent labeling process, however, presents several challenges. First, in order to ensure comprehensibility and to capture the broader meaning of a topic, researchers may introduce terms beyond the original descriptors. For instance, Maier et al. (2018) labeled food safety topic as foodborne diseases, a phrase not included among the top ten descriptors (see Table 1). This illustrates that effective labeling requires substantial domain knowledge and subjective judgment (D. He et al., 2021), which can easily lead to inconsistency across annotators (Lau & Baldwin, 2016). Second, due to the inherent randomness of LDA (Chen et al., 2020; Chuang et al., 2015; Mantyla et al., 2018), descriptors of the same topic may lack semantic coherence, making interpretation difficult. Moreover, there are no clear standards regarding the number of descriptors to be used. In the food safety example, the top ten descriptors were mostly generic terms, which is insufficient for precise topic characterization. Adding more descriptors may provide detail but also introduce semantic noise, complicating interpretation. These challenges have motivated researchers to explore automated methods for topic labeling.

Based on the form of topic labels, automated methods can be broadly classified into three categories: term (or phrase) labels, sentence labels, and summary labels. These methods generally extract or generate candidate labels from the analyzed text or external resources (e.g., Wikipedia). A relevance score is then calculated for each candidate label based on its similarity to the corresponding topic, and the most relevant labels are ultimately selected. For example, Mei et al. (2007) used pointwise mutual information (PMI) to compute the semantic similarity between candidate terms and topic descriptors, thereby identifying suitable labels. Similarly, D. He et al.

Table 1. Top 20 Descriptors for the Food Safety Topic

Rank	Descriptor	Rank	Descriptor	Rank	Descriptor	Rank	Descriptor
1	food	6	people	11	disease	16	infection
2	outbreak	7	case	12	bacterium	17	investigation
3	salmonella	8	state	13	contaminate	18	sick
4	illness	9	eat	14	raw	19	chicken
5	report	10	ill	15	foodborne	20	egg

Note. Adapted from Maier et al. (2018).

(2021) employed a Doc2Vec embedding model to measure the semantic similarity between sentences and topics, selecting appropriate sentences and further composing them into summary labels to provide richer topical information.

Nevertheless, different forms of topic labels exhibit their own advantages and disadvantages, and no single form can meet all needs. Aletras et al. (2017) found that while lists of descriptors provide higher topical precision, term labels are more easily understood by users. Wan and Wang (2016) and D. He et al. (2021) also noted that although term labels are concise and easy to memorize, they may lack sufficient informational depth. By contrast, summary labels can capture topical meanings and scope more comprehensively, while also enhancing differentiation among topics. However, summaries impose a higher cognitive load, as users must read longer passages. Wan and Wang acknowledged that their method did not explicitly address the fluency of summaries and often make them more difficult to interpret.

To overcome the limitations of single-form labels, Chaudhary et al. (2024) proposed a method combining semantic networks and language models to generate multiple forms of labels—including term, sentence, and summary labels. The novelty of this approach lies in providing diverse label forms, enabling adaptation to different application levels. However, the method faces several challenges in practice, as it relies heavily on semantic networks, fine-tuning of language models, and customized algorithms, all of which demand substantial human and computational resources. Moreover, the range of generated labels is constrained by the vocabulary and relationships predefined in the semantic network. In addition, since the different label forms are generated independently, consistency may be insufficient, reducing users' holistic understanding of topic meanings. Although this method offers a promising direction for diversified topic labeling, issues regarding resource demands, adaptability, and label consistency remain to be addressed.

In response to these challenges, this study leverages the powerful text generation capabilities of large language models (LLMs) to simultaneously produce term labels and summary labels, thereby adapting to different application needs. LLMs excel at encoding contextual semantics and generating coherent natural language outputs, enabling them to produce labels that capture not only lexical overlap with topic descriptors but also deeper semantic abstractions of the original topics—all without task-specific training. This study employs pretrained LLMs that operate independently of semantic networks or other external resources and are not restricted to a specific LLM, which substantially lowers the technical barriers to application while reducing the human effort required for manual labeling and mitigating subjective inconsistencies. Rijcken et al. (2023) also explored the use of ChatGPT for automated topic labeling; however, their work remained at an exploratory stage, with prompts designed in a relatively simplistic manner and evaluation limited to a single domain expert's comparison with her own annotations. Consequently, the generated labels showed limited usefulness. Hence, this study investigates strategies to enhance both the quality and applicability of generated labels while maintaining strong relevance to topics.

Another challenge lies in the evaluation of topic labeling itself, which is inherently subjective and thus lacks a standardized assessment methodology. To better understand the practical utility of automated labeling methods, this study not only adopts conventional evaluation metrics but also introduces an innovative approach, that is, using LLMs to evaluate generated labels by integrating them with topic assignment tasks. This provides a novel, reasonable, and practical framework for assessing topic labeling methods.

The remainder of this paper is organized as follows. The sections "Related Work on Automated Topic Labeling" and "Related Work on Generative LLMs in Topic Modeling" review and synthesize relevant literature. The section "LLM-Based Topic Labeling Strategies and Evaluation" presents the strategies and evaluation methods proposed in this study, along with a series of experiments. "Experimental Results" reports the findings. Finally, the paper concludes with a summary and future directions.

2. Related Work on Automated Topic Labeling

To enhance the interpretability of topic modeling results, the development of effective automated topic labeling methods is essential. The pioneering work of Mei et al. (2007) laid a crucial foundation for this field by proposing a procedural framework that has since been adopted and extended by numerous studies. Their procedure consists of the following key steps:

(1) Selecting reference texts as sources for topic labels.
(2) Extracting candidate labels potentially representative of topics from the reference texts.
(3) Calculating a semantic relevance score between each candidate label and its corresponding topic.
(4) Generating the final topic labels based on the candidates and their relevance scores.
(5) Evaluating the quality of the generated topic labels.

This section follows the above framework and systematically reviews relevant studies for each step.

2.1 Label source selection

Research on term-based labeling has often focused on extracting candidate labels directly from the analyzed texts used to construct the topic models, as demonstrated by Kou et al. (2015) and Mei et al. (2007). However, some researchers argue that the analyzed texts may not always contain appropriate terms for labeling, therefore suggest using external resources with broad knowledge coverage. For example, Alokaili et al. (2020), Bhatia et al. (2016), and Lau et al. (2011) employed Wikipedia entry titles as candidate sources. Other studies have turned to structured knowledge bases, such as DBpedia (Hulpus et al., 2013) or the semantic network ConceptNet (Chaudhary et al., 2024), to capture semantically richer candidates.

By contrast, research on summary labels—for instance, Cano Basave et al. (2014), D. He et al. (2019, 2021), and Wan and Wang (2016)—has pointed out that even large external resources cannot adequately reflect real-time events or domain-specific information. Consequently, candidate sentences are directly extracted from the analyzed texts, leveraging contextual richness

to preserve more accurate topical semantics and thereby enhance descriptive power.

2.2 Candidate label extraction

When sentences or summaries serve as labels, candidate sentences can be directly selected from the text. For term-based labeling, however, appropriate terms must first be extracted. One common approach is to rely on chunk parsers to identify noun phrases based on syntactic information, as in Kou et al. (2015), Lau et al. (2011), and Mei et al. (2007). Alternatively, Mei et al. also proposed a statistical bigram extraction method that does not require syntactic information; in this method, bigrams were evaluated using t-tests, and statistically significant pairs were selected as candidate labels.

2.3 Relevance score computation

In early work on term-based relevance scoring, Mei et al. (2007) adopted statistical measures such as topic-specific word probabilities and PMI to evaluate the association between candidates and topics. With advances in machine learning and NLP, Lau et al. (2011) went beyond these statistical measures (PMI, *t*-test, Chi-square test) by integrating them into a support vector regression model, thereby improving labeling performance through supervised learning. Building on this approach, Bhatia et al. (2016) incorporated richer linguistic features, including letter trigram vectors, the PageRank scores of source documents, the length of candidate terms, and the frequency of overlap between candidates and topic descriptors. The use of letter trigram vectors traces back to Kou et al. (2015), who also explored word embeddings to capture deeper semantic relations between terms. Other researchers, such as Chaudhary et al. (2024) and Hulpus et al. (2013), examined structured knowledge bases and semantic networks as candidate sources. They employed graph-based algorithms—including centrality measures and maximal connected subgraphs—to compute relevance scores from the relational structure of terms.

For sentence-based labels, D. He et al. (2019) and Wan and Wang (2016) used Kullback-Leibler (KL) divergence between candidate sentences and topic descriptors, while D. He et al. (2021) further adopted Doc2Vec embeddings to refine sentence–topic similarity and improve accuracy.

2.4 Topic label generation

Most term-based methods select candidates with the highest relevance scores as labels. However, with the rapid rise of deep learning, Alokaili et al. (2020) applied sequence-to-sequence neural architectures to generate term labels. Trained on Wikipedia article content, their model generated concise and informative term labels by conditioning on both the topic descriptors and the keywords extracted from representative documents in which the topic was highly prevalent.

For summary labels, D. He et al. (2019, 2021) and Wan and Wang (2016) selected representative sentences from candidate pools to form topic summaries. Given length constraints, these methods had to balance semantic relevance with diversity, avoiding redundancy while maintaining inter-topic discrimination. To optimize sentence ranking, Wan and Wang employed submodular function maximization, while D. He et al. (2019, 2021) used graph-based methods such as Markov

chains and random walks. However, as extractive summarization approaches, they often produce fragmented summaries composed of loosely connected sentences, resulting in limited fluency and readability and potentially failing to capture nuanced thematic relationships (Supriyono et al., 2024).

Chaudhary et al. (2024) advanced this field by proposing a unified framework that generates term, sentence, and summary labels simultaneously using state-of-the-art techniques. For term labels, they selected high-scoring words from ConceptNet via graph-based methods. For sentence labels, they introduced the Graph2Corpus algorithm, transforming maximal connected subgraphs from ConceptNet into sentence corpora and selecting those that are most semantically aligned with representative documents. For summary labels, they combined the sentence corpus with representative documents and employed a pretrained language model to generate abstractive summaries. Compared with extractive approaches by D. He et al. (2019, 2021) and Wan and Wang (2016), their deep learning–based abstractive summaries provided more coherent and interpretable topic representations. However, the framework relies on multiple specialized linguistic resources and carefully designed graph-based algorithms, which increases methodological complexity and may limit scalability and broader adoption.

In summary, despite the growing body of work on topic label generation, existing studies largely focus on producing a single form of label, such as terms, sentences, or summaries, in isolation. Although Chaudhary et al. (2024) represent a notable exception by generating term, sentence, and summary labels within a unified framework, these label forms are still produced through separate generation processes, which may limit their semantic coherence across representations.

2.5 Evaluation of generated labels

The seminal work of Mei et al. (2007) not only established a procedural framework but also proposed evaluation criteria for topic labels:

- The *understandability* of topic labels to users,
- The *relevance* of topic labels to their corresponding topics,
- The *coverage* of topic labels in capturing the scope of topics, and
- The *discriminativeness* of topic labels across topics.

In practice, Mei et al. (2007) computed semantic relevance using word probabilities and PMI, while also addressing the challenge of labels highly relevant to multiple topics by proposing strategies to enhance the inter-topic discriminativeness of labels. They further recommended Maximal Marginal Relevance (MMR) to select multiple diverse terms, thereby improving coverage.

For term labels, many studies have relied on human evaluations, in which annotators rated the overall quality of labeling results produced by different methods and subsequently ranked the methods for comparison. This approach was used by Bhatia et al. (2016), Hulpus et al. (2013), Lau et al. (2011), and Mei et al. (2007). Kou et al. (2015) supplemented human judgment with WordNet-based comparisons against manually assigned labels, while Alokaili et al. (2020) applied BERTScore to automatically assess semantic relevance. Notably, most studies emphasize relevance and coverage, with

comparatively less attention to understandability or discriminativeness.

For summary labels, D. He et al. (2019) and Wan and Wang (2016) evaluated relevance, coverage, and discriminativeness through both human and automated assessments. Automated measures included KL divergence between topic and summary word distributions for relevance, coverage of the top 20 topic descriptors, and cosine similarity between summary labels of different topics for discriminativeness. D. He et al. (2021) advanced this by using Doc2Vec embeddings to capture contextual semantics for relevance and discriminativeness. However, none of these studies explicitly evaluated the understandability of summaries.

Finally, Chaudhary et al. (2024) conducted automated evaluations across all three label forms. They used BERTScore for term and sentence labels, KL divergence for summaries, and coverage measures for sentence and summary labels, assessing the proportion of top 20 topic descriptors captured. Nevertheless, despite these advances, ensuring labels are simultaneously understandable, relevant, comprehensive in coverage, and discriminative across topics remains a key challenge.

Overall, despite the availability of diverse evaluation methods, comprehensive evaluation remains underexplored. While Mei et al. (2007) identified four core quality dimensions for topic labeling—understandability, relevance, coverage, and discriminativeness—most subsequent work has focused primarily on semantic relevance, with limited and uneven attention to coverage, discriminativeness, and user-centered interpretability.

3. Related Work on Generative LLMs in Topic Modeling

3.1 Introduction to generative large language models

Generative LLMs are sequence-to-sequence language models based on the transformer architecture, capable of generating corresponding text from input text. These models are pre-trained on large-scale corpora that learn rich linguistic rules, syntactic structures, and semantic patterns, thereby acquiring advanced abilities in natural language understanding and generation. Building on this foundation, techniques such as instruction fine-tuning and reinforcement learning from human feedback (RLHF) have further enhanced their accuracy and adaptability in text generation.

The advantages of LLMs lie in their powerful language processing capacity and wide applicability. Users only need to provide task requirements and contextual information in prompts, then the model can generate highly relevant and fluent responses. Since LLMs inherently contain vast linguistic knowledge, they can often be applied to diverse NLP tasks without requiring additional domain-specific fine-tuning. For example, de Paoli (2024) employed generative LLMs to automatically extract keywords from texts, while Zhang et al. (2024) used LLMs to produce highly readable summaries. Whether commercial models (e.g., GPT, Gemini) or open-source ones (e.g., LLaMA, Mistral), generative LLMs have already demonstrated performance comparable to or even surpassing that of humans in text generation, machine translation, semantic analysis, and other NLP tasks.

With the rapid adoption of LLM technologies, their applications have expanded into the field of

topic modeling, offering new possibilities and solutions for text analysis. The following discusses two main directions in applying LLMs to topic modeling.

3.2 Using generative LLMs as topic modeling tools

Mu et al. (2024), Pham et al. (2023), and Wang et al. (2023) applied generative LLMs to topic modeling by guiding the models with prompts to produce topic labels that capture textual semantics. Despite these promising advances, these approaches still face notable challenges arising from inherent limitations of LLMs.

One major issue is the restriction on token length, which prevents feeding all documents into the model at once. When texts are processed in multiple batches, the generated outputs often include semantically similar yet lexically different labels—such as illness, sick, and disease in a food safety topic—resulting in redundancy and overlap. This outcome necessitates additional post-processing to consolidate frequent labels and merge similar ones into representative core topics.

Another challenge lies in the high computational costs associated with LLM inference. These costs often limit the analysis to partial subsets of the corpora rather than the full dataset, which risks omitting important topics and introduces potential bias in the results. Moreover, most study employ few-shot prompting strategies that provide only a small set of example labels to guide the LLMs. While such strategies can be effective in a certain case, they struggle to adapt to different requirements for topical granularity across tasks. As a result, multiple iterations of prompt design are often needed to achieve suitable labels, posing practical challenges for real-world application.

3.3 Post-processing applications with existing topic models

Other studies have investigated LLMs for automated evaluation or interpretation of topic models.

Stammbach et al. (2023) conducted two experiments: assessing topic coherence and determining the optimal number of topics. They compared LLM-based evaluations with the traditional Normalized Pointwise Mutual Information (NPMI) metric, and found that LLMs aligned more closely with human judgments. They also used LLMs to assign topics to documents, then computed label purity to estimate optimal topic numbers, showing that models with higher purity were generally more interpretable.

Reuter et al. (2024) developed GPTopic, a package based on the ChatGPT API to assist users in interpreting and adjusting topic models. The naming and explanatory functions of GPTopic are roughly equivalent to term-based and summary-based topic labeling. However, their paper primarily described system functions without detailing its implementation.

Rijcken et al. (2023) examined how LLMs could aid human interpretation of topic models. An expert compared her own labels with ChatGPT-generated summaries, evaluating whether automatically generated summaries improved interpretability. Results were mixed, showing that only about 50% of ChatGPT outputs were deemed useful by experts. While the authors acknowledged the small scale and uncertainty of ChatGPT, the study suggested that simply adding topic descriptors as prompts does not guarantee satisfactory results.

Kozlowski et al. (2024) compared three generative LLMs—FLAN, GPT-4o, and GPT-4 mini—

using four metrics: number of distinct labels, stability, similarity, and accuracy. The number of distinct labels measured redundancy, stability assessed consistency of outputs across runs, similarity was computed via cosine similarity, and accuracy was judged qualitatively. Results showed that although GPT-4o and GPT-4 mini were slightly less stable and similar than FLAN, they generated highly accurate labels with minimal redundancy, especially for three-word outputs, highlighting their practical usability.

Khandelwal (2025) proposed and compared four LLM-based topic labeling strategies differing in how topic-related documents were used as prompts. The first method selected documents with high overlap with topic descriptors; the second used term frequency–inverse document frequency (TF-IDF) to identify the most representative documents; the third and fourth created subtopics, with the third selecting from the largest subtopic and the fourth combining documents from multiple subtopics for broader coverage. Effectiveness was evaluated with Sentence-BERT, comparing label-document similarity. Experiments on the BBC News and 20 Newsgroups corpora showed that the third method yielded the best results.

4. LLM-based Topic Labeling Strategies and Evaluation

The purpose of this study is to explore how LLMs can be applied to the task of automated topic labeling for topic models in a practical and methodologically coherent manner. Although recent studies have demonstrated the potential of LLMs for topic modeling, directly using LLMs as topic modeling tools remains challenging due to token length limitations, sensitivity to prompt design, high computational costs, and the risk of generating excessive or overlapping topics. Accordingly, this study adopts LLMs specifically for topic labeling rather than topic discovery, combining LLMs with traditional topic models to achieve a more feasible and scalable solution.

Building on the limitations identified in prior work on automated topic labeling, this study proposes an LLM-based labeling framework that addresses several open challenges. First, term labels and summary labels are generated jointly to ensure semantic consistency across label forms. Second, prompt-based generation reduces reliance on complex algorithms and extensive domain-specific expertise. Third, summary labels are produced via LLM-based abstractive summarization, enabling more fluent and coherent topic descriptions than extractive approaches. Finally, a set of evaluation metrics combining semantic similarity and task-oriented assessments is adopted to evaluate relevance, coverage, discrimination, and stability, consistent with the evaluation principles articulated by Mei et al. (2007).

Unlike some prior studies that treat sentence labels as a distinct category (e.g., Chaudhary et al., 2024), sentence labels are not treated as a separate category in this study. Term labels are constrained to short phrases of one to three words to support rapid recognition and differentiation across topics, while summary labels are limited to approximately 30 words to provide concise natural-language explanations that facilitate interpretation. Given these length constraints and functional overlap, the distinction between

sentence labels and summary labels is not clear-cut. To avoid redundant experimentation and to maintain conceptual clarity, sentence labels are subsumed under summary labels, and the evaluation focuses exclusively on term labels and summary labels.

This section therefore introduces the proposed topic labeling strategies, prompt design and evaluation metrics, followed by a description of the experimental procedure, including the corpus, the topic modeling method, and the resulting topic model.

4.1 Topic labeling strategies

This study developed two topic labeling strategies—*Concise Descriptor Labeling* (CDL) and *Context-Enriched Labeling* (CEL)—whose designs stem from distinct theoretical perspectives on topic interpretation. CDL reflects the traditional view in topic modeling that the high-probability descriptors of a topic are sufficient to summarize its meaning. This approach prioritizes efficiency and surface-level interpretability. CEL, by contrast, draws on contextualist perspectives in NLP, which argue that incorporating representative document context can produce labels that better capture fine-grained semantics and technical specificity. These complementary perspectives motivated the design of the two strategies.

In practice, the two strategies differ in how the input is organized and how the LLM processes it. CDL uses only topic descriptors as input, allowing the model to generate both term and summary labels for all topics in a single prompt. This concise input makes batch processing feasible and highly efficient. CEL, on the other hand, augments descriptors with representative documents to enrich contextual information. Because the added document content greatly increases input length, the LLM generates labels for one topic at a time, with the descriptors of that topic and representative documents provided together. For a topic model containing K topics, CEL therefore requires K separate generation runs.

4.2 Prompt design

To implement these strategies, prompts were carefully designed to guide the generation of term and summary labels. Term labels consist of short phrases of one to three words, while summary labels are brief statements of approximately 30 words.

Every prompt consists of two components: a system instruction and a user input. The system instruction comprises the following elements:

(1) Introduction to LDA: A brief introduction to the basic concepts and applications of LDA topic models, defining key terms such as topic descriptors, representative documents, and topic labels.
(2) Text Collections: An overview of the corpus sources and text types analyzed in this study.
(3) Task: Instructions for the LLM to generate term and summary labels based on the user-provided topic model results, outlining the steps and precautions for each strategy.
(4) Input Format: Specification of the JSON input structure, which must include required fields such as topic ID and descriptors, with representative documents added if required by the strategy.
(5) Output Format: Specification of the JSON output structure, including topic ID and the corresponding term and summary labels for each topic.

The user input differs across strategies. In CDL, it consists of the descriptors of all topics entered together, whereas in CEL, it consists of the descriptors and representative documents of a single topic, provided one topic at a time.

4.3 Evaluation methods and metrics

To evaluate the two strategies, this study primarily adopts quantitative measures, supplemented with qualitative observations of label features. Traditional approaches typically rely on either qualitative assessments, which depend on expert judgment, or quantitative assessments, which involve objective comparisons (e.g., semantic similarity between labels and descriptors). To provide a more comprehensive and application-oriented evaluation, this study designs four quantitative metrics: *Topic Labeling Stability* (TLS), *Topic Relevance* (TR), *Topic Coverage* (TC), and *Topic Discriminativeness* (TD).

4.3.1 Topic Labeling Stability (TLS)

Since LLMs are probabilistic text generation models, repeated generations may yield different outputs even under identical inputs. To account for this inherent randomness, TLS is proposed to evaluate the consistency of term labels for the same topic across multiple generations under a fixed LLM, prompt, and generation setting. It captures intrinsic stochastic variation within a single configuration, rather than consistency across different model architectures, versions, or deployment platforms. To assess the stability of a labeling method, we therefore compare variations in term labels across repeated generations of the same topic model. Following Kozlowski et al. (2024), this study defines two indicators.

The first indicator is TLS_l *Lexical Stability*, with its calculation shown in Formula (1). In this formula, K is the number of topics in the topic model, I is the number of labeling experiments, t_{ik} represents the term label generated for the k-th topic in the i-th experiment, and $\left|\bigcup_{i=1}^{I}\{t_{ik}\}\right|$ represents the cardinality of the union of all term labels generated for topic k across the I runs, i.e., the total number of distinct term labels produced for topic k over repeated generations.

$$TLS_l = 1 - \frac{\sum_{k=1}^{K}(\left|\bigcup_{i=1}^{I}\{t_{ik}\}\right|-1)}{K(I-1)} \quad (1)$$

The second indicator is TLS_s *Semantic Stability*, as shown in Formula (2). It calculates the consistency of term labels at the semantic level across multiple generations. In the formula, all generated term labels are first transformed into semantic embeddings using an embedding function $E(\cdot)$, which maps textual inputs into a shared semantic vector space. The cosine similarity, $Cos(E(t_{ik}), E(t_{jk}))$ is then computed between the semantic embeddings of the k-th term labels generated in the i-th and j-th experiments. Finally, all pairwise similarity scores are averaged across topics and runs to obtain the overall semantic stability score.

$$TLS_s = \frac{\sum_{k=1}^{K}\sum_{i=1}^{I}\sum_{j=i+1}^{I} Cos(E(t_{ik}),\ E(t_{jk}))}{KI(I-1)/2} \quad (2)$$

The results of both indicators range between 0 and 1. The closer the value is to 1, the higher the stability of the labeling method across multiple generations.

4.3.2 Topic Relevance (TR)

Mei et al. (2007) argued that generated labels should maintain a high degree of semantic relevance to their corresponding topics. Since this study generates both term and summary labels, both forms are evaluated by comparing with the

topic descriptors and representative documents. The proposed TR metrics map all textual elements—including labels, topic descriptors, and representative documents—into embedding vectors in a shared semantic space and compute their semantic similarity. While these similarity-based measures do not aim to directly capture human interpretability or perceived meaningfulness, TR is intended to serve as a scalable proxy indicator for assessing the semantic alignment between generated labels and their underlying topics under controlled evaluation settings.

The formulas of TR are shown in (3)–(6). Here, K is the number of topics, N is the number of descriptors or representative documents within each topic, t_k and s_k denote the term label and the summary label for the k-th topic, and d_{kn} and r_{kn} denote the n-th descriptor and the n-th representative document of the k-th topic. Semantic relevance is computed by measuring cosine similarity between the corresponding embeddings of labels and topic-related texts.

$$TR_{TD} = \frac{\sum_{k=1}^{K}\sum_{n=1}^{N} Cos(E(t_k),\ E(d_{kn}))}{KN} \quad (3)$$

$$TR_{TR} = \frac{\sum_{k=1}^{K}\sum_{n=1}^{N} Cos(E(t_k),\ E(r_{kn}))}{KN} \quad (4)$$

$$TR_{SD} = \frac{\sum_{k=1}^{K}\sum_{n=1}^{N} Cos(E(s_k),\ E(d_{kn}))}{KN} \quad (5)$$

$$TR_{SR} = \frac{\sum_{k=1}^{K}\sum_{n=1}^{N} Cos(E(s_k),\ E(r_{kn}))}{KN} \quad (6)$$

All four indicators range between 0 and 1. Values closer to 1 indicate stronger semantic association between the generated labels and the original topics.

4.3.3 Topic Coverage (TC)

Previous studies often evaluated coverage by measuring the proportion of the top 20 descriptors of a topic that were included in the generated summary labels (Chaudhary et al., 2024; D. He et al., 2019; Wan & Wang, 2016). Such lexical-level matching offers a convenient proxy but fails to capture whether the labels truly cover the full semantic scope of the topic. A label may overlap lexically with several descriptors yet still miss essential aspects of the topic, especially when synonyms or context-specific terminology are involved.

To address this limitation, this study adopts topic assignment—the core downstream application of topic models—as the basis for evaluating coverage. Using LLMs, test documents are assigned to the most appropriate topics based solely on the generated term and summary labels, and we then verify whether the assigned topics align with the documents' dominant topics (i.e., those with the highest proportion in the topic distribution). While this evaluation relies on the language understanding capabilities of LLMs and does not directly measure human judgment, TC is intended as a task-oriented and scalable proxy for assessing whether generated labels sufficiently capture the semantic scope of topics under a consistent evaluation setting. If documents dominated by a given topic can be correctly reassigned to that topic based only on the labels, then these labels are considered to exhibit good coverage.

Recall is used as the coverage metric, and macro recall (the average recall across all topics) measures overall coverage. The formula is shown in (7), where K is the number of topics, TP_k is the number of dominant-topic documents correctly assigned by the LLM for the k-th topic, and FN_k is the number of dominant-topic documents not correctly assigned.

$$TC = \frac{1}{K}\sum_{k=1}^{K}\frac{TP_k}{TP_k+FN_k} \quad (7)$$

To reflect different application need, both Recall@1 (LLM assigns only one topic) and Recall@3 (LLM may assign up to three topics) are reported as *TC*@1 and *TC*@3 respectively.

4.3.4 Topic Discriminativeness (TD)

Good labels should not only cover their own topic comprehensively but also distinguish it clearly from others. Prior studies have typically assessed discriminativeness by calculating the cosine similarity between embeddings of labels (D. He et al., 2019, 2021; Wan & Wang, 2016). However, such approaches focus only on textual or semantic distance among labels, and do not directly reflect how well labels support practical topic separation in document classification.

Building on the topic-assignment framework introduced above, TD is evaluated by examining whether documents assigned to a given topic according to its labels are in fact dominated by other topics. A low rate of such misassignments indicates stronger discriminative power of the generated labels. Similar to TC, this metric leverages LLM-based document assignment as a task-driven proxy rather than a direct measure of human interpretability. By grounding discriminativeness in document-level topic separation, TD aims to assess whether labels differentiate topics meaningfully in practical usage scenarios under controlled evaluation conditions.

Precision is used as the TD metric, and macro precision (average precision across all topics) measures overall discriminativeness. The formula is shown in (8), where K is the number of topics, TP_k is the number of documents correctly assigned to the k-th topic, and FP_k is the number of documents incorrectly assigned to it. Higher TD values indicate that the labels provide sharper topical distinctions.

$$TD = \frac{1}{K}\sum_{k=1}^{K}\frac{TP_k}{TP_k+FP_k} \quad (8)$$

In summary, these four metrics provide a multi-faceted basis for evaluating automated topic labeling. TLS captures the stability and reproducibility of the LLM-based generation process across repeated runs, ensuring that the labels remain consistent under the same input conditions. TR focuses on the semantic relevance between the generated labels and the original topics. However, it primarily reflects embedding-based semantic proximity and therefore does not directly measure the interpretability or usefulness of the labels in downstream applications. By contrast, TC and TD emphasize practical utility by testing whether the labels can effectively guide accurate topic assignment in unseen documents, thereby capturing their capacity for both semantic coverage and discriminative power in task-oriented settings. Together, these metrics balance stability, semantic alignment, and practical classification performance, offering a more comprehensive assessment of automated topic labeling strategies.

4.4 Experimental environment and procedure

To evaluate the feasibility of the proposed LLM-based automated topic labeling method, and to compare the two generation strategies, this study first performed topic modeling, then input the resulting topic model data into the LLM to generate topic labels, and finally carried out comparison and evaluation. The details of each step are explained below.

This study used the publicly available Kaggle dataset "NIPS Conference Papers" as the corpus for topic modeling. The dataset consists of academic papers from the Neural Information Processing Systems (NIPS) conference and covers a wide range of research topics within machine learning and related fields. Although the experiments focus on a single academic corpus, the proposed topic labeling framework is not inherently domain-specific. Because it relies on topic descriptors, representative documents, and general-purpose LLMs, the approach is in principle applicable to corpora from other domains. The NIPS corpus was selected primarily for its substantial topical diversity, rich textual content, and its alignment with the authors' domain expertise, which together facilitate a focused examination of the topic labeling process in this initial study.

Each document corresponds to a single research paper in text format. For experimental purposes, author information was removed from the full text. Each document was then segmented into sentences and tokenized, with the first 300 words extracted. Standard preprocessing procedures were applied, including lemmatization and stop-word removal.

The processed texts were then fed into a self-developed LDA topic modeling program, with Gibbs Sampling employed for parameter inference. To determine the number of topics, models ranging from 5 to 100 topics were generated, and the model with the highest topic coherence was selected. The final model contained 15 topics. For each topic, the top 10 descriptors and the 10 representative documents with the highest dominant topic proportion (i.e., ranked 1–10) were extracted for subsequent labeling. To ensure independence between label generation and evaluation, additional documents ranked 11–20 by dominant topic proportion were reserved for the topic assignment experiments described later.

Topic label generation was performed using Google's Gemini 2.5 Flash API with fixed decoding parameters (temperature = 0, top_p = 0, top_k = 1) to reduce randomness. However, as noted by *Defeating Nondeterminism in LLM Inference* (H. He & Thinking Machines Lab, 2025), outputs can still vary due to nondeterminism in inference servers. To address this, five independent generation rounds were conducted for both CDL and CEL strategies, and the resulting labels were used as the basis for evaluation.

On this basis, both qualitative observation and quantitative comparison were carried out. The qualitative analysis focused on comparing the semantic features and stylistic expressions of the generated labels under different strategies. The quantitative analysis included the four evaluation metrics described earlier: TLS, TR, TC, and TD.

For TLS and TR evaluation, Sentence Transformers (Reimers & Gurevych, 2019) were used to perform semantic vectorization, with thext-cs-scibert (https://huggingface.co/morenolq/thext-cs-scibert), a model fine-tuned on computer science research papers (La Quatra & Cagliero, 2022), applied to enhance accuracy. Since label generation produces five sets of results, the four TR metrics (TR_{TD}, TR_{TR}, TR_{SD}, and TR_{SR}) were computed for each set, and the median values were reported as representative scores to ensure comparability and avoid skew from outliers.

For the evaluation of TC and TD, a topic assignment task was conducted. The test set

comprised 150 representative documents (10 per topic, ranked 11–20 by dominant topic proportion) that were excluded from the label generation stage. These documents ensured strong topical association while avoiding overlap with training inputs. In each round, the Gemini 2.5 Flash API, with the same fixed decoding parameters (temperature = 0, top_p = 0, top_k = 1), assigned the most appropriate topic to each document using the generated term and summary labels. Recall and precision were then computed to assess coverage (TC) and discriminativeness (TD). As with TR, the median values across five runs were reported as representative scores.

5. Experimental Results

The topic model used for the topic labeling experiments is shown in Table 2. In addition to listing the top 10 descriptors for each topic, the table also presents example term labels generated by the two strategies (CDL and CEL).

5.1 Qualitative observation

This section qualitatively compares the labels generated by CDL and CEL, which exhibit clear stylistic and semantic differences. The illustrative examples examine both term labels and summary labels to reveal their complementary roles and semantic alignment across labeling strategies, as well as how CDL and CEL differ at both the term and summary levels.

In Topic 1, the associated descriptors include mixture, posterior, and nonparametric, which together indicate a focus on statistical inference within generative modeling. Based solely on these descriptors, CDL generates the term label "Generative Models," a broad disciplinary label that captures the general modeling paradigm but omits explicit methodological cues. This abstraction is also evident in the corresponding CDL summary label: "This topic focuses on generative models, particularly those using mixture models and latent variables. It involves techniques for density estimation, likelihood maximization, variational inference, and nonparametric Bayesian methods to model complex data distributions." While this summary provides a coherent overview of generative modeling, it emphasizes general modeling components and inference techniques without clearly foregrounding a specific methodological perspective.

In contrast, CEL produces the term label "Bayesian Nonparametric Inference," directly echoing key descriptors such as posterior and nonparametric and foregrounding the statistical inference perspective of the topic. This focus is consistently reinforced at the summary level, as reflected in the CEL summary label: "This topic covers advanced Bayesian and nonparametric inference methods for complex probabilistic models. It includes techniques like variational inference and MCMC for estimating latent variables and densities." Together, the CEL term and summary labels form a coherent semantic pair, with the term label serving as a concise methodological identifier and the summary label providing a complementary explanation that elaborates on the same conceptual focus.

A similar but more concise pattern is observed for Topic 10. CDL assigns the term label "Deep Neural Networks" with a summary emphasizing general neural architectures and learning mechanisms, thereby capturing the model class

Table 2. Topic Descriptors and Term Labels Generated by the Two LLM-based Labeling Strategies

Topic ID	Descriptors	CDL label	CEL label
1	mixture, density, likelihood, variational, latent, posterior, conditional, maximum, generative, nonparametric	Generative Models	Bayesian Nonparametric Inference
2	accuracy, year, attention, significant, amount, stream, quality, despite, advance, become	Performance Advances	ML System Performance
3	distribute, parallel, computation, analog, implementation, implement, circuit, sensor, communication, power	Distributed Hardware	Neurocomputing Hardware
4	spike, response, activity, brain, cell, stimulus, population, sensory, cortex, fire	Neural Activity	Neural Sensory Processing
5	sparse, observation, nonlinear, source, measurement, filter, independent, covariance, PCA, noisy	Signal Processing	Component Analysis & Signal Recovery
6	kernel, classifier, generalization, supervise, selection, support, margin, semi, positive, unlabeled	Kernel Methods	Classifier Generalization
7	convex, loss, rank, regression, bind, risk, norm, minimization, regularization, constraint	Convex Optimization	Regularized Convex Optimization
8	gradient, convergence, descent, rate, update, step, iteration, converge, batch, objective	Gradient Descent	Accelerated Gradient Descent
9	decision, policy, action, reinforcement, agent, reward, game, regret, expert, plan	Reinforcement Learning	Reinforcement Learning & Bandits
10	deep, layer, speech, hide, unit, recurrent, architecture, net, back, forward	Deep Neural Networks	Neural Speech Models
11	memory, rule, synaptic, dynamical, self, connection, capacity, differential, change, delay	Neural Dynamics	Associative Memory Dynamics
12	graph, tree, graphical, nod, item, belief, social, user, causal, product	Graphical Models	Belief Propagation & Graphs
13	cluster, distance, similarity, metric, group, dimensionality, embed, manifold, spectral, projection	Clustering & Manifolds	Spectral Manifold Learning
14	detection, motion, vision, visual, scene, face, segmentation, video, pose, shape	Computer Vision	Visual Object Understanding
15	word, language, topic, document, search, query, text, active, instance, category	Natural Language Processing	Text Topic Modeling

Note. For each topic, the top 10 descriptors identified by the LDA model are listed. Each topic is associated with a single term label produced by the Concise Descriptor Labeling (CDL) strategy and a single term label produced by the Context-Enriched Labeling (CEL) strategy.

but leaving the speech-related application implicit. In contrast, CEL produces the term label "Neural Speech Models" and a corresponding summary that explicitly situates neural architectures within the context of speech recognition, reinforcing alignment between term-level specificity and summary-level elaboration.

Overall, the two strategies differ in semantic orientation. CDL tends toward generality and accessibility, suitable for contexts requiring quick identification of topic domains, but risks blurring topic boundaries. CEL, by contrast, produces term labels that are more technical and precise, faithfully reflecting descriptors and highlighting topic differences, though potentially narrower in scope.

5.2 Quantitative comparison

The performance of both strategies under various quantitative metrics is presented in Table 3. Unless otherwise specified, the reported TR, TC, and TD results are computed using both term labels and summary labels, as defined in the Section "LLM-Based Topic Labeling Strategies and Evaluation."

5.2.1 Topic Labeling Stability (TLS)

TLS consists of two indicators, TLS_L and TLS_S. For CDL, the generated labels were completely consistent across the five runs, yielding perfect scores of 1.0 for both indicators, demonstrating very high stability. In contrast, CEL showed some variation across topics. For example, Topic 1 produced "Bayesian Nonparametric Inference" in three runs and "Approximate Bayesian Inference" in two runs. Similarly, Topic 14 alternated between "Object and Scene Analysis" and "Visual Object Understanding." Such variations reduced its lexical consistency score (TLS_L = 0.8333). However, since these variants were semantically close, CEL still achieved a high semantic consistency score (TLS_S = 0.9851).

5.2.2 Topic Relevance (TR)

TR includes four indicators: TR_{TD} (Term labels vs. Descriptors), TR_{TR} (Term labels vs. Representative documents), TR_{SD} (Summary labels vs. Descriptors), and TR_{SR} (Summary labels vs. Representative documents). As shown in Table 3, CDL outperforms CEL in TR_{TD} and

Table 3. Measurement Results of Evaluation Metrics for the Two Strategies

Metric	Indicator	CDL	CEL
TLS	TLS_L	1.0000	0.8333
	TLS_S	1.0000	0.9851
TR	TR_{TD}	0.9409	0.9001
	TR_{TR}	0.8327	0.8151
	TR_{SD}	0.8139	0.8139
	TR_{SR}	0.9812	0.9825
TC	$TC@1$	0.7200	0.7800
	$TC@3$	0.9733	0.9800
TD	TD	0.7452	0.7964

TR_{TR}, indicating stronger alignment of term labels with surface-level semantics. For instance, the descriptors of Topic 13 include cluster, distance, similarity, and manifold. CDL's label "Clustering & Manifolds" directly reflects these terms, whereas CEL's "Spectral Manifold Learning" highlights only a technical aspect, thus reducing surface alignment. Similarly, the descriptors of Topic 14 span detection, motion, vision, segmentation, and video. CDL's "Computer Vision" broadly encompasses these terms, while CEL's "Visual Object Understanding" narrows the scope to objects.

For TR_{SD} and TR_{SR}, both strategies perform almost identically, with TR_{SR} values approaching 1.0. This indicates that regardless of strategy, the generated summary labels maintain strong semantic consistency with representative documents. These findings are consistent with Wan and Wang (2016) and D. He et al. (2021), who argued that summary labels capture the broader semantic scope of topics.

5.2.3 Topic Coverage (TC)

TC is evaluated using recall when assigning either a single topic (*TC*@1) or three topics (*TC*@3). In the single-topic setting (*TC*@1), CEL achieves slightly higher recall than CDL, with both strategies exceeding 0.8. This suggests that although single-topic assignments are inherently difficult when documents often involve multiple topics, the proposed approaches still provide reasonably strong coverage. The performance gap between CEL and CDL may be attributable to their respective characteristics. CDL's broader labels tend to capture diverse contexts, whereas CEL's more precise labels align more closely with technical foci. For example, CEL's "Neural Sensory Processing" for Topic 4 better reflects representative documents on place cells, visual encoding, and spatial navigation than CDL's broader "Neural Activity." Conversely, CDL's "Kernel Methods" for Topic 6 encompasses SVM, PAC-Bayes, feature construction, and universal kernels, while CEL's "Classifier Generalization" highlights the central concept but omits breadth. When multiple assignments are permitted (*TC*@3), both strategies achieve scores near 1.0, indicating robust topic coverage.

5.2.4 Topic Discrimination (TD)

In TD, CEL performs slightly better than CDL, though differences are modest. This relates to the high semantic relatedness of certain topic groups. For example, Topic 1 focuses on statistical inference and generative modeling, and Topic 12 pays attention to graph structures and reasoning. Since the former often relies on the latter as formal expression, both CDL ("Generative Models," "Graphical Models") and CEL ("Bayesian Nonparametric Inference," "Belief Propagation & Graphs") are prone to confusion. A similar situation also occurs in Topics 7, 8, and 6: Topic 7 ("Convex Optimization" / "Regularized Convex Optimization") provides the theoretical foundation, Topic 8 ("Gradient Descent" / "Accelerated Gradient Descent") represents its core algorithms, and Topic 6 ("Kernel Methods" / "Classifier Generalization") corresponds to related applications. Their hierarchical relationship complicates discrimination. Nevertheless, CEL's labels are generally more precise. For example, in neural system-related topics, CEL's "Neural Sensory Processing" (Topic 4) and "Associative Memory Dynamics" (Topic 11) better capture representative document foci than

CDL's broader "Neural Activity" and "Neural Dynamics." This precision accounts for CEL's slight advantage in TD.

5.3 Section summary

Synthesizing the qualitative and quantitative analyses, CDL and CEL each demonstrate distinct yet complementary advantages. CDL produces short, general noun phrases that are highly stable (TL_S = 1.0) and better aligned with surface-level semantics (TR_{TD}, TR_{TR}), thereby offering broad semantic coverage. However, this breadth can blur topic boundaries, leading to lower performance in coverage ($TC@1$) and discriminativeness (TD).

By contrast, CEL generates more technical terms. Despite its less consistent lexical level (lower TLS_L), it maintains strong semantic consistency (TLS_S). Its use of representative document context sharpens focus on technical elements, yielding superior performance in both coverage and discriminativeness.

Overall, the two strategies highlight different application values. CDL emphasizes breadth and stability, make it suitable for non-specialist audiences seeking quick topic overviews; whereas CEL emphasizes precision and technical detail, providing fine-grained distinctions and accuracy for expert use. These findings suggest that LLM-based labeling need not pursue a single "best" solution, but should flexibly adopt strategies according to research and application contexts.

6. Conclusions

This study proposes an innovative and practical method for automated topic labeling in topic models, and demonstrates its effectiveness through empirical experiments. The method leverages LLM-based text generation, requiring neither prior training nor customized linguistic resources or programs. It can simultaneously produce semantically consistent, fluent, and interpretable labels in the forms of both terms and summaries. To this end, two generation strategies were designed: Concise Descriptor Labeling (CDL) and Context-Enriched Labeling (CEL). Both were evaluated comprehensively through qualitative observation and quantitative comparison using metrics including TLS, TR, TC, and TD.

The experimental results reveal that CDL typically generates general disciplinary terms in the form of single noun phrases. These labels are highly stable and maintain strong surface-level semantic connections. CEL, by contrast, generates compound technical or professional terms more often. Although lexical forms may vary across runs, semantic consistency remains high. Both strategies yield labels that are strongly aligned with the original topic descriptors and representative documents, confirming that the method faithfully captures the semantic meaning of topics. In topic assignment experiments, CEL performed better overall, though both strategies demonstrated acceptable levels of coverage and discriminativeness. Collectively, these findings confirm the practicality of the proposed approach that LLMs can efficiently generate meaningful labels in both term and summary form. Moreover, the two strategies complement each other, enabling researchers to flexibly choose the appropriate approach depending on application needs, thereby enhancing the interpretability and subsequent analysis of topic models.

This study makes three contributions. First, it introduced an LLM-based approach to automated topic labeling that requires no task-specific training or external linguistic resources. Compared with prior studies such as Chaudhary et al. (2024), this approach is more straightforward to apply. Moreover, unlike Rijcken et al. (2023), whose prompts are relatively narrow in scope, the present study develops a more structured prompt design that incorporates explicit explanations of topic modeling concepts, input/output formats, and task instructions, thereby improving transparency and reproducibility.

Second, it proposes and evaluates two generation strategies—CDL and CEL. Drawing on both established practices in topic modeling and recent advances in LLM prompting, the two strategies were systematically compared through qualitative observation and multiple quantitative metrics (TLS, TR, TC, TD). The results indicate that each has distinct advantages, providing alternative options suitable for different application needs.

Third, it develops an evaluation framework based on topic assignment. While topic relevance is assessed through semantic similarity between labels and textual elements, topic coverage and discriminativeness are evaluated by assigning topics to unseen documents. This framework moves beyond lexical-level comparisons often used in earlier studies, and provides a closer approximation to the practical use of topic labels in real text analysis.

Building on these contributions, future research may proceed in several directions. First, applying the proposed framework to corpora from broader domains, genres, and languages would enable a more comprehensive evaluation of its generalizability and robustness beyond a single academic dataset. Second, while this study examines label stability under fixed LLM configurations, future work could systematically investigate the effects of different LLM architectures, versions, and deployment settings on label generation, thereby further assessing the reliability of LLM-based topic labeling. Third, refinements in prompt design merit are worth continuing exploration, including the selection and organization of descriptors and representative documents, as well as the use of more advanced prompting strategies to improve stability and semantic alignment. Finally, future studies could extend the evaluation framework by incorporating human judgments or hybrid human–LLM assessments, in order to better examine interpretability and practical usefulness beyond semantic similarity and task-based performance alone.

References

Aletras, N., Baldwin, T., Lau, J. H., & Stevenson, M. (2017). Evaluating topic representations for exploring document collections. *Journal of the Association for Information Science & Technology, 68*(1), 154-167. https://doi.org/10.1002/asi.23574

Alokaili, A., Aletras, N., & Stevenson, M. (2020). Automatic generation of topic labels. In *SIGIR'20: Proceedings of the 43rd International ACM SIGIR Conference on Research and Development in Information Retrieval* (pp. 1965-1968). https://doi.org/10.1145/3397271.3401185

Bhatia, S., Lau, J. H., & Baldwin, T. (2016). Automatic labelling of topics with neural embeddings. In Y. Matsumoto & R. Prasad (Eds.), *Proceedings of COLING 2016, the 26th international conference on computational linguistics: Technical papers* (pp. 953-963). The COLING 2016 Organizing Committee. https://aclanthology.org/C16-1091/

Cano Basave, A. E., He, Y., & Xu, R. (2014). Automatic labelling of topic models learned from Twitter by summarisation. In K. Toutanova & H. Wu (Eds.), *Proceedings of the 52nd annual meeting of the Association for Computational Linguistics (Volume 2: Short Papers)* (pp. 618-624). Association of Computational Linguistics. https://doi.org/10.3115/v1/P14-2101

Chaudhary, A., Milios, E., & Rajabi, E. (2024). Top2Label: Explainable zero shot topic labelling using knowledge graphs. *Expert Systems with Applications, 242*, Article 122676. https://doi.org/10.1016/j.eswa.2023.122676

Chen, S., Andrienko, N., Andrienko, G., Adilova, L., Barlet, J., Kindermann, J., Nguyen, P. H., Thonnard, O., & Turkay, C. (2020). LDA ensembles for interactive exploration and categorization of behaviors. *IEEE Transactions on Visualization & Computer Graphics, 26*(9), 2775-2792. https://doi.org/10.1109/TVCG.2019.2904069

Chuang, J., Roberts, M. E., Stewart, B. M., Weiss, R., Tingley, D., Grimmer, J., & Heer, J. (2015). TopicCheck: Interactive alignment for assessing topic model stability. In R. Mihalcea, J. Chai, & A. Sarkar (Eds.), *Proceedings of the 2015 conference of the North American Chapter of the Association for Computational Linguistics: Human language technologies* (pp. 175-184). Association for Computational Linguistics. https://doi.org/10.3115/v1/N15-1018

de Paoli, S. (2024). Performing an inductive thematic analysis of semi-structured interviews with a large language model: An exploration and provocation on the limits of the approach. *Social Science Computer Review, 42*(4), 997-1019. https://doi.org/10.1177/08944393231220483

He, D., Ren, Y., Khattak, A. M., Liu, X., Tao, S., & Gao, W. (2021). Automatic topic labeling using graph-based pre-trained neural embedding. *Neurocomputing, 463*, 596-608. https://doi.org/10.1016/j.neucom.2021.08.078

He, D., Wang, M., Khattak, A. M., Zhang, L., & Gao, W. (2019). Automatic labeling of topic models using graph-based ranking. *IEEE Access, 7*, 131593-131608. https://doi.org/10.1109/ACCESS.2019.2940516

He, H., & Thinking Machines Lab. (2025, September 10). Defeating nondeterminism in LLM inference. *Thinking Machines Lab: Connectionism*. https://doi.org/10.64434/tml.20250910

Hulpus, I., Hayes, C., Karnstedt, M., & Greene, D. (2013). Unsupervised graph-based topic labelling using DBpedia. In *WSDM'13: Proceedings of the Sixth ACM*

International Conference on Web Search and Data Mining (pp. 465-474). https://doi.org/10.1145/2433396.2433454

Khandelwal, T. (2025). *Using LLM-based approaches to enhance and automate topic labeling*. arXiv. https://doi.org/10.48550/arXiv.2502.18469

Kou, W., Li, F., & Baldwin, T. (2015). Automatic labelling of topic models using word vectors and letter trigram vectors. In G. Zuccon, S. Geva, H. Joho, F. Scholer, A. Sun, & P. Zhang (Eds.), *Information retrieval technology (AIRS 2015)* (pp. 253-264). Springer. https://doi.org/10.1007/978-3-319-28940-3_20

Kozlowski, D., Pradier, C., & Benz, P. (2024). *Generative AI for automatic topic labelling*. arXiv. https://doi.org/10.48550/arXiv.2408.07003

La Quatra, M., & Cagliero, L. (2022). Transformer-based highlights extraction from scientific papers. *Knowledge-Based Systems, 252*, Article 109382. https://doi.org/10.1016/j.knosys.2022.109382

Lau, J. H., & Baldwin, T. (2016). The sensitivity of topic coherence evaluation to topic cardinality. In *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies* (pp. 483-487). Association for Computational Linguistics. https://doi.org/10.18653/v1/N16-1057

Lau, J. H., Grieser, K., Newman, D., & Baldwin, T. (2011). Automatic labelling of topic models. In *HLT'11: Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies* (Vol. 1, pp. 1536-1545). https://dl.acm.org/doi/10.5555/2002472.2002658

Maier, D., Waldherr, A., Miltner, P., Wiedemann, G., Niekler, A., Keinert, A., Pfetsch, B., Heyer, G., Reber, U., Häussler, T., Schmid-Petri, H., & Adam, S. (2018). Applying LDA topic modeling in communication research: Toward a valid and reliable methodology. *Communication Methods & Measures, 12*(2/3), 93-118. https://doi.org/10.1080/19312458.2018.1430754

Mantyla, M. V., Claes, M., & Farooq, U. (2018). Measuring LDA topic stability from clusters of replicated runs. In *ESEM'18: Proceedings of the 12th ACM/IEEE International Symposium on Empirical Software Engineering and Measurement* (Article 49). Association for Computational Linguistics. https://doi.org/10.1145/3239235.3267435

Mei, Q., Shen, X., & Zhai, C. (2007). Automatic labeling of multinomial topic models. In *KDD'07: Proceedings of the 13th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining* (pp. 490-499). Association for Computational Linguistics. https://doi.org/10.1145/1281192.1281246

Mu, Y., Dong, C., Bontcheva, K., & Song, X. (2024). Large language models offer an alternative to the traditional approach of topic modelling. In N. Calzolari, M.-Y. Kan, V. Hoste, A. Lenci, S. Sakti, & N. Xue (Eds.), *Proceedings of the 2024 joint international*

conference on computational linguistics, language resources and evaluation (*LREC-COLING 2024*) (pp. 10160-10171). ELRA Language Resources Association, International Committee on Computational Linguistics. https://aclanthology.org/2024.lrec-main.887/

Pham, C. M., Hoyle, A., Sun, S., Resnik, P., & Iyyer, M. (2023). *TopicGPT: A prompt-based topic modeling framework*. arXiv. https://doi.org/10.48550/arXiv.2311.01449

Reimers, N., & Gurevych, I. (2019). *Sentence-BERT: Sentence embeddings using Siamese BERT-networks*. arXiv. https://doi.org/10.48550/arXiv.1908.10084

Reuter, A., Thielmann, A., Weisser, C., Fischer, S., & Säfken, B. (2024). *GPTopic: Dynamic and interactive topic representations*. arXiv. https://doi.org/10.48550/arXiv.2403.03628

Rijcken, E., Scheepers, F., Zervanou, K., Spruit, M., Mosteiro, P., & Kaymak, U. (2023, August 20–24). *Towards interpreting topic models with ChatGPT* [Conference paper]. The 20th World Congress of the International Fuzzy Systems Association, Daegu, Republic of Korea. https://research.tue.nl/en/publications/towards-interpreting-topic-models-with-chatgpt/

Stammbach, D., Zouhar, V., Hoyle, A., Sachan, M., & Ash, E. (2023). Revisiting automated topic model evaluation with large language models. In H. Bouamor, J. Pino, & K. Bali (Eds.), *Proceedings of the 2023 conference on empirical methods in natural language processing* (pp. 9348-9357). Association for Computational Linguistics. https://doi.org/10.18653/v1/2023.emnlp-main.581

Supriyono, Wibawa, A. P., Suyono, & Kurniawan, F. (2024). A survey of text summarization: Techniques, evaluation and challenges. *Natural Language Processing Journal, 7*, Article 100070. https://doi.org/10.1016/j.nlp.2024.100070

Wan, X., & Wang, T. (2016). Automatic labeling of topic models using text summaries. In K. Erk & N. A. Smith (Eds.), *Proceedings of the 54th annual meeting of the Association for Computational Linguistics (Volume 1: Long Papers)* (pp. 2297-2305). Association for Computational Linguistics. https://doi.org/10.18653/v1/P16-1217

Wang, H., Prakash, N., Hoang, N. K., Hee, M. S., Naseem, U., & Lee, R. K.-W. (2023). Prompting large language models for topic modeling. In *2023 IEEE International Conference on Big Data* (*BigData*) (pp. 1236-1241). Institute of Electrical and Electronics Engineers. https://doi.org/10.1109/BigData59044.2023.10386113

Zhang, T., Ladhak, F., Durmus, E., Liang, P., McKeown, K., & Hashimoto, T. B. (2024). Benchmarking large language models for news summarization. *Transactions of the Association for Computational Linguistics, 12*, 39-57. https://doi.org/10.1162/tacl_a_00632

(Received: 2025/9/29; Accepted: 2026/1/6)

圖書資訊學刊　第24卷第1期　（2026.6）　頁17-40
https://doi.org/10.6182/jlis.202606_24(1).017

基於大型語言模型的自動化主題標示研究

A Study of Automated Topic Labeling Based on Large Language Models

林頌堅[1]

Sung-Chien Lin[1]

摘　要

本研究提出一種基於大型語言模型（LLM）的自動化主題標示方法，能夠為主題模型中的每個主題產生有意義的詞語和摘要標籤，並提出兩種生成策略：簡要描述語標示（CDL）和上下文增強標示（CEL）。除了對生成結果進行質性觀察和傳統的穩定性和主題相關性測量外，還以主題指派任務評估涵蓋性和區別性。實驗結果顯示，CDL傾向產生簡明的學科術語，具有高穩定性並相當接近描述語的語意；CEL則常生成專業術語，儘管多次生成的詞彙有些微變化，語意仍保有相當高的一致性。雖然CEL在涵蓋性與區分性上表現較優，但兩種策略的結果都可接受且具互補性，並可根據應用情境選擇適合策略。未來研究建議著重生成策略的進一步優化和整合，增強主題模型的理解與應用。

關鍵字：主題模型、大型語言模型、自動化主題標示

[1] 世新大學資訊傳播學系
Department of Information and Communications, Shih-Hsin University, Taipei, Taiwan
E-mail: scl@mail.shu.edu.tw

註：本中文摘要由作者提供。
以APA格式引用本文：Lin, S.-C. (2026). A study of automated topic labeling based on large language models. *Journal of Library and Information Studies, 24*(1), 17-40. https://doi.org/10.6182/jlis.202606_24(1).017
以Chicago格式引用本文：Lin, Sung-Chien. “A Study of Automated Topic Labeling Based on Large Language Models.” *Journal of Library and Information Studies* 24, no. 1 (2026): 17-40. https://doi.org/10.6182/jlis. 202606_24(1).017