https://doi.org/10.51408/issi2025_123

Automatic Literature Review Generation by Integrating Large and Small Models

Xiaofei Li[1], Guo Chen[2]

[1]358246618@qq.com,[2]delphi1987@qq.com
School of Economics & Management, Nanjing University of Science & Technology, Nanjing (China)

Abstract

This study proposes an innovative method for the automatic literature review generation, specifically designed for small-scale literature corpus in niche domains. First, the proposed method enhances the BERTopic with LLM to extract macro-level topic information. It then leverages LLM-based instruction fine-tuning to identify micro-level move structures. Finally, the method employs systematic LLM fine-tuning and a template-based generation strategy to automatically generate review that are thematically clear and logically coherent. Experimental results demonstrate that this method generates high-quality, well-structured review texts with clear topics when applied to small-scale citation analysis corpus, offering a new reference and practical example for automatic literature review generation in specialized fields.

Introduction

The development of automatic literature review generation techniques has significantly enhanced researchers' ability to efficiently acquire knowledge by synthesizing concise review texts from thematically related studies (Asmussen & Møller 2019). Current approaches to automatic literature review generation can be broadly classified into two categories: (1) extractive and generative methods based on small-scale deep learning models and (2) natural language generation methods leveraging LLMs. The first category integrates extractive and generative techniques, exemplified by the method proposed by Vaishali et al.(Vaishali et al., 2024), which employs an improved TextRank algorithm to extract key sentences from multiple documents, followed by a Seq2Seq model for review generation. While these methods are effective, they heavily depend on large-scale corpora and often struggle with maintaining consistency between the generated content and the original text. The second category, represented by retrieval-augmented generation approaches such as the one introduced by Han et al. (Han et al., 2024), incorporates relevant literature as an external knowledge source to enhance LLM-based review generation. Although these methods achieve superior fluency and logical coherence, they remain constrained by the inherent limitations of LLMs, including restricted context window sizes, outdated knowledge representations, and susceptibility to generating "hallucinated" information (Wang et al., 2024). Additionally, existing methods struggle to balance accuracy and comprehensiveness when processing small-scale literature corpus that are continuously updated in niche domains. To address these challenges, this paper proposes a novel hybrid framework that integrates small models with LLMs for automatic scientific review generation. The proposed approach leverages LLMs' strengths in language understanding and knowledge

synthesis while incorporating topic model to uncover latent topics, enabling high-precision review tailored to small-scale niche literature corpus.

Methodology

The proposed automatic review generation framework consists of three steps (as shown in Figure 1). First, a BERTopic model enhanced by LLM is used to identify the macro-level topic component of the document set. At the same time, the LLM is instruction-tuned to generate the micro-level move component of the documents. Finally, the LLM integrate two components based on a predefined template to generate a review. Each step is detailed as follows.

LLM-Enhanced BERTopic Zero-Shot Topic Modeling

Topic models are effective tool for extracting topic information from document set, providing a macro-level perspective for literature review. However, both traditional topic models such as LDA, Top2Vec, and BERTopic and LLMs face challenges when applied to the small-scale niche literature corpus. These challenges include sparse topic distributions due to the limited number of documents and high semantic overlap, which makes it difficult to distinguish topics. Moreover, these models struggle to capture deep topic relationships between documents, leading to unclear or even distorted topic results.

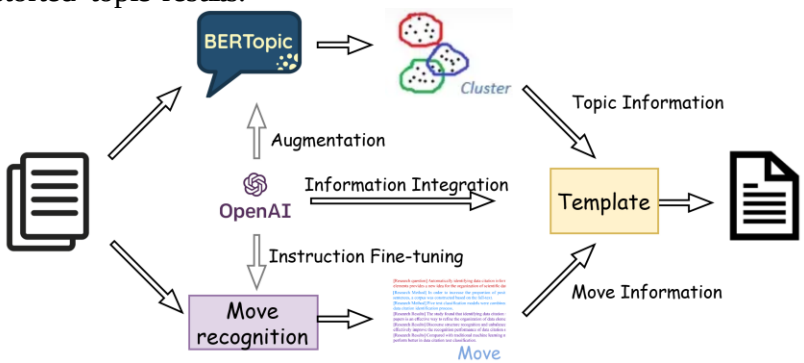

Figure 1. Framework of the Automatic Literature Review Generation Method.

To address these issues, this study proposes an LLM-enhanced zero-shot BERTopic modeling approach. This method integrates LLMs into three stages of topic modeling: enhancing document topic representation in the text embedding phase, assisting topic identification in the modeling phase, and refining topic distribution and representation in the post-modeling phase. This approach improves the overall performance of topic model.

In the text embedding phase, traditional word embedding models often fail to adequately capture intricate topic relationships between documents when processing small-scale literature collections in specialized domains. This limitation subsequently undermines the performance of topic model. To address this issue, this study leverages LLM to enhance document topic representation (as illustrated in Figure 2). The proposed methodology consists of two key stages: First, LLMs are utilized to generate high-quality topical phrases, tags, and descriptions from raw

document content, thereby extracting critical topic information. Second, distinct embedding representations are created for both the topic information and the original documents. These representations are subsequently fused via vector concatenation, thus creating a unified document embedding that highlights thematic features. This method improves the effectiveness of topic model by integrating topic-focused information with the original semantic content.

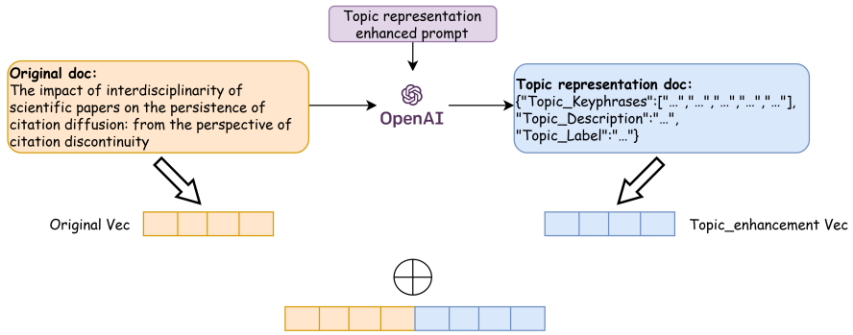

Figure 2. Method for Enhancing document Topic Representation.

In the topic identification phase, this study departs from the conventional approach of mining topics from scratch and instead leverages LLM-generated prior knowledge to guide BERTopic's zero-shot topic modeling (as illustrated in Figure 3). Specifically, LLMs are employed to extract salient topics from the document set, which are subsequently used as zero-shot topics for BERTopic. Following this, we calculate the similarity between each document and the zero-shot topics, then apply hierarchical processing based on similarity threshold: documents with similarity scores exceeding the threshold are directly assigned to their corresponding topics, while the remaining documents undergo further topic identification via BERTopic. This layered strategy effectively integrates the prior knowledge provided by LLMs with the adaptability of BERTopic, thereby improving the accuracy of topic recognition while ensuring broad coverage of topic distribution.

During the experiments, we observed that BERTopic's results exhibited loose topic representations and ambiguous boundaries. To address this issue, this study integrates LLMs in the post-modeling phase for refining topic representations and adjusting distributions. First, we utilize LLM to generate semantically compact and coherent topic labels based on the original outputs, thereby replacing BERTopic's native topic representations. Subsequently, by leveraging LLMs' zero-shot classification capability, we reassign the topic affiliations of boundary documents using the optimized topic labels as classification criteria. This two-step optimization strategy enhances both the accuracy and consistency of the final results.

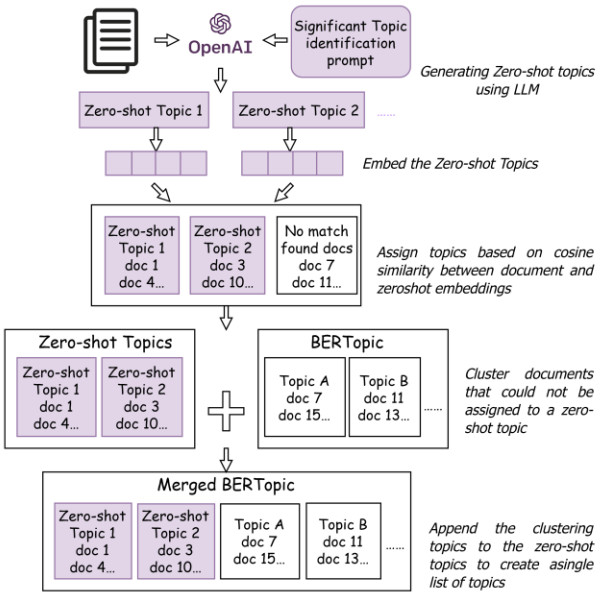

Figure 3. LLM-Guided BERTopic Zero-Shot Topic Modeling.

By integrating LLM into the entire workflow of BERTopic, our method achieves that the generated topics are not only semantically coherent and compact but also robust against noise and ambiguity in small-scale corpus.

Instruction fine-tuning LLM for Move Recognition

Move recognition effectively deconstructs sentence-level knowledge units in scientific literature, providing structured knowledge—such as research problems, methods, and results—that is essential for comprehensive review.

This study proposes an move recognition method that combines In-Context Learning (Agarwal et al., 2024) and Chain-of-Thought (Wei et al., 2022) LLMs prompting techniques, leveraging instruction-tuned LLM to accurately extract three types of knowledge units from abstracts: research problems, methods, and results. As illustrated in Figure 4, the prompt design for this method consists of four key modules: role setting and task description: guides the model to define its role and construct tasks based on instructions; Chain-of-Though: offers guided reasoning steps to help the model establish a clear logical chain during move recognition; In-Context Learning: provides examples of move recognition; and input integration: presents abstracts of scientific literature as input.

To comprehensively evaluate the performance of LLM in move recognition, we developed a systematic validation framework. The evaluation dataset is derived from our research group's previous move recognition projects, which include high-quality human-annotated data (Chen & Xu, 2019). Recognizing that human-annotated moves and LLM-generated moves may differ in wording but maintaining semantic equivalence, we introduced BERTScore(Zhang et al. 2020), a deep semantic matching metric, to effectively assess the semantic consistency of LLM-generated rhetorical moves.

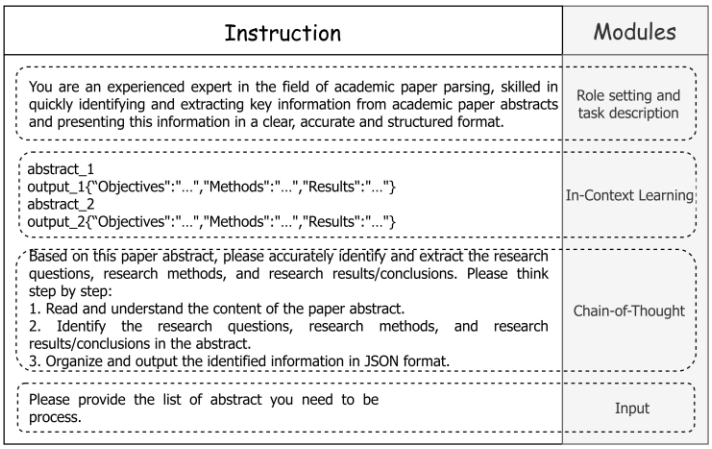

Figure 4. Move Recognition Prompt Design.

The comparative experimental results, summarized in Table 1, demonstrate the superior performance of instruction-tuned LLM in the move recognition task. Incorporating advanced prompt engineering strategies, the LLM-generated functional sentences exhibit significantly higher semantic consistency with human-annotated sentences. This enhanced performance establishes a reliable technical foundation for the task of automatic review generation.

Table 1. Comparative Results of Move Recognition.

Model	*Precision*	*Recall*	*F1-score*
SVM	0.834	0.834	0.834
CNN	0.839	0.838	0.839
Bi-LSTM	0.846	0.845	0.846
LLM	0.848	0.865	**0.853**

Template-Based Automatic Literature Review Generation with LLM

Through topic modeling and move recognition, we extracted both the macro-level topics and the micro-level move structures from the literature corpus. Subsequently, it is necessary to further investigate the internal connections between papers and construct a concise yet in-depth review. LLMs possess robust text generation and semantic integration capabilities. With instruction tuning, they can be effectively customized for literature review tasks. To enhance the effectiveness of LLMs in analysing topic connections and integrating moves, this study proposes a "Topic-Move" review template (see Figure 5) to standardize input data. Based on this, the automated review generation process consists of two stages. First, during the preprocessing stage, we organize the results of topic modeling and move recognition into a standardized and hierarchical format to ensure structured input. Then, in the generation stage, we employ chain-of-thought prompting combined with a modular generation strategy, completing the review in three steps: (1) feeding text segments with the same move under the same topic into the LLM to generate a move-level summary, (2) aggregating all move-level summaries under the same topic to produce

a topic-level summary, and (3) synthesizing all topic-level summaries to generate the complete review text.

In summary, the proposed automatic literature review method combines the strengths of LLMs and small models while overcoming their respective limitations. First, we leveraged the text comprehension and generation capabilities of LLMs to enhance topic representation and identify moves, thereby compensating for the semantic understanding shortcomings of small models when processing niche literature corpora. Second, the computational efficiency of small models is utilized for topic modeling, structuring raw literature into topic-level and move-level knowledge units — this dual approach reduces computational burdens on LLMs, thereby suppresses their hallucination tendencies. Furthermore, the framework adopts a phased generation architecture (move-topic-full-text) with modular strategies, effectively circumventing the context window constraints of LLMs.

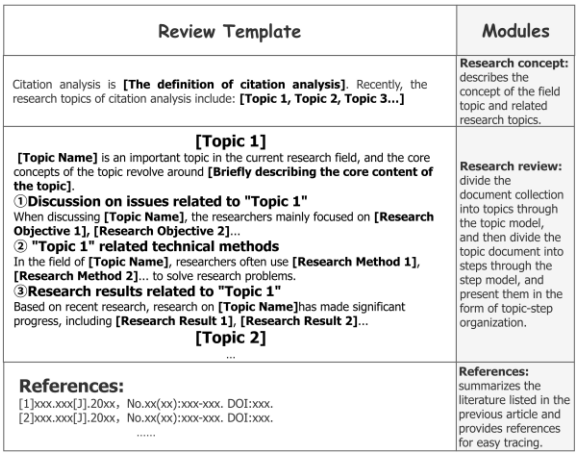

Figure 5. Template for Automatic Literature Review Generation.

Experiment – A Case Study in Citation Analysis

To evaluate the effectiveness of the proposed method, this study selected 24 papers in the field of "citation analysis" published in SSCI and CSSCI journals between September and December 2024 as experimental samples.

First, we utilized LLMs to enhance the topic representation of the original documents. (All LLMs used in this study were accessed via the OpenAI GPT-4 API). By inputting the titles, keywords, and abstracts of the papers, the LLM generated topical phrases, tags, and detailed descriptions. Subsequently, the model conducted a preliminary identification of significant topics within the dataset, recognizing three prominent topics (see Table 2). These identified topics were then employed as prior knowledge for zero-shot topic modeling using BERTopic. The preliminary modeling results (Figure 6) revealed one outlier topic, one zero-shot topic, and three topics derived through BERTopic.

Table 2. Significant Topics and Representations Identified by LLM (Partial Display).

Topic	*Topic Words*
Topic 0	['Impact', 'Measurement', 'Performance', 'Influence', 'Evaluation', …]
Topic 1	['Advanced Methods', 'Analytics', 'Novel Techniques', …]
Topic 2	['Data Management', 'Open Citation Data', 'Information Retrieval', …]

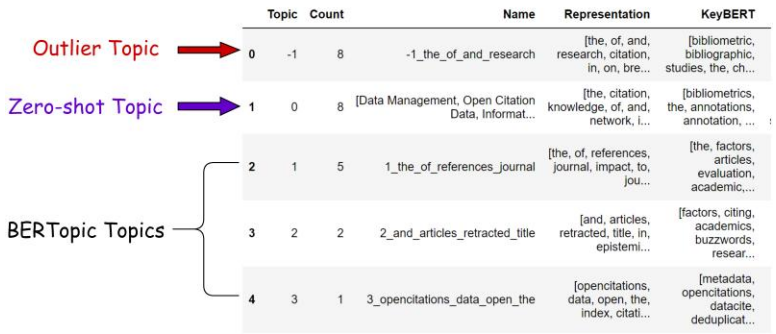

Figure 6. Preliminary Topic Modeling Results.

Following this, the LLM was employed to refine the topic representations and distributions. Specifically, the experiment integrated the topic keywords extracted by KeyBERT with the three most confident documents from each topic cluster into the LLM to generate more precise topic labels. Subsequently, documents with confidence scores below 0.8 or classified as outlier topics were treated as pending classification samples, using the optimized topic labels as classification labels. Finally, we utilized LLM fine-tuned with chain-of-thought reasoning to reclassify these samples, thereby enhancing the accuracy and consistency of topic division. The final optimized topic distributions and labels are presented in Table 3.

Table 3. Final Topic Modeling Results.

Topic	*Label*	*Count*
Topic 0	Citation diffusion	5
Topic 1	Citation measurement and prediction	8
Topic 2	Citation data management	5
Topic 3	Citation analysis across disciplines	4
Other	\	2

After completing the topic modeling, the study performed move recognition on the abstracts within each topic. Subsequently, the categorized move-specific texts for each topic were fed into the LLM for semantic integration and summarization, resulting in the generation of topic-based literature review modules. Finally, these modular reviews were synthesized into a comprehensive review following a "Topic–Move" structure. Figure 7 displays the input data and the LLM-generated literature review content for the "research problem" move under the "Citation Diffusion" topic. The results indicate that the generated literature review texts exhibit a clear focus on topics and strong coherence among documents within each topic. They effectively capture the research focal points and logical relationships in the literature, thereby

validating the effectiveness of the proposed methods in producing structured, high-quality literature reviews.

Template Filling Draft	Literature Review
1.1 Research issues related to "citation diffusion" Recent research issues on citation diffusion include the following: Study the role of citation range in developed economies and emerging powers on academic impact (especially in the social sciences), develop new citation range measurement methods, and explore the role of these ranges in improving academic impact. [1] Reveal the complex citation relationship and basic characteristics of retracted papers in molecular biology, including temporal and spatial patterns, reasons and publishers. [2] Analyze the similarities and differences in the knowledge base and diffusion trajectories of IPSUMGD and LDA to determine whether there are general patterns of successful innovation diffusion. [3] Study how policy information (policy design, topics and attributes) affects policy adoption. [4] Address the problem that existing methods are not suitable for measuring the preference for popular papers, and quantify this preference by calculating the popularity obtained in the references of papers. [5]	1.1 Research issues related to "citation diffusion" At present, citation diffusion research mainly focuses on identifying the knowledge base and its dissemination trajectory, quantifying the preference for popular papers, exploring the role of citation scope on academic impact and its measurement, the impact of policy information on policy adoption, and analyzing the citation relationship of retracted papers [1][2][3][4][5]. Through comprehensive discussion, these research issues are centered on measuring and understanding the role of citations in academic dissemination and influence, and provide a variety of methods for analyzing citations from diversified perspectives (such as innovation success, policy adoption, the impact of retracted papers, etc.) [1][3][5]. These studies are of great significance to the optimization of scientific research impact assessment methods, for example, exploring the academic impact of social sciences through new citation scope measurement methods [3].

Figure 7. Overview of Automatic Literature Review Generation Content.

Discussion

This study introduces an innovative method for automatic literature review generation that integrates both large and small models. It demonstrates unique advantages and significant potential in scenarios involving small-scale niche literature corpus. The proposed approach can be widely applied to periodic review tasks in niche fields, enhancing the efficiency of researchers in tracking the dynamic trends within their areas of study.

References

Agarwal, R., Singh, A., et al. 2024. Many-shot in-context learning. arXiv.

Asmussen, Claus Boye, and Charles Møller. 2019. "Smart Literature Review: A Practical Topic Modelling Approach to Exploratory Literature Review." *Journal of Big Data* 6 (1): 93.

Chen, Guo, and Tianxiang Xu. 2019. "Sentence Function Recognition Based on Active Learning." *Data Analysis and Knowledge Discovery* 3 (8): 53–61.

Han, Binglan, Teo Susnjak, and Anuradha Mathrani. 2024. "Automating Systematic Literature Reviews with Retrieval-Augmented Generation: A Comprehensive Overview." *Applied Sciences* 14 (19): 9103.

Vaishali, Ginni Sehgal, and Prashant Dixit. 2024. "A Comprehensive Study of Automatic Text Summarization Techniques." International *Conference on Emerging Innovations and Advanced Computing, Sonipat*, India 2024.

Wang, Yidong, Qi Guo, et al. 2024. "AutoSurvey: Large Language Models Can Automatically Write Surveys." arXiv.

Wei, Jason, Xuezhi Wan, et al. 2022. "Chain-of-Thought Prompting Elicits Reasoning in L arge Language Models." *Advances in Neural Information Processing Systems* 35

Zhang, Tianyi, Varsha Kishore, et al. 2020. "BERTScore: Evaluating Text Generation with BERT." arXiv.