Instruct Large Language Models to Generate Scientific Literature Survey Step by Step

Yuxuan Lai[1,2](✉), Yupeng Wu[3], Yidan Wang[1], Wenpeng Hu[4], and Chen Zheng[1,2]

[1] The Open University of China, Beijing, China
laiyx@ouchn.edu.cn, erutan@pku.org.cn
[2] Engineering Research Center of Integration and Application of Digital Learning Technology, Ministry of Education, Beijing, China
[3] FloatMiracle, Beijing, China
[4] PLA Academy of Military Science, Beijing, China

Abstract. Automatically generating scientific literature surveys is a valuable task that can significantly enhance research efficiency. However, the diverse and complex nature of information within a literature survey poses substantial challenges for generative models. In this paper, we design a series of prompts to systematically leverage large language models (LLMs), enabling the creation of comprehensive literature surveys through a step-by-step approach. Specifically, we design prompts to guide LLMs to sequentially generate the title, abstract, hierarchical headings, and the main content of the literature survey. We argue that this design enables the generation of the headings from a high-level perspective. During the content generation process, this design effectively harnesses relevant information while minimizing costs by restricting the length of both input and output content in LLM queries. Our implementation with Qwen-long achieved third place in the NLPCC 2024 *Scientific Literature Survey Generation* evaluation task, with an overall score only 0.03% lower than the second-place team. Additionally, our soft heading recall is 95.84%, the second best among the submissions. Thanks to the efficient prompt design and the low cost of the Qwen-long API, our method reduces the expense for generating each literature survey to 0.1 RMB, enhancing the practical value of our method.

Keywords: Automatic literature survey · Large language model · Prompt design

1 Introduction

Conducting a literature survey is a crucial component of scientific research [1]. However, as the volume of scientific publications continues to grow exponentially and the rate of new publications accelerates [2], it costs the researchers increasing

D. F. Wong et al. (Eds.): NLPCC 2024, LNAI 15363, pp. 484–496, 2025.
https://doi.org/10.1007/978-981-97-9443-0_43

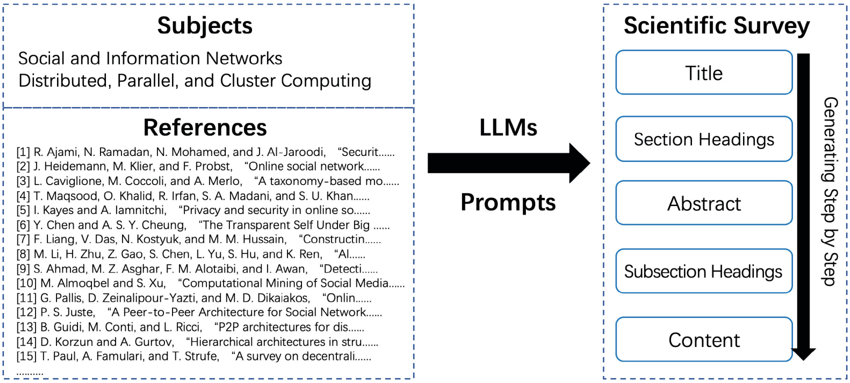
Subjects
Social and Information Networks
Distributed, Parallel, and Cluster Computing
References
[1] R. Ajami, N. Ramadan, N. Mohamed, and J. Al-Jaroodi, "Securit......
[2] J. Heidemann, M. Klier, and F. Probst, "Online social network......
[3] L. Caviglione, M. Coccoli, and A. Merlo, "A taxonomy-based mo......
[4] T. Maqsood, O. Khalid, R. Irfan, S. A. Madani, and S. U. Khan......
[5] I. Kayes and A. Iamnitchi, "Privacy and security in online so......
[6] Y. Chen and A. S. Y. Cheung, "The Transparent Self Under Big ......
[7] F. Liang, V. Das, N. Kostyuk, and M. M. Hussain, "Constructin......
[8] M. Li, H. Zhu, Z. Gao, S. Chen, L. Yu, S. Hu, and K. Ren, "AI......
[9] S. Ahmad, M. Z. Asghar, F. M. Alotaibi, and I. Awan, "Detecti......
[10] M. Almoqbel and S. Xu, "Computational Mining of Social Media......
[11] G. Pallis, D. Zeinalipour-Yazti, and M. D. Dikaiakos, "Onlin......
[12] P. S. Juste, "A Peer-to-Peer Architecture for Social Network......
[13] B. Guidi, M. Conti, and L. Ricci, "P2P architectures for dis......
[14] D. Korzun and A. Gurtov, "Hierarchical architectures in stru......
[15] T. Paul, A. Famulari, and T. Strufe, "A survey on decentrali......
....
LLMs
Prompts
Scientific Survey
Title
Section Headings
Abstract
Subsection Headings
Content
Generating Step by Step

Fig. 1. An illustration of the scientific survey generation task, and our step-by-step generation approach.

amounts of time on this task. With the advanced capabilities of large language models (LLMs) [3–5], AI-generated literature surveys offer a promising solution by significantly reducing the time and effort required.

However, generating scientific literature surveys presents several challenges. Firstly, these surveys are typically very lengthy, and even advanced models like GPT-4 [3] and Claude 3 [6] are constrained by a 4k-8k token output limit, making it impossible to generate a survey in a single pass. Attempting to generate the entire content sequentially based on all previous content can also be problematic, as it may cause the LLM to forget earlier information or fail to maintain a consistent structure without a comprehensive plan. Secondly, scientific literature surveys cover a vast amount of information. Transmitting this information to LLMs and generating a coherent survey can be very expensive in terms of LLM API usage, which diminishes its practical value.

In this paper, we propose a novel approach to scientific literature survey generation that leverages LLMs in a step-by-step prompting manner. As shown in Fig. 1, given the subjects and several reference papers, we design a series of prompts that guide LLMs to sequentially generate the title, abstract, hierarchical headings, and the main content of a literature survey. By breaking down the task into manageable steps, LLMs can maintain a high-level perspective while generating headings, thereby improving the coherence and relevance of the generated surveys. Additionally, after generating the headings, the main content generation can be conditioned on this structure rather than on all previous content, significantly reducing the cost of API usage.

We implemented our system using the Qwen-long version of Alibaba's Tongyi Qianwen model [5]. Our system was submitted in the NLPCC 2024 *Scientific Literature Survey Generation* task, where it achieved third place, with an overall score of 61.11, just 0.03% behind the second-place team. Additionally, our method demonstrated a soft heading recall of 95.84%, the second highest among all submissions and only 1.17% lower than the first-place team. These results highlight the effectiveness of our approach in maintaining logical and coherent survey structures, which can be attributed to the plan-then-generate strategy employed in our sequential generation process. Thanks to the efficient prompt design and the low cost of the Qwen-long API, our method reduces the expense of generating each literature survey to 0.1 RMB, enhancing its practical value.

The remainder of this paper is organized as follows: Sect. 2 reviews related works. Section 3 details our prompt design methodology and the step-by-step generation process. Section 4 presents the experimental setup and evaluation results, as well as discussing the advantages and limitations of our method. Finally, Sect. 5 concludes the paper.

2 Related Works

Automatic literature survey generation has garnered increasing attention in recent years due to its potential to enhance research efficiency and reduce the time required for literature review processes. However, because of the inherent challenges involved, few works have provided a systematic methodology to leverage AI tools to address this task comprehensively. Recently, AutoSurvey [7], a contemporary work to this paper, proposes equipping LLMs with initial retrieval and outline generation capabilities to automatically generate literature surveys. Additionally, some surveys [8] are claimed to be written with the help of LLMs.

In this paper, we focus on the NLPCC 2024 *Scientific Literature Survey Generation* evaluation task [9]. To the best of our knowledge, this is the first public evaluation task on automatic literature survey generation. Our system participated in this contest and achieved third place, with an overall score only 0.03% lower than the second-place team.

Leveraging LLMs for generating long-form text has become a rapidly evolving area of research. Some efforts have explored novel attention mechanisms to model long contexts effectively [10–12]. However, recent studies [13,14] point out that these long-context LLMs cannot fully utilize the information embedded in the extensive context window.

Instead of extending the context window of LLMs, some works explore the use of outlines to organize long text generation. [15] employ detailed outlines to improve the coherence of long story generation. [16] propose a pipeline that includes outline generation, information retrieval augmentation, augmented outline generation, and content generation for long blog creation. In this paper, we also adopt an outline-based generation strategy to manage the long context required for generating scientific literature surveys in a step-by-step manner.

3 Methodology

3.1 Task Formulation

In this paper, we address the NLPCC 2024 *Scientific Literature Survey Generation* evaluation task, which focuses on generating literature surveys based on given subjects and reference papers.

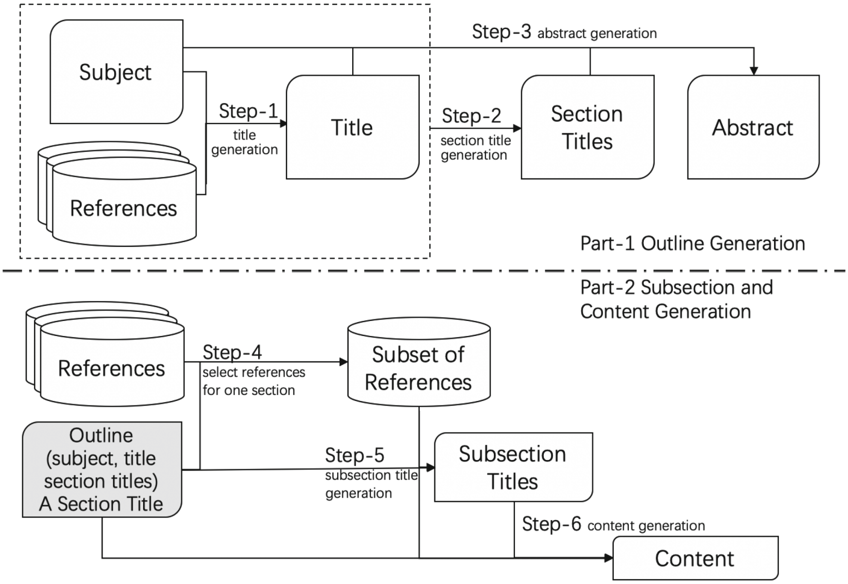
Subject
References
Step-1
title generation
Title
Step-2
section title generation
Section Titles
Step-3 abstract generation
Abstract
Part-1 Outline Generation
Part-2 Subsection and Content Generation
References
Step-4
select references for one section
Subset of References
Outline (subject, title section titles) A Section Title
Step-5
subsection title generation
Subsection Titles
Step-6 content generation
Content

Fig. 2. An illustration of our step-by-step literature survey generation framework.

Specifically, illustrated in Fig. 1, given a list of reference papers $R = \{r_1, r_2, ..., r_n\}$ and a list of key subjects (topics) $S = \{s_1, s_2, ..., s_m\}$, our goal is to produce a well-structured literature survey $\langle T, A, M\rangle$. The output includes the title T, abstract A, and the main body $M = \{(h_1, c_1), (h_2, c_2), ..., (h_k, c_k)\}$, where h_i represents the headings at different levels and c_i denotes the corresponding contents.

Notice that we did not utilize the *reference content* field provided in the test set because the NLPCC 2024 task guideline claimed that only the *subject* and *reference* were provided for testing. Additionally, we did not employ any retrieval-based methods to augment the reference content based on the titles of the reference papers, as the guidelines explicitly state that *data other than the training set cannot be used in the process of model development.*

3.2 Overview

The framework for our step-by-step literature survey generation is illustrated in Fig. 2. The generation pipeline consists of six steps, divided into two main phases.

In the first phase, outline generation, we instruct the LLM to sequentially generate the title, section headings, and abstract. Each step is conditioned on all previous outputs to ensure a coherent planning of the entire literature survey.

In the second phase, subsection and content generation, we select references for each section and then generate subsection headings and content based on the generated outline and the selected references. This design ensures that the generation of each subsection is aware of the overall structure of the paper, while also reducing the input length to save costs on LLM API usage.

3.3 Prompt Design

In this section, we will elaborate on the prompt designs for each of the six steps in Fig. 2.

Step 1: Title Generation - We instruct the LLM to generate a suitable title for the literature survey based on the provided subject and reference papers using the following prompt. We use semicolons (;) and newline characters (\n) to concatenate multiple subjects and references, respectively.

Prompt-1, title generation:

There is an academic puzzle for you.

I will give you a list of references. In subject “{*subject*}”, a survey paper existing with these references. You will give me a guess of the title of the survey paper.

Here are the references: {*references*}

The output should be in several lines, and the content in the last one is your answer (the title that you guess).

Only one guess is required. The title should start with “Title:”.

We observe that a chain-of-thought strategy [17] is generally integrated into existing LLM products, causing the model likely to generate explanations as well. We do not penalize the generation of additional information. Instead, we instruct the model to generate the title on the last line with a specific marker, *Title:*. We verify the output format, and if the specific marker is not found, we reinstruct the LLM while retaining the memory. The same prompt is used, but we add a prefix: *The response format is incorrect. Note that only one guess is required. The title should start with “Title:”.*

Step 2: Section Heading Generation - We use a follow-up question to generate the section headings while retaining the previous information, including references, subjects, title, and potentially, their explanations. We instruct the model to generate between 6 and 10 headings. Although the LLM might not always adhere strictly to this range, it is only required to regenerate the headings if the initial count falls outside the range of 3 to 25. We concatenate these headings and their indices with newline characters (\n) to form the *outline* for the following prompts.

Step 3: Abstract Generation - We use the following prompt to generate the abstract of the paper by summarizing the generated title and outline under the given subjects.

Prompt-2, section heading generation (a follow-up question):

Can you guess the outline of this paper? Just generate a list of first-level headings.

About 6–10 first-level headings are good! Just output the first-level headings, do not generate any other content.

No item number is required, each first-level heading begins with "*".

Prompt-3, abstract generation:

You are an academic paper writing assistant in the subject *"{subject}"*.

I am writing a survey paper titled with *"{title}"*.

Here is my outline : {*outline*}

Can you write an abstract for me?

You may write only 1 paragraph with about 200–500 words, do not include more detailed headings, or any lists.

You should focus on the content of the abstract, do not repeat the word "abstract" or including any other content.

Step 4: Selecting References for Each Section - The full reference list is so extensive that using it in the subsequent generation steps would incur high API usage costs. Additionally, conditioning each section on the same group of references may lead to similarity across sections. Therefore, we select a subset of references for each section before generating the detailed structure and content.

Specifically, for each section title generated in step 2, we instruct the LLM to generate relevant references using the following prompt. The generated references will be used in the generation of subsection headings and contents for the specific section.

Prompt-4, reference selection:

You are an academic paper writing assistant in the subject *"{subject}"*.

I am writing a survey paper titled with *"{title}"*.

Here is my outline : {*outline*}

I will give you a list of references. Can you help me choose from these references that might be useful when I write this chapter the section *"{sec heading}"*?

Here are the references: {*references*}

The output should be in several lines. References that you think may be useful should be on one line each, beginning with "*".

Please retain the square bracketed numbers (like [1], [20]) I give for each reference.

Step 5: Subsection Heading Generation - We use prompt-5 to generate the subsection headings. The LLM leverages information from the title, outline, and the selected references to generate subsection headings for a given section. We then extend the outline generated in step 2 by incorporating the subsection headings for all sections, providing a comprehensive outline for content generation in the next step.

Prompt-5, subsection heading generation:

You are an academic paper writing assistant in the subject *"{subject}"*.

I am writing a survey paper titled with *"{title}"*.

Here is my outline: {*outline*}

Here a list of reference papers: {*subset of references*}

I am working on the section *"{sec heading}"*.

Can you write the second-level headings for this section? 3–5 second-level headings are cool!

Just output the corresponding second-level headings, do not generate any other content.

No item number is required, each second-level heading begins with "* ".

Step 6: Content Generation - Finally, detailed content is generated for each subsection using prompt-6. To encourage the content is coherent and aligned with the overall structure, we provide an outline consisting of all the section and subsection headings. We also supply the subset of relevant references generated in step-4 and encourage the model to cite them appropriately by an example.

We observed that the prompt can become quite lengthy, causing the LLM to lose focus on the main task of generating subsection content. To address this, we emphasize the subsection title several times within the prompt to ensure the model remains focused.

Prompt-6, subsection content generation:

You are an academic paper writing assistant in the subject *"{subject}"*.

I am writing a survey paper titled with *"{title}"*.

Here is my outline : {*outline*}

Here a list of reference papers. I hope you will cite these references in the generated content. For example, [123] indicates a reference to the reference that begins with [123]. {*subset of references*}

Can you write the content of *"{subsec heading}"* part in the section *"{sec heading}"*? You may write 3–5 paragraphs with about 1000 words, do not include more detailed headings. You should focus on the subsection *"{subsec heading}"*, do not repeat the headings or including other content.

4 Experiments

4.1 Dataset

We use the dataset from the NLPCC 2024 *Scientific Literature Survey Generation* evaluation task, provided by Kexin Technology [9]. The dataset comprises randomly crawled arXiv survey papers along with their references, containing 500 survey papers in the training set and 200 in the test set. Since the evaluation task did not designate specific instances for validation rather than training, we have not included a separate validation set in our statistics.

Table 1. Dataset statistics, including the total number of papers, the average number of subjects, references, references with content, and the average content length, respectively.

	#paper	avg.subject	avg.ref	avg.ref_content	avg.content
Train	500	1.7	98.5	21.1	72.5k
Test	200	1.9	126.8	27.2	–

In the training set, the fields including *title*, *article id*, *subject*, *abstract*, *content*, *reference*, and *reference content* are provided. While in the test set, only *subject*, *reference*, and *reference content* are provided. The statistics of the dataset is shown in Table 1.

In this paper, we only use the training set for prompt design, instead of using them for parameter tuning or in-context learning. As discussed in §3.1, because of the requirements in task guideline, we did not utilize the *reference content* field provided in the test set, nor did we employ any retrieval-based methods to augment the reference titles.

4.2 Implement Detail

We implemented our system using the Qwen-long version of Alibaba's Tongyi Qianwen model [5], with a cost of 0.5 RMB per million tokens for input and 2 RMB per million tokens for output. For instances where the reference list is too long for Qwen-long to process (with the role of *user*), we truncate the references, leaving only the first 80–100 references. Additionally, in cases where the reference list contains inappropriate content that Qwen-long refuses to process, we try to use references with odd or even indices to circumvent this issue. Generating all 200 surveys in the test set cost 20.86 RMB, approximately 0.10 RMB per paper. This low cost enhances the practicality of our method.

4.3 Evaluation Metric

To evaluate the quality of the generated literature survey, we use three kinds of metrics:

ROUGE [18] is a recall-oriented reference-based metrics. Specifically, we employ ROUGE1, ROUGE2, and ROUGEL scores. We use the implement provided by Google[1].

[1] https://github.com/google-research/google-research/tree/master/rouge.

Table 2. Evaluation Results of Participating Teams

	S-H Recall	ROUGE1	ROUGE2	ROUGEL	Human	Score
Flying	**97.01**	43.99	12.79	13.07	**70.62**	**63.64**
CNKI-Research-Group	90.21	**45.95**	14.07	**14.16**	68.48	61.14
ID (Ours)	95.84	43.53	**14.68**	12.91	63.78	61.11
Literature Summarizer Wizards	95.06	41.58	13.29	13.41	57.66	58.49
BaseLine	91.03	37.07	10.70	12.25	57.85	56.30
Literature hunter	94.03	37.07	11.54	12.35	54.45	56.27
DUFL2024NLP	87.48	26.71	7.81	9.54	52.22	51.46
IDEA of NUDT	84.59	27.26	5.93	9.95	35.97	44.98

Soft Heading Recall (S-H Recall for short) is used to evaluate the structure of the generated survey.

$$\mathrm{Sim}(t_i, t_j) = \cos\left(\mathrm{embed}(t_i), \mathrm{embed}(t_j)\right)$$

$$\mathrm{card}(T) = \sum_{i=1}^{|T|} \frac{1}{\sum_{j=1}^{|T|} \mathrm{Sim}(t_i, t_j)}$$

$$\mathrm{card}(R \cap G) = \mathrm{card}(R) + \mathrm{card}(G) - \mathrm{card}(R \cup G)$$

$$\text{soft heading recall} = \frac{\mathrm{card}(R \cap G)}{\mathrm{card}(R)}$$

$T = \{t_1, t_2, t_3, \ldots, t_K\}$ represents a group of the chapter titles/heads in a generated/reference survey. R and G are the chapter titles of the generated and reference survey, respectively. The bge-large-en-v1.5[2] model is used for text embedding. This score encourages the similarity between generated and reference chapter titles while punishes the similarity of titles within the generated survey.

Human evaluation is also used to manually evaluate the following aspects:

- Fluent language with clear expression;
- Logical article structure;
- Ample, reliable, and accurate citations;
- Consistency of content with the theme, staying on-topic;
- Broad analytical scope.

The overall **score** is calculated as the average of the ROUGE score, the soft heading recall, and the human score. Note that the ROUGE score here is the average of ROUGE1, ROUGE2, and ROUGEL.

[2] https://huggingface.co/BAAI/bge-large-en-v1.5.

4.4 Evaluation Result

The official evaluation results is shown in Table 2. We can see that our system, **ID**, achieved the third place in the evaluation task, with an overall score only 0.03% lower than the second-place team.

Additionally, our system demonstrates competitive performance across other metrics. Notably, in soft heading recall, our system achieves a score of 95.84%, securing second place and trailing the leading team by just 1.17%. This result highlights its effectiveness in generating high-quality headings. We attribute this success to our sequential generation design, which includes an individual heading generation step, providing a high-level perspective that enhances the ability of large language models to generate meaningful headings.

However, human evaluation remains a relatively weak aspect of our method. Without leveraging the content of the reference papers, the hallucination phenomenon of LLMs becomes significant. Consequently, our method cannot ensure the accuracy and reliability of the citations and analysis in the generated survey, thereby impairing our human evaluation score. In the future, we will incorporate the reference content into our framework to improve the factual accuracy and reliability of our outputs, aiming to generate more accurate and trustworthy literature surveys.

4.5 Case Study

An example of our generated survey paper and the corresponding reference one is shown in Table 3. Our system successfully captures the subject of the paper, *interpretability in neural networks*, and generates a relevant title.

For the outline generation, this example achieves a soft heading recall of 97.78%. While the outline produced by our system is somewhat lengthy and too detailed, it successfully captures key aspects of the survey, such as *explainable AI*, *Saliency Maps*, and *Black-Box Models*. The discussion on *ethical considerations* and *future directions* at the end of the survey is also well-founded. In the future, some specific instruction strategies can be designed to encourage the LLM to generate more concise outlines.

For the abstract and content (citation), we can see that our generated paper captures several common points with the reference one (**bold** in Table. 3), resulting in appropriate abstracts and citations. However, some citations are still influenced by hallucinations, which could be mitigated by incorporating the detailed content of the reference papers.

Table 3. A comparison between our generated survey and the reference paper [19].

Title **Reference:** On Interpretability of Artificial Neural Networks: A Survey **Ours:** A Comprehensive Survey on Explainable Machine Learning and Artificial Intelligence: Methods, Applications, and Challenges
Outline **Reference:** I. INTRODUCTION A. What Does Interpretability Mean? B. Why Is Interpretability Difficult? C. How to Build a Good Interpretation Method? II. A SURVEY ON INTERPRETATION METHODS A. Taxonomy Definition B. Post-hoc Interpretability Analysis C. Ad-hoc Interpretable Modeling III. INTERPRETABILITY IN MEDICINE IV. PERSPECTIVE V. CONCLUSION VI. ACKNOWLEDGEMENT **Ours (partial):** 1. Introduction 1.1 Background and Importance of Explainable ML/AI 1.2 Historical Perspective and Evolution of XAI 1.3 Key Challenges and Objectives in Explainability ... 4. Saliency Maps and Visualization Techniques 5. Auditing and Fairness in Black-Box Models ... 15. Ethical Considerations and Regulations 16. Future Directions and Open Challenges
Abstract **Reference** : ... In this paper, we propose a simple but comprehensive taxonomy for interpretability, systematically review recent studies in improving interpretability of neural networks, **describe applications of interpretability in medicine**... **Ours:** ... The paper explores various XAI techniques, including Case-Based Reasoning, Saliency Maps, and Shapley Values, while examining their applications and limitations. It also discusses the **use of Deep Learning in healthcare** and the interpretability aspects of Generative Adversarial Networks (GANs) and Recurrent Neural Networks (RNNs) in Natural Language Processing. ...
Content & Citation **Reference** : Also, interpretability plays an important role in ethic use of deep learning techniques [57]. **To build patients' trust in deep learning**, interpretability is needed to hold a deep learning system accountable [57]. **Ours:** Additionally, the ethical concerns surrounding data privacy, bias, and algorithmic accountability need to be addressed. Regulatory frameworks, such as those outlined in [57], must adapt to accommodate the integration of **AI technologies in healthcare while maintaining patient safety and privacy.** **Citation information:** [57] J. R. Geis, et al., "Ethics of artificial intelligence in radiology: summary of the joint European and North American multisociety statement," Canadian Association of Radiologists Journal, vol. 70, no. 4, pp. 329–34, 2019.

5 Conclusion

In this paper, we presented a novel approach to generating scientific literature surveys with LLMs. We designed a series of instructions to generate well-structured and relevant content in a step-by-step manner. Our system with Qwen-long achieved third place in the NLPCC 2024 *Scientific Literature Survey Generation* evaluation task, with an overall score only 0.03% lower than the second-place team. However, challenges remain in ensuring the accuracy and reliability of citations and content due to the hallucination of LLMs and the absence of reference content. Future work will focus on addressing these limitations to generate more robust and trustworthy automatic literature survey.

Acknowledgements. This work is supported by NSFC (62206070) and the Innovation Fund Project of the Engineering Research Center of Integration and Application of Digital Learning Technology, Ministry of Education (1221014). The datasets used in this paper is supported from Kexin Technology.

References

1. Fink, A.: Conducting research literature reviews: from the internet to paper. Sage publications (2019)
2. Liang, W., et al.: Mapping the increasing use of LLMs in scientific papers. arXiv preprint arXiv:2404.01268 (2024)
3. OpenAI: GPT-4 technical report. arXiv preprint arXiv:2303.08774 (2023)
4. Ouyang, L., et al.: Training language models to follow instructions with human feedback. In: NeurIPS (2022)
5. Bai, J., et al.: Qwen technical report. arXiv preprint arXiv:2309.16609 (2023), https://ar5iv.org/abs/2309.16609
6. Anthropic, A.: The claude 3 model family: Opus, sonnet, haiku. Claude-3 Model Card **1** (2024)
7. Wang, Y., et al.: AutoSurvey: large language models can automatically write surveys. arXiv preprint arXiv:2406.10252 (2024)
8. Schulhoff, S., et al.: The prompt report: a systematic survey of prompting techniques. arXiv preprint arXiv:2406.06608 (2024)
9. Tian, Y., et al.: Overview of the NLPCC2024 shared task6: scientific literature survey generation. In: Proceedings of the NLPCC 2024 Conference (2024)
10. Dao, T., Fu, D., Ermon, S., Rudra, A., Ré, C.: FlashAttention: Fast and memory-efficient exact attention with IO-awareness. Adv. Neural. Inf. Process. Syst. **35**, 16344–16359 (2022)
11. Ding, J., et al.: LongNet: scaling transformers to 1,000,000,000 tokens. arXiv preprint arXiv:2307.02486 (2023)
12. Xiao, G., Tian, Y., Chen, B., Han, S., Lewis, M.: Efficient streaming language models with attention sinks. In: ICLR (2023)
13. Li, J., Wang, M., Zheng, Z., Zhang, M.: LooGLE: can long-context language models understand long contexts? arXiv preprint arXiv:2311.04939 (2023)
14. Li, T., Zhang, G., Do, Q.D., Yue, X., Chen, W.: Long-context LLMs struggle with long in-context learning. arXiv preprint arXiv:2404.02060 (2024)

15. Yang, K., Klein, D., Peng, N., Tian, Y.: Doc: Improving long story coherence with detailed outline control. arXiv preprint arXiv:2212.10077 (2022)
16. Lee, Y., Ka, S., Son, B., Kang, P., Kang, J.: Navigating the path of writing: outline-guided text generation with large language models. arXiv preprint arXiv:2404.13919 (2024)
17. Wei, J., et al.: Chain-of-thought prompting elicits reasoning in large language models. Adv. Neural. Inf. Process. Syst. **35**, 24824–24837 (2022)
18. Lin, C.Y.: Rouge: a package for automatic evaluation of summaries. In: Text Summarization Branches Out: Proceedings of the ACL-04 Workshop, pp. 74–81. Association for Computational Linguistics, Barcelona, Spain (2004)
19. Fan, F.L., Xiong, J., Li, M., Wang, G.: On interpretability of artificial neural networks: a survey. IEEE Trans. Radiat. Plasma Med. Sci. **5**(6), 741–760 (2021)