Generation of Scientific Literature Surveys Based on Large Language Models (LLM) and Multi-Agent Systems (MAS)

Ruihua Qi[1,2(✉)], Weilong Li[1], and Haobo Lyu[1]

[1] School of Software Engineering, Dalian University of Foreign Languages, Dalian 116044, China
{236493401,236493404}@student.dlufl.edu.cn

[2] Research Center for Language Intelligence, Dalian University of Foreign Languages, Dalian 116044, China
rhqi@dlufl.edu.cn

Abstract. With the rapid increase in the number and speed of scientific publications, researchers face significant time pressure when conducting literature reviews. This paper presents an automatic literature review generation method leveraging large language models (LLMs) and multi-agent systems (MAS). By designing multiple agent roles, including reference parsing, analysis, generation, and integration agents-this method fully utilizes the natural language processing capabilities of LLMs and the collaborative strengths of MAS to produce high-quality literature reviews. In the NLPCC2024 evaluation task, our method excelled in multiple automatic evaluation metrics (such as SoftHeadingRecall and ROUGE) and manual evaluations, showcasing its great potential for practical applications.

Keywords: Large Language Models · Multi-Agent Systems · Literature Review Generation · Natural Language Processing · Collaborative Generation

1 Introduction

With the rapid increase in the number and speed of scientific publications, researchers face significant time pressure when conducting literature reviews. According to Johnson et al. (2018) [1], as of August 2018, there were 33,100 active peer-reviewed English journals, publishing approximately 3 million papers annually, with an annual growth rate of about 5%. Scientific literature surveys are foundational in advancing scholarly understanding and innovation. They play a crucial role in academic research by providing a comprehensive synthesis of existing knowledge within a specific field, enabling researchers to identify

R. Qi, W. Li and H. Lyu—These authors contributed equally to this work.

D. F. Wong et al. (Eds.): NLPCC 2024, LNAI 15363, pp. 169–180, 2025.
https://doi.org/10.1007/978-981-97-9443-0_14

gaps, trends, and key findings, and guiding future research directions. By systematically evaluating and integrating previous studies, literature surveys contribute to the validation and refinement of theoretical frameworks. These surveys enable researchers to gain a thorough understanding of existing work within their field, uncover gaps in the literature, and discover potential directions for future research. By questioning established assumptions, pinpointing key issues, and stimulating scientific dialogue, literature surveys significantly contribute to the progression of the discipline. They systematically analyze and synthesize the literature using various methods, such as critical evaluation, exploratory description, and integrative assessment. This process not only consolidates current knowledge but also suggests new pathways for advancing research.

Generating a scientific literature survey requires meticulous planning, extensive searching, and rigorous screening of a large body of literature. This task involves systematically reviewing and synthesizing existing research within a specified domain. It requires identifying, evaluating, and integrating relevant studies to provide a comprehensive overview of the current state of knowledge [2].

Traditional methods for conducting literature reviews primarily rely on manual searching, reading, and synthesizing literature. Although this approach is thorough and comprehensive, it has significant drawbacks. First, manually processing a large volume of literature requires substantial time and effort [3]. Second, due to personal biases and knowledge gaps, researchers may overlook important studies when selecting and evaluating literature [4]. As the volume of scientific publications rapidly grows, this process becomes increasingly complex, and accurately synthesizing and interpreting collected data becomes more time-consuming and labor-intensive.

In recent years, advancements in artificial intelligence, particularly natural language processing (NLP) technologies, have provided new tools to enhance the efficiency of literature reviews [5]. The emergence of large language models (LLMs) offers new opportunities for automating and streamlining the systematic literature review (SLR) process. LLMs are trained on extensive text datasets and excel at understanding and generating human-like language [6]. Research has shown that LLMs perform exceptionally well in understanding context, identifying key points, and generating summaries, making them highly effective in assisting literature reviews [7]. They can quickly process and analyze large volumes of text, generating structured summaries and analyses, significantly improving the efficiency of literature reviews. Multi-agent systems (MAS) further enhance this process by facilitating distributed problem-solving and parallel processing, thereby improving the efficiency and accuracy of handling extensive literature.

This paper aims to enhance the efficiency and quality of scientific literature surveys by leveraging large language models (LLMs) and multi-agent systems (MAS). By systematically integrating natural language processing technologies, this research addresses inherent challenges in traditional literature review processes, such as time consumption, potential biases, and overwhelming volume of literature. The main contributions are as follows:

1. Automated scientific literature survey generation with LLMs: Utilizing LLMs to automatically generate literature surveys, improving coherence and accuracy.
2. Design of Multi-Agent System for automated scientific literature survey generation, implementing MAS to integrate diverse methods and perspectives, enriching the analytical depth of literature surveys. This approach enhances efficiency and scalability through collaborative task division, and provides robust error detection and correction capabilities.

2 Related Work

2.1 Automation of Systematic Literature Reviews

According to Ananiadou et al. (2009) [8], literature reviews involve searching for relevant studies, screening to focus on key evidence, mapping research activities to engage stakeholders and set priorities, and synthesizing evidence from diverse sources. Preparing a systematic review includes creative tasks for developing questions and protocols, and technical tasks that can be automated (Tsafnat et al., 2014) [9].

In the current landscape of systematic literature review (SLR) automation, natural language processing (NLP) preprocessing involves various techniques. Traditional SLR automation uses shallow machine learning methods, requiring manual feature engineering and fine-tuning for each specific domain or dataset. Many legacy NLP representation techniques are now replaced by word embeddings, which capture the meaning of words and reduce the need for extensive text cleaning and manual feature creation.

SLR automation is typically achieved through classification, a supervised machine learning task, with models commonly evaluated using cross-validation. Key metrics include Work Saved over Sampling (WSS), Recall, Precision, and F-measure, with WSS encompassing the essence of the latter three metrics, making it the preferred metric for SLR automation. Support Vector Machines (SVM) and Bayesian Networks, such as Naïve Bayes classifiers, are predominantly used across various steps of the SLR process. However, there is a significant lack of evidence supporting the use of deep learning techniques in SLR automation [10].

Van Dinter et al. (2021) [10] conducted a systematic review of literature related to literature reviews to collect and summarize the current state of research. Their study is the first systematic literature review focusing on automated systems for literature reviews, emphasizing all stages of the review process, including NLP and ML techniques from retrospective methods.

2.2 Automatic Literature Review Generation Using LLMs

Automatic literature review generation is a complex challenge in Natural Language Processing (NLP). Traditional methods are time-consuming and labor-intensive, often biased and incomplete due to the vast amount of information.

Recent advancements in Large Language Models (LLMs) have shown significant potential in automating this process.

Kasanishi et al. (2023) [11] introduced SciReviewGen, a large-scale dataset for automatic literature review generation, containing over 10,000 literature reviews and 690,000 cited papers, primarily in computer science. This dataset facilitates the evaluation of transformer-based summarization models for literature reviews. They also proposed the Query-weighted Fusion-in-Decoder (QFiD) model, which weights cited papers based on query relevance. Experiments showed that QFiD generates more relevant and coherent reviews than other models, though challenges like hallucinations and insufficient detail remain.

The introduction of SciReviewGen and QFiD underscores LLMs' potential to enhance literature review efficiency and reduce manual effort. This dataset and model lay the groundwork for future research in automatic literature review generation, addressing issues like citation network integration and full-text information enhancement to improve review accuracy and comprehensiveness.

2.3 Utilizing LLMs to Improve Literature Review Efficiency

Antu et al. (2023) [12] explored using LLMs like OpenAI's ChatGPT to help students complete literature reviews for their theses. They found that ChatGPT significantly enhances review efficiency and comprehensiveness by quickly processing large volumes of text and providing structured summaries. However, limitations include generating non-existent citations (hallucinations) and using outdated or irrelevant sources.

In case studies in computer science and communication, Antu et al. demonstrated ChatGPT's ability to generate research topics and identify relevant literature [12]. Despite drawbacks such as overgeneralization and inaccuracy, ChatGPT serves as a useful starting point for literature reviews. The authors emphasized the importance of "human-AI collaboration," where researchers critically assess and refine AI-generated outputs to ensure quality and relevance.

The study shows that while LLMs like ChatGPT can streamline the literature review process, human oversight is crucial. This combined approach of AI and human expertise helps undergraduates complete their research more effectively, balancing the strengths of both. The comparative analysis highlights the transformative potential of LLMs in academic research, especially in automating tedious literature review tasks. However, both studies stress the importance of human supervision to mitigate AI-generated content's limitations, such as hallucinations and biases [13].

3 Methodology

3.1 Task Definition

The objective of the Scientific Literature Survey Generation competition in NLPCC 2024 is to develop a retrieval-enhanced text generation model capable of

producing comprehensive literature reviews. A series of topic-related references and provided metadata (including titles, topics, abstracts, and references) are utilized to generate high-quality literature reviews. The tasks involve training and fine-tuning the model using the supplied training and validation sets. The goal is to create a model that produces literature reviews meeting specific evaluation criteria. In the final stage, literature reviews are generated based on the test dataset, and the resulting texts generated by the model are submitted for assessment.

3.2 Motivation

In the scientific research field, the rapid increase in the number and speed of publications imposes significant time and workload pressures on researchers conducting literature reviews. Traditional manual review methods are time-consuming and labor-intensive, hindering the ability to quickly and efficiently cover the latest research. Review papers typically follow a structured format: title, abstract, main text, and conclusion. This paper introduces an automated literature review generation method leveraging large language models (LLM) and multi-agent systems (MAS) to address this issue. Our system is especially suited for fields like computer science and life sciences that require rapid processing and summarization of extensive literature. The design aims to enhance review efficiency and quality by intelligently parsing, analyzing, generating, and integrating literature, distributing the review process among various LLM-based agents.

To tackle these challenges, we designed a multi-agent system where each agent is responsible for tasks such as literature parsing, analysis, generation, and integration. These agents collaborate, utilizing the robust text generation capabilities of large language models to automatically produce structured, high-quality literature reviews.

Our approach is grounded in the theories of natural language processing and multi-agent systems. Large language models, like the qwen [14] series from Tongyi Qianwen and OpenAI's GPT [15], excel in text generation and understanding, producing coherent and high-quality text. For this experiment, we employed the qwen-long [16] model, known for its strong generation capabilities, cost-effectiveness, and suitability for large-scale applications. Multi-agent systems manage complex tasks efficiently through division of labor and collaboration. By combining these technologies, we achieve efficient and effective automatic literature review generation.

3.3 The Proposed Model

The system introduced in this study, named the "Automated Literature Review Generation System" (ALRGS), is depicted in Fig. 1. The system consists of five main agents: the Reference Analysis Agent, the Title & Abstract Writer Agent, the Section Content Writer Agent, the Conclusion Writer Agent, and the Integrator Agent. Each agent is responsible for a specific task, and through their

collaborative efforts, they enable efficient and automated literature review generation. Below is a detailed description of each agent.

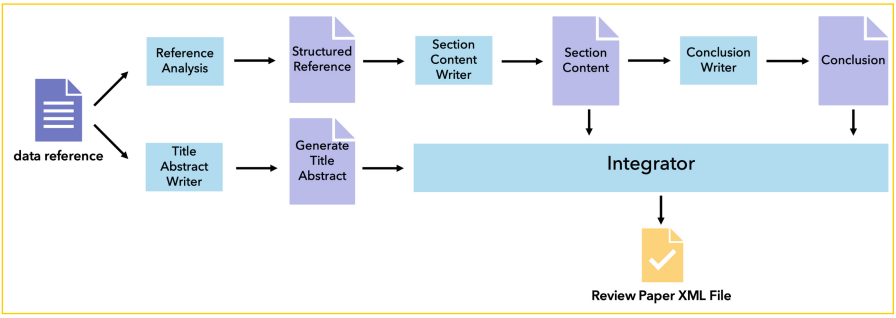

Fig. 1. Workflow of the ALRGS.

Each of these agents performs its specific task, utilizing predefined workflows to achieve efficient and automated literature review generation. The following section provides a detailed description of each agent and its core algorithm.

Reference Analysis Agent. The primary function of the Reference Analysis Agent is to extract references and related content from the given JSON file. Using predefined prompts, this agent analyzes the references, identifies their logical relationships within the review, and classifies them. The classification results are presented as structured chapter titles and related reference lists, ensuring that each reference is individually listed and described, avoiding any merging or summarizing. The prompt used by this agent is shown in **Prompt1**.

Prompt1:Reference Analysis Agent

Please analyze the following references and their abstracts, identify their logical relationships in the review, and classify them. All references must be listed individually, without omission or merging. Output structured section titles and list the related references under each section title. Ensure that each reference is clearly associated with a specific section, and avoid summarizing multiple references together. The output must be organized, with accurate and concise section titles that reflect the thematic grouping of the references.

Title and Abstract Writer Agent. This agent aims to automatically generate high-quality titles and abstracts based on the analysis of structured references. By reading classified references from the specified file, the agent identifies themes and logical relationships, generating suitable titles and concise abstracts for the

review paper. This process leverages the robust text understanding and logical reasoning capabilities of large language models to ensure academic accuracy. The prompt used by this agent is shown in **Prompt2**.

Prompt2:Title and Abstract Writer Agent

Based on the following section titles, please generate an appropriate title and a brief abstract for a review paper: Section Titles: Title: "title_here" Abstract: "abstract_here". Please generate the paper title and abstract.

Section Content Writer Agent. The Section Content Writer Agent generates the main text content for each section by leveraging the structured and classified reference texts. Utilizing the text generation capabilities of large models, it fine-tunes LLM parameters (Prompt Engineering) and employs generation strategies to produce high-quality review content that adheres to academic standards. The prompt used by this agent is shown in **Prompt3**.

Prompt3:Section Content Writer Agent

The following are section titles and related references. Please write the text content for each section. "structured_references" Please write the section content. Only the body text needs to be outputted, omitting all other parts. Do not include #, and do not reply in Markdown format. Respond as per the specified requirements. Ensure that each section's content is sufficiently detailed and provides comprehensive analysis. The output should be at least 3000 characters long. Without hashtag #.

Conclusion Writer Agent. This agent receives the section content generated by the previous agent and systematically organizes and analyzes the review content. It identifies and integrates key points and research findings from each section to create a precise and comprehensive conclusion. The Conclusion Writer Agent synthesizes these points into a coherent and logical conclusion, ensuring the review's completeness and academic integrity. The prompt used by this agent is shown in **Prompt4**

Prompt4:Conclusion Writer Agent

Based on the following section contents, please write the Conclusion section for the review paper. Section Contents: "chapter_contents" Do not include #. Please write the Conclusion:

Integrator Agent. This agent compiles the content generated for each section into a complete literature review and produces the final document in XML format. By outputting in XML, the system ensures the review is structured and standardized, facilitating further evaluation and publication.

In summary, our automated literature review generation system, based on LLM and MAS, achieves efficient and accurate literature review generation through intelligent parsing, analysis, generation, and integration. This significantly reduces researchers' workload and enhances the coverage and accuracy of literature reviews.

4 Experiments

4.1 Data

In this experiment, we used a dataset of 500 English scientific review papers, with 400 for the training set and 100 for the validation set [17]. Each sample includes the article's title, article_id, subject, abstract, references, reference content, and review content. The training data samples are provided in JSON format, with "reference_content" containing the titles and abstracts of some references. The test dataset consists of 200 review papers, formatted as JSON files with the structure {"subject": "", "reference": [...]}. All data were supplemented with missing reference abstracts through web scraping to ensure completeness and quality.

4.2 Experimental Setup

In this experiment, we employed multiple large language models (qwen-long, gpt-4o [18], gpt-3.5-turbo [19]) to generate literature reviews. All models utilized the dataset supplemented with reference abstracts through web scraping, ensuring more effective use of reference content during the generation process.

4.3 Evaluation Metrics

To evaluate the quality of the generated literature reviews, we adopted multiple evaluation metrics, as shown in (1)-(5), including automated and structural assessments:

- **Automated Evaluation**: We used ROUGE-1/2/L to assess the content quality of the generated reviews [20].
- **Structural Evaluation**: We employed Soft Heading Recall to evaluate the structure of the generated reviews, embedding all generated chapter titles using the bge-large-en-v1.53 model.

$$\mathrm{Sim}(t_i, t_j) = \cos(\mathrm{embed}(t_i), \mathrm{embed}(t_j)) \tag{1}$$

$$\mathrm{count}(t_i) = \frac{1}{\sum_{j=1}^{K} \mathrm{Sim}(t_i, t_j)} \tag{2}$$

$$\text{card}(\text{T}) = \sum_{i=1}^{K} \text{count}(t_i) \tag{3}$$

$$\text{card}(\text{R} \cap \text{G}) = \text{card}(\text{R}) + \text{card}(\text{G}) - \text{card}(\text{R} \cup \text{G}) \tag{4}$$

$$\text{soft heading recall} = \frac{\text{card}(\text{R} \cap \text{G})}{\text{card}(\text{R})} \tag{5}$$

– **Manual Evaluation**: LLMs and human experts evaluated the following aspects: language fluency and clarity; logical structure; sufficiency, reliability, and accuracy of citations; consistency with the theme, ensuring no deviation; and comprehensiveness of the analysis.

4.4 Experimental Results and Analysis

In this experiment, we compared the effectiveness of the qwen-long model with the combined use of gpt-4o and gpt-3.5-turbo. The results of this comparison are summarized in Table 1, which presents the performance of the models across various evaluation metrics, including Soft Heading Recall and ROUGE scores. Analyzing these results, we derived the following key conclusions.

Firstly, as shown in Table 1, the qwen-long model excelled in the Soft Heading Recall metric, demonstrating high accuracy and consistency in generating chapter titles and content structures for reviews. This advantage is largely due to the qwen-long model's excellent performance in handling long texts and complex contextual relationships. Its strong contextual understanding ensures logically coherent and well-structured content.

Table 1. Experiments Result.

Model	SoftHeadingRecall	Rouge1	Rouge2	RougeL	Human
qwen-long	0.8646	0.3760	0.1070	0.1141	-
gpt-3.5-turbo+gpt4o	0.8736	0.3451	0.0964	0.1110	-
qwen-long(Final)	0.8747	0.2670	0.0781	0.0954	0.5222

Additionally, the qwen-long model's training data spans a wide range of fields and topics, providing it with strong generalization capabilities. This allows the model to adapt flexibly to various themes and types of literature. The broad adaptability and low computational cost make the qwen-long model highly cost-effective for large-scale applications, especially in research fields such as computer science and life sciences that require processing large volumes of literature.

On the other hand, although the qwen-long model excels in structure generation, the combination model of gpt-4o and gpt-3.5-turbo slightly outperforms it in the details and diversity of content generation. This advantage is reflected in the combination model of ChatGPT, which, by integrating different versions of

models and leveraging their strengths, enhances the diversity and detail processing capabilities of the generated text. The synergistic effect of multiple models makes the generated content richer and more vivid, better matching the vocabulary and phrases of the references, thereby performing excellently on ROUGE-1 and ROUGE-2 metrics. This diverse expression and detailed content generation capability make the ChatGPT combination model perform better in tasks requiring high-precision content generation.

Nevertheless, each model has its applicable scenarios. The qwen-long model, while ensuring generation quality, has a higher cost-effectiveness, suitable for tasks that require rapid processing and summarization of large volumes of literature. Its advantage in structure generation allows it to provide high-quality chapter divisions and content organization. The ChatGPT combination model, on the other hand, excels in the details and diversity of content generation, making it suitable for tasks that require high-precision content generation. For example, for literature review tasks that need detailed analysis and description, the ChatGPT combination model can provide richer and more detailed content.

Additionally, our system design features a significant advantage with its highly decoupled, independent agents. This allows the selection of the most suitable model and agent based on the specific task's difficulty and requirements. This flexibility enhances the system's applicability and scalability, enabling dynamic adjustment and optimization of model usage for different types of literature reviews. For example, for tasks requiring rapid large-scale literature structure generation, the qwen-long model is preferred; for tasks needing high-detail content generation, the ChatGPT combination model is used.

In summary, this study provides new insights and methods for automated literature review generation by comparing the performance of different models, offering significant application value and research implications. Future research will continue to explore combining and optimizing the strengths of different models to further improve the performance and practical application of literature review generation systems.

5 Conclusion

This paper presents a method that combines Multi-Agent Systems (MAS) and Large Language Models (LLMs) to enhance the efficiency and quality of literature reviews. By breaking down the task into multiple subtasks and assigning specific responsibilities to each agent, we optimize parameter settings, and improve process interpretability. This method not only addresses traditional issues such as time consumption, bias, and information overload but also showcases the potential of LLMs in academic research. The main contribution of this paper is the introduction of a more efficient and reliable literature review method, providing researchers with a new tool to improve the comprehensiveness and accuracy of their reviews.

Despite its many advantages, the method has certain limitations. First, LLM-generated summaries of complex concepts may still be inaccurate. Second, the

method's effectiveness relies on the quality and diversity of the training data. Future research directions include incorporating Retrieval-Augmented Generation (RAG) technology, allowing LLMs to automatically retrieve full texts based on titles and abstracts, thus reducing hallucinations. Additionally, future agent systems could evolve independently, dynamically optimizing their performance by adjusting goals and strategies based on both historical data and current needs. These improvements are expected to further enhance the quality of literature reviews, providing stronger support for academic research.

In conclusion, this study provides an efficient and reliable method for literature reviews and suggests new directions for future research. We hope this method offers valuable insights, fostering further development and optimization of literature review techniques.

Acknowledgement. We would like to express our sincere gratitude to Kexin Technology for providing the invaluable data support and assistance for this event. Their contribution significantly enhanced the quality and scope of our study. This work is partially supported by grant from the Applied Basic Research Project of Liaoning Province (No. 2022JH2/101300270); Project of Humanities and Social Sciences Research of the Ministry of Education (No.23YJAZH109); Promote scientific research cooperation with Canada, Australia, New Zealand and Latin America and High-level Talent Training Program (LiuJin Mei 2023 No.25).

References

1. Johnson, R., Watkinson, A., Mabe, M., Ware, M.: The STM Report: an overview of scientific and scholarly publishing. Technical and Medical Publishers, International Association of Scientific (2018)
2. Kraus, S., Breier, M., Lim, W.M., et al.: Literature reviews as independent studies: guidelines for academic practice. RMS **16**(8), 2577–2595 (2022)
3. Boote, D.N., Beile, P.: Scholars before researchers: on the centrality of the dissertation literature review in research preparation. Educ. Res. **34**(6), 3–15 (2005)
4. Ridley, D.: The literature review: a step-by-step guide for students (2012)
5. da Silva Júnior, E.M., Dutra, M.L.: A roadmap toward the automatic composition of systematic literature reviews. Iberoamerican J. Sci. Meas. Commun. (2021)
6. Carlini, N., et al.: Extracting training data from large language models. In: 30th USENIX Security Symposium (USENIX Security 21), pp. 2633–2650 (2021)
7. Hariri, W.: Unlocking the potential of ChatGPT: a comprehensive exploration of its applications, advantages, limitations, and future directions in natural language processing. arXiv preprint learning, commerce, and politics. Wiley
8. Ananiadou, S., et al.: Supporting systematic reviews using text mining. Soc. Sci. Comput. Rev. **27**(4), 509–523 (2009)
9. Tsafnat, G., et al.: Systematic review automation technologies. Syst. Control Found. Appl. **3**, 1–15 (2014)
10. Van Dinter, R., Tekinerdogan, B., Catal, C.: Automation of systematic literature reviews: a systematic literature review. Inf. Softw. Technol. **136**, 106589 (2021)
11. Kasanishi, T., Isonuma, M., Mori, J., Sakata, I.: SciReviewGen: a large-scale dataset for automatic literature review generation. In: Findings of the Association for Computational Linguistics: ACL 2023, pp. 6695–6715. Toronto, Canada. Association for Computational Linguistics (2023)

12. Antu, S.A., Chen, H., Richards, C.K.: Using LLM (large language model) to improve efficiency in literature review for undergraduate research. LLM@ AIED, 8–16 (2023)
13. Williamson, S.M., Prybutok, V.: The era of artificial intelligence deception: unraveling the complexities of false realities and emerging threats of misinformation. Information **15**, 299 (2024). https://doi.org/10.3390/info15060299
14. Bai, J., Bai, S., Chu, Y., et al.: Qwen technical report. arXiv preprint arXiv:2309.16609 (2023)
15. Brown, T.B., et al.: Language models are few-shot learners. In: Advances in Neural Information Processing Systems (NeurIPS). https://arxiv.org/abs/2005.14165
16. Alibaba DAMO Academy.: Qwen-Long: Large Language Model. https://damo.alibaba.com/technology/large-language-model/qwen-long
17. Yangjie, T., et al.: Overview of the NLPCC2024 Shared Task6: scientific literature survey generation. http://tcci.ccf.org.cn/conference/2024/taskdata.php
18. OpenAI.: Hello GPT-4o. OpenAI. https://openai.com/index/hello-gpt-4o/
19. OpenAI.: GPT-3.5 Turbo Overview. OpenAI. https://platform.openai.com/docs/models/gpt-3-5
20. Lin, C-Y.: Rouge: a package for automatic evaluation of summaries. Text summarization branches out, 74–81 (2004)