WikiREVIEW: A Multi-Perspective Review Framework for Automatic Wiki-Style Article Generation

Guo-Biao Zhang, Zhijing Wu*, Tian Lan, Ding-Yuan Liu, Yu-Shi Zhu, Xian-Ling Mao

School of Computer Science and Technology, Beijing Institute of Technology, China
zgb_stubborn@bit.edu.cn, wuzhijing.joyce@gmail.com, lantiangmftby@gmail.com,
3220251327@bit.edu.cn, zhuyushi@bit.edu.cn, maoxl@bit.edu.cn

Abstract

As a knowledge-intensive and challenging task, automatic generation of long-form wiki-style articles has garnered increasing attention from researchers due to its ability to efficiently integrate, organize and present vast amounts of both structured and unstructured knowledge. To the best of our knowledge, most of the existing mainstream state-of-the-art methods for automatic wiki-style article generation typically follow a "one-shot generation" paradigm: given a topic, (1) first generating a structured outline, (2) then independently and in parallel generating the content of each outline chapter in a one-shot using the chapter title and references. However, the core limitation of the paradigm lies in its disregards inter-chapter correlation and lacks post-generation revision and refinement, resulting in content redundancy, weak relevance and logical inconsistency. To address these issues, we propose **WikiREVIEW**, a novel multi-perspective review framework for automatic wiki-style article generation. Specifically, our proposed method introduces multi-perspective experts to review the content of each outline chapter at both chapter and paragraph levels following the initial generation, offering evaluation feedback and continuously refining the numerous deficiencies in the initial long-form article, ultimately achieving high-quality wiki-style article generation. Extensive experimental results on the public English dataset FreshWiki and our own constructed high-quality Chinese dataset ChineseWiki, demonstrate that our proposed WikiREVIEW significantly outperforms existing state-of-the-art automatic wiki-style article generation methods across all automatic evaluation metrics and human evaluation.

1 Introduction

The task of automatically generating wiki-style articles aims to provide readers with comprehensive, reliable, clearly structured and easily understandable information of a given topic, with content that adheres to the stylistic conventions of Wikipedia, posing a significant challenge in Natural Language Processing (NLP) (Sauper and Barzilay 2009; Minguillón et al. 2017; Li et al. 2022; Fan and Gardent 2022; Yang et al. 2025). For instance, given the topic *"Taylor Hawkins"*, the task first generates an outline in the format of *" # Early Life ## Birth and Family # ..."*, and then generates a complete wiki-style article based on the outline and retrieved references. Moreover, owing to its capacity to provide comprehensive, reliable and clearly structured information, the task of automatically generating wiki-style articles holds significant potential for various downstream applications, including knowledge base construction (Tian et al. 2024; Nandy and Bandyopadhyay 2025; Yuan et al. 2025), multi-document summarization (Liu, Wang, and Yuan 2024; Li, Zhang, and Chaturvedi 2025) and AI-assisted writing (Zeng et al. 2024; Kim et al. 2025).

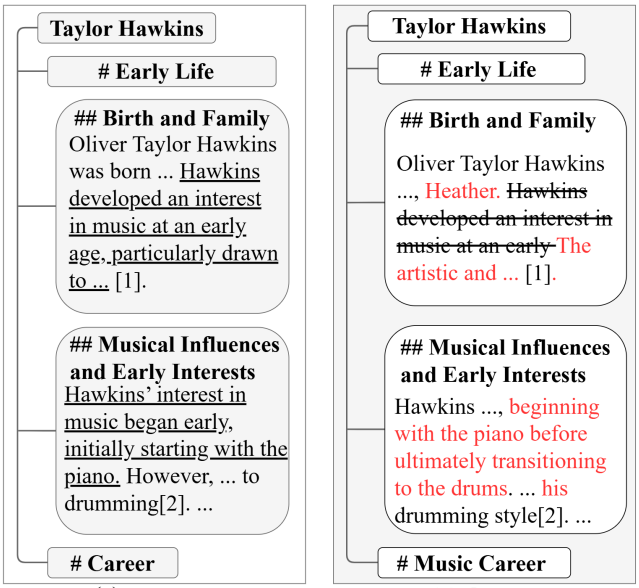

Figure 1: Comparison of wiki-style articles generated by (a) OmniThink and (b) our method. The underlined sentences in (a) highlight redundant content, while the struck-through sentences and red text in (b) indicate the revisions following the multi-perspective review.

Recently, existing methods for automatically generating wiki-style articles have typically combined large language models (LLMs) (Zeng et al. 2022; Achiam et al. 2023; Team et al. 2023; Dubey et al. 2024; Liu et al. 2024) with retrieval-augmented generation (RAG) techniques (Chen et al. 2024;

*Corresponding author

Yang et al. 2024; Xia et al. 2025; Wang et al. 2025), and adopted a "one-shot generation" paradigm. This paradigm first generates an outline based on a given topic, then independently and in parallel generates the content for each chapter using the outline and retrieved references, and finally concatenates these chapters to form a complete article. For example, Storm (Shao et al. 2024), built on the RAG framework, first generates a structured outline for a given topic by leveraging both the internal knowledge of a LLM and information retrieved through multi-perspective conversation (Ram 1991). It then expands the content of each chapter in parallel based on the outline with retrieved references and connects these chapters together to form a complete article. Building on Storm, OmniThink (Xi et al. 2025) introduces an iterative expansion and reflection mechanism to gather a more comprehensive references and construct a more informative outline for a given topic. It similarly expands the chapters in parallel and integrates them to complete the article.

Although the aforementioned methods have made significant progress in automatically generating wiki-style articles, they overlook inter-chapter correlation and lack post-generation content quality review, leading to content redundancy and logical inconsistency (underlined content in Figure 1 (a)), which ultimately undermines the overall quality of articles.

These issues essentially reflect the inadequacy of "one-shot generation" paradigm in simulating the human writing process, particularly the review stage. Notably, Cognitive writing theory (Flower and Hayes 1981) particularly emphasizes that high-quality writing depends on continuous evaluation, feedback, and revision during the "reviewing" stage to ensure logical rigor, structural coherence, and comprehensiveness of the article's content.

Motivated by the cognitive writing theory, we propose **WikiREVIEW**, a novel multi-perspective review framework for automatic wiki-style article generation, to address the limitations of "one-shot generation" methods. Figure 1 shows a comparison of a article on the topic *Taylor Hawkins*, generated by our method and by OmniThink. Specifically, OmniThink often produces articles with redundant information and weak inter-chapter relevance (Figure 1(a)). In contrast, our method introduces a novel multi-perspective review mechanism that employs multi-perspective experts to evaluate the content of each chapter according to predefined evaluation criteria, and provides corresponding feedback. The feedback is then utilized to refine the article's content, reduce content redundancy and improve logical coherence, thereby resulting in high-quality articles (Figure 1(b)). In addition, to further ensure high-quality article generation, we use query rewriting and re-ranking mechanisms to minimize noise during retrieval process in the outline generation phase, and employ an internal-external knowledge fusion approach during the chapter content generation phase, which leverages the extensive prior knowledge of the LLM to mitigate the limited breadth of generated content.

We conduct both automatic and human evaluations of WikiREVIEW on the public FreshWiki dataset our curated ChineseWiki dataset. Extensive experimental results demonstrate that our framework for automatically generating wiki-style articles outperforms existing state-of-the-art methods in both English and Chinese.

In conclusion, our main contributions are as follows:

- We propose WikiREVIEW, a novel framework for automatic wiki-style article generation that simulates the human cognitive writing process. It incorporates a multi-perspective review mechanism following the initial content generation to evaluate feedback and continuously refine the article.
- We employ keyword query optimization and re-ranking mechanisms to reduce noise introduced during retrieval in the outline generation phase. Additionally, we introduce an internal-external knowledge fusion method to alleviate the limited content breadth in the chapter generation stage.
- Extensive experiments on both English and Chinese datasets demonstrate that our proposed WikiREVIEW consistently outperforms existing methods in both automatic and human evaluations.

2 Related Work

With the rapid development of LLMs, recent research has increasingly combined LLMs with RAG to automatically generate high-quality wiki-style articles, achieving significant progress. STORM (Shao et al. 2024) is the first wiki-style article generation method that integrates LLMs with RAG. It generates an outline based on a given topic and references retrieved via multi-perspective question-answering (Lan et al. 2020). The outline is then expanded in parallel using the retrieved references to produce chapters content, which are subsequently form a article. Building upon STORM, Co-STORM (Jiang et al. 2024) introduces an interactive environment that allows users to observe and participate in the writing process (Lan et al. 2025), thereby simplifying complex information retrieval and enhancing both the efficiency and quality of outline construction and article generation. In contrast to the above two methods, OmniThink (Xi et al. 2025) further extends STORM by simulating cognitive behaviors of learners. Through iterative expansion and reflection mechanism (Lan et al. 2024a,b), it progressively deepens topic understanding and broadens the scope of references. The resulting references are then used to parallel expand the outline to generate chapters and integrate them to complete the article. Additionally, WikiAutoGen (Yang et al. 2025) integrates visual and textual content and optimizes retrieval references through a multi-angles self-reflection mechanism, aiming to automatically generate high-quality multi-modal wiki-style articles and introduces the WikiSeek benchmark to evaluate multi-modal knowledge generation on more challenging topics.

3 Method

3.1 Task Definition

The task of automatic wiki-style article generation aims to generate a comprehensive, well-structured, and logically coherent article based on a given topic. Specifically, given a

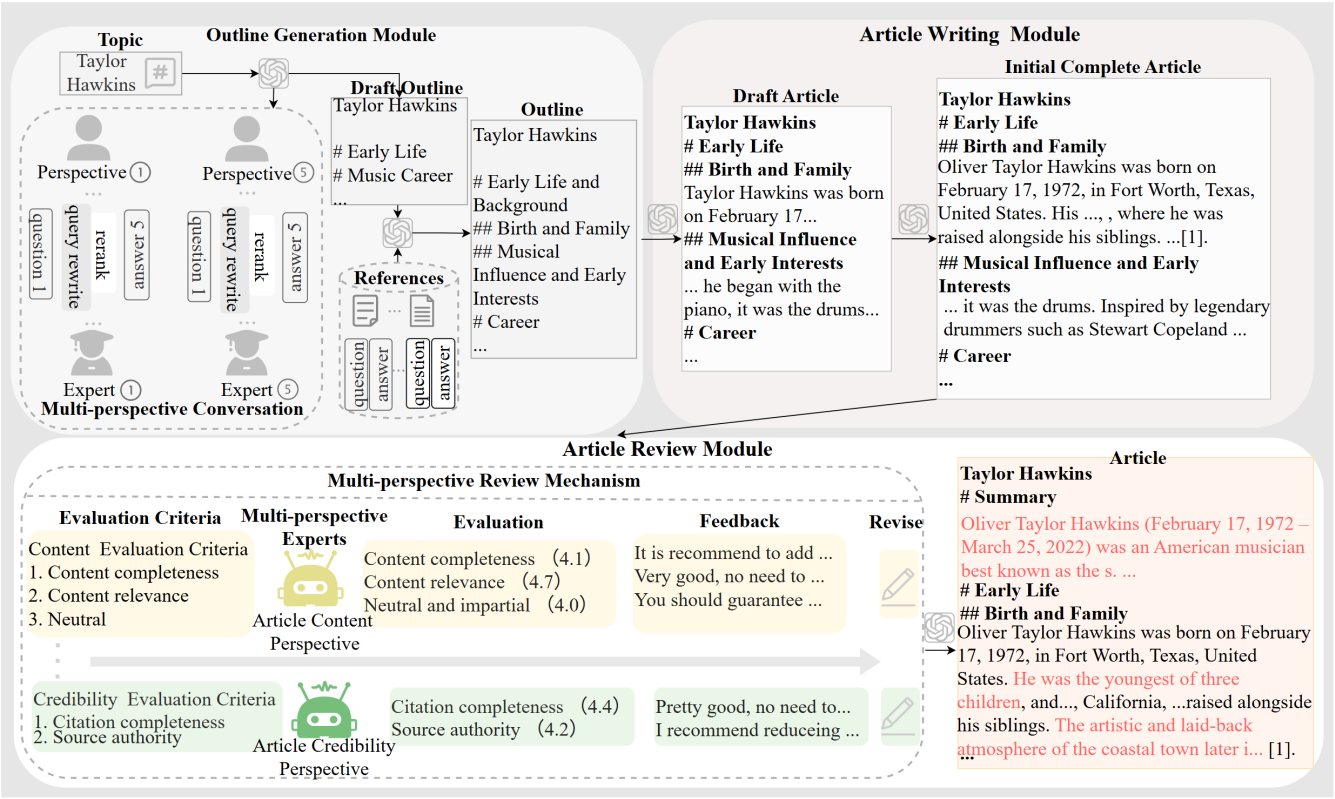

Figure 2: Overview of the **WikiREVIEW** framework. It consists of three main modules: **(1) Outline Generation Module** first generates a draft outline for a given topic and then refines it using references composed of multi-perspective conversation history and retrieved content to produce a complete outline; **(2) Article Writing Module** generates an initial complete article by refining the draft article paragraph by paragraph based on the references; **(3) Article Review Module** incorporates multi-perspective review mechanism, which uses multi-perspective experts to evaluate the articles based on the evaluation criteria, provide feedback to revise the content, and subsequently prompt the LLM to polish and generate a complete wiki-style article.

topic t, a LLM retrieves relevant external references $\mathcal{R}$ using a search engine. An outline $\mathcal{O}$ is then generated based on t and $\mathcal{R}$, and finally, the complete article $\mathcal{A}$ is generated using both the outline $\mathcal{O}$ and the external references $\mathcal{R}$.

3.2 The WikiREVIEW Framework

The overall architecture of our proposed WikiREVIEW framework is shown in Figure 2, comprising three modules: **Outline Generation Module**, which generates an outline for a given topic by leveraging a draft article and references, including multi-perspective conversation history and retrieved web content. **Article Writing Module**, which prompts the LLM to generate a draft article based on the outline, then refines it using relevant references to produce a initial complete article. **Article Review Module**, which introduces multi-perspective experts to evaluate the initial complete article according to predefined evaluation criteria and provide feedback to refine article's content, and prompt the LLM to generate a summary of the complete article.

Outline Generation Module. An outline determines the content direction, structural hierarchy, and logical progression of an article. Therefore, a well-structured outline is crucial for generating high-quality articles. Drawing on the multi-perspective dialogue in previous work (Shao et al. 2024), we prompt a LLM to analyze relevant Wikipedia articles and identify N distinct perspectives $\mathcal{P} = \{p_1, ..., p_N\}$[1] on a given topic t. Each perspective is instantiated as a role-specific wiki-style author who poses an initial question query about the topic. These queries are then refined using a keyword query rewriting strategy to enhance their precision. Each domain expert retrieve answers from web pages based on the refined queries and apply a re-ranking mechanism to select the most relevant answer, which are returned to the respective wiki-style author. Drawing from prior conversation history, the authors deepen their understanding of the topic and pose more detailed follow-up questions. This multi-perspective conversation process continues for a maximum of M turns. Ultimately, we integrate the conversation history from all perspectives with the retrieved web content to form a comprehensive reference $\mathcal{R}$. Using this reference and the initial outline $\mathcal{O}_{\mathcal{D}}$ generated by the LLM for the topic t, we produce a final outline $\mathcal{O} = \text{LLM}(\mathcal{O}_{\mathcal{D}}, \mathcal{R})$ with

[1]Add p_0 to $\mathcal{P}$ as a basic factual perspective, responsible for ensuring broad coverage of basic factual information on the topic.

improved structure and enriched information.

Article Writing Module. The extensive prior knowledge embedded in LLMs forms the cornerstone of their text generation abilities. Unlike previous approaches that rely solely on external knowledge to generate chapter content, we adopt a method that combines the model's internal knowledge with external knowledge to alleviate the issue of limited breadth of generated content. First, given the outline $\mathcal{O}$, the LLM is prompted to generate a diverse and trustworthy draft article $\mathcal{A}_{\mathcal{D}}$. Subsequently, the draft article $\mathcal{A}_{\mathcal{D}}$ and the references $\mathcal{R}$ are jointly utilized to refine the content of each chapter. Specifically, for each chapter, the chapter title and the titles of its hierarchical sub-sections are used to retrieve relevant content from $\mathcal{R}$, based on semantic similarity computed using Sentence-BERT embeddings (Reimers and Gurevych 2019). During the writing of each chapter, the retrieved content will be used to verify, supplement, and revise the chapter content, ensuring both its richness and factual accuracy. Once all chapters content are generated, they are connected to form the initial complete article $\mathcal{A}_{\mathcal{I}} = \text{LLM}(\mathcal{R}, \mathcal{A}_{\mathcal{D}})$.

Article Review Module. Cognitive writing theory (Flower and Hayes 1981; Frisoni et al. 2024) emphasizes that high-quality writing depends on continuous evaluation, feedback, and revision during the "reviewing" stage. Accordingly, we first develop a more comprehensive post-generation evaluation criteria (see Appendix D) based on wikipedia's universal page rating criteria and evaluation criteria (§4.4) in Storm (Shao et al. 2024). Then we introduced a multi-perspective review mechanism following the generation of article $\mathcal{A}_{\mathcal{I}}$. This mechanism adopts multi-perspective experts, including article content perspective, wiki-style perspective, and article credibility perspective, to conduct a comprehensive multi-dimensional evaluations of articles based on the post-generation evaluation criteria. It produces quantitative scores and corresponding feedback $\mathcal{F}$, which are used to revise and enhance the article content.

Specifically, article content perspective expert assess each chapter's content across three dimensions: content completeness, content relevance, and neutral. It then assign a score for each dimension based on the evaluation criteria. If the content completeness score is 4.1, indicating room for improvement, It may suggest adding key facts and details. If the content relevance score is 4.7, it will recommend maintaining the current content without modification. Targeted revisions are then made to the chapter content based on this feedback $\mathcal{F}$. Finally, we prompt the LLM to assess the overall logical flow of the article at a macro level, insert appropriate transition words between chapters when necessary, and generate a summary of the article to obtain a high-quality wiki-style article $\mathcal{A} = \text{LLM}(\mathcal{A}_{\mathcal{I}}, \mathcal{F})$. Similar to the article content perspective expert, the remaining perspective experts follow the same review process.

4 Experiments

4.1 Dataset

We use the publicly available English dataset FreshWiki (Shao et al. 2024) to evaluate the effectiveness of WikiREVIEW. In addition, to verify the generalization of our approach, we curate a high-quality Chinese dataset ChineseWiki. The ChineseWiki dataset construction process and the statistics of both datasets are provided in Appendix A[2].

4.2 Baselines

We compare our approach with all LLM-based baselines, including oRAG, Storm (Shao et al. 2024), Co-STORM (Jiang et al. 2024) and OmniThink (Xi et al. 2025), for automatic wiki-style article generation on both FreshWiki and ChineseWiki datasets. Detailed descriptions of these baselines are provided in Appendix B.

4.3 Implementation Details

We build WikiREVIEW with zero-shot prompting based on the DSPy framework (Khattab et al. 2023), and conducted experiments using Qwen-2.5 (Team 2024), GPT-4 (Achiam et al. 2023), GPT-4o (Hurst et al. 2024), and DeepSeek-R1 (Guo et al. 2025). The same parameter settings were maintained, with a temperature of 1.0 and top_p set to 0.9. And we employed the Serper's API to retrieve real-time web page content, returning 5 web pages per query. Additionally, both hyperparameters *N* and *M* were set as 5. More implementation details are provided in Appendix C.

4.4 Evaluation Metrics

Following previous work (Xi et al. 2025), we employ the Prometheus-7b-v2.0 model (Kim et al. 2024) to evaluate both outline and article quality, which has been shown to have a strong correlation with human preferences. For the outlines, we assessed them across three dimensions: Content Guidance, Hierarchical Clarity and Logical Coherence, using a 1–5 scale rubric developed by OmniThink. For the generated articles, we evaluated them across four dimensions: Relevance, Breadth, Interest Level and Organization, using a 1–5 scale rubric developed by Storm. We also utilize the traditional ROUGE metric (Lin 2004) to evaluate the quality of the generated articles. For article verifiability, we calculate the citation recall and citation precision as defined by ALCE (Gao et al. 2023), and use Mistral-7B-Instruct model (Jiang et al. 2023) to verify whether the cited paragraph contains the corresponding sentence. Additionally, we conducted a human evaluation (§5.3) using a more fine-grained 7-point scale based on the same dimension of article quality. More details for evaluation criteria are provided in Appendix D.

5 Results and Analysis

5.1 Main Results

The results of article evaluation on the FreshWiki and ChineseWiki datasets are presented in Table 1 and Table 2 [3], respectively, indicating several findings: (1) Our method consistently achieves the highest performance across all evalua-

[2] All appendices are available in our code link.

[3] The results presented in Tables 1 and 2 are the averages of three independent experiments, and paired t-tests showed a significant difference between WIkiREVIEW and the best baseline OmniThink ($p < 0.05$).

Backbones	Methods	Rubric Grading				Verifiability		ROUGE-1
		Relevance	Breadth	Interest Level	Organization	CR	CP	
Qwen-2.5	oRAG	3.91	4.08	3.96	3.63	74.21	76.02	44.86
	STORM	4.10	4.19	4.21	3.89	79.76	81.82	48.41
	Co-STORM	4.26	4.22	4.30	4.07	81.61	82.10	48.94
	OmniThink	4.31	4.38	4.33	4.15	81.62	83.73	51.45
	WikiREVIEW (Our)	**4.51**	**4.40**	**4.56**	**4.28**	**83.02**	**83.98**	**54.12**
GPT-4	oRAG	3.97	4.10	4.12	3.72	75.13	75.70	45.19
	STORM	4.16	4.23	4.33	4.01	82.10	83.60	47.89
	Co-STORM	4.34	4.28	4.42	4.10	82.37	83.32	48.01
	OmniThink	4.37	4.43	4.45	4.19	82.71	83.64	53.80
	WikiREVIEW (Our)	**4.60**	**4.44**	**4.62**	**4.35**	**84.61**	**85.86**	**55.86**
GPT-4o	oRAG	4.02	4.10	4.18	3.81	75.72	76.60	45.35
	STORM	4.18	4.29	4.35	4.07	83.28	84.71	48.26
	Co-STORM	4.37	4.26	4.46	4.25	84.01	85.27	48.03
	OmniThink	4.40	4.50	4.51	4.23	84.72	85.83	54.51
	WikiREVIEW (Our)	**4.65**	**4.54**	**4.66**	**4.42**	**85.52**	**86.88**	**54.52**
DeepSeek-R1	oRAG	4.03	4.08	4.17	3.86	75.01	76.40	45.20
	STORM	4.19	4.27	4.30	4.08	80.57	81.52	47.17
	Co-STORM	4.34	4.29	4.37	4.21	81.61	82.01	48.92
	OmniThink	4.37	4.42	4.48	4.18	82.10	83.28	52.39
	WikiREVIEW (Our)	**4.61**	**4.45**	**4.60**	**4.35**	**83.89**	**84.71**	**53.36**

Table 1: Results of article quality evaluation on the FreshWiki dataset. CR and CP denotes Citation Recall and Citation Precision, respectively.

Backbones	Methods	Rubric Grading				Verifiability		ROUGE-1
		Relevance	Breadth	Interest Level	Organization	CR	CP	
GPT-4o	oRAG	3.92	3.96	4.02	3.82	72.81	73.68	41.00
	STORM	4.14	4.22	4.25	3.98	80.52	81.74	45.28
	Co-STORM	4.25	4.30	4.28	4.10	81.79	82.27	47.80
	OmniThink	4.31	4.37	4.32	4.16	82.06	82.88	49.62
	WikiREVIEW (Our)	**4.40**	**4.42**	**4.45**	**4.30**	**83.10**	**84.21**	**51.71**
DeepSeek-R1	oRAG	4.00	4.04	4.10	3.91	75.86	76.30	44.11
	STORM	4.24	4.32	4.32	4.11	81.17	82.75	48.10
	Co-STORM	4.38	4.39	4.40	4.24	82.21	83.36	49.00
	OmniThink	4.47	4.44	4.46	4.30	83.90	84.62	53.48
	WikiREVIEW (Our)	**4.63**	**4.50**	**4.63**	**4.45**	**84.62**	**85.88**	**55.17**

Table 2: Results of article quality evaluation on the ChineseWiki dataset. We selected GPT-4o, which performs best for English wiki-style generation, and DeepSeek-R1, known for its superior performance in Chinese, to generate Chinese wiki-style articles.

tion dimensions on both datasets, based on different backbone models. Among these backbones, GPT-4o demonstrates the best performance in generating Wiki-style articles in English, while DeepSeek-R1 achieves the highest performance for Chinese Wiki-style article generation. (2) WikiREVIEW exhibits a notable advantage in Relevance and Organization. For instance, when using GPT-4o as the backbone for English Wiki-style article generation, it achieved relevance and organization scores of 4.65 and 4.42, respectively, outperforming the strongest baseline OmniThink (4.40/4.23), by 0.25 and 0.19. With DeepSeek-R1 as the backbone for Chinese wiki-style article generation, it attained scores of 4.63 and 4.45, surpassing OmniThink (4.47/4.30) by 0.16 and 0.15, respectively. These results highlight the value of our multi-perspective review mechanism in enhancing the content comprehensiveness, logical rigor and structural coherence of the generated articles. (3) WikiREVIEW also demonstrates advantages in Verifiability and ROUGE metrics, further validating the authenticity and reliability of the generated content. Further details on Verifiability are provided in Appendix E. It is worth noting that LLM evaluators may overestimate machine-generated text, so we supplemented the human evaluation (§5.3).

In addition, we evaluate the quality of the outlines of the generated English wiki-style articles by GPT-4o, with the results presented in Table 3. WikiREVIEW achieves the highest performance across all three evaluation dimensions. This can be primarily attributed to our proposed multi-perspective review mechanism, which continuously refines the content and structure of articles post-generation, thereby indirectly making the outline more logical and better covering the chapter content.

5.2 Ablation Studies

To investigate the importance of the multi-perspective review mechanism as a core component of WikiREVIEW, we conducted an ablation study on the FreshWiki dataset using GPT-4o as the backbone model. Specifically, we compared the performance of the WikiREVIEW framework with its variant, WikiREVIEW w/o R, which the multi-perspective

Method	Content Guidance	Hierarchical Clarity	Logical Coherence
oRAG	3.96	3.92	3.92
STORM	3.98	3.98	3.95
Co-STORM	3.98	3.96	3.97
OmniThink	4.01	4.00	3.97
WikiREVIEW (Our)	**4.03**	**4.01**	**4.01**

Table 3: Results of outline quality evaluation on the FreshWiki dataset.

review mechanism is removed after the initial article generation. As shown in Figure 3, the variant performs significantly worse than WikiREVIEW across all four evaluation metrics, even falling below the best baseline OmniThink. Figure 4 illustrates that it also exhibits varying degrees of decline in the Verifiability and ROUGE metrics, again lagging behind OmniThink. These results strongly affirm the critical role of the multi-perspective review mechanism following initial article generation, and further highlight that outline generation, article drafting, article review, and polishing are tightly interconnected and essential components of the complete writing process.

5.3 Human Evaluation

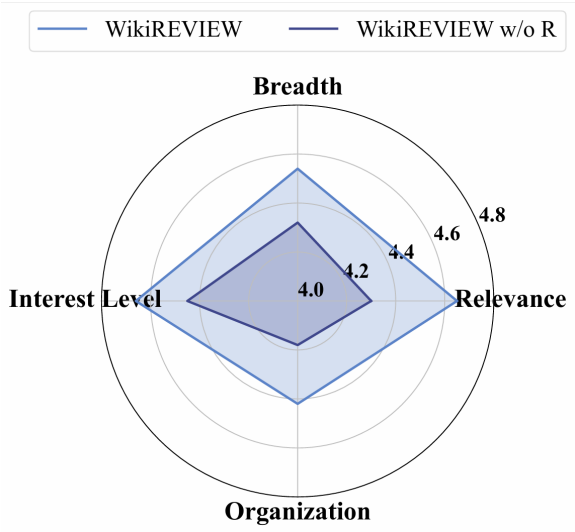

Figure 3: Ablation results of WikiREVIEW on rubric grading (5-point scale).

To further assess the quality of articles generated by WikiREVIEW and gain deeper insights into its strengths and weaknesses, we randomly selected 20 topics from the FreshWiki dataset. Three graduate students (including one doctoral student) from well-known universities were invited to evaluate the articles produced by WikiREVIEW and those generated by the strongest baseline OmniThink. Each evaluator used a 7-point scale to assess the article pairs across the four dimensions in §4.4. The results of the human evaluation are presented in Table 4. Additionally, the three evaluators

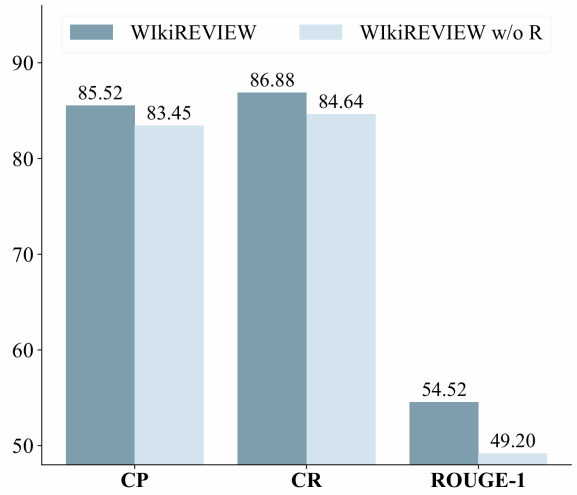

Figure 4: Ablation results of WikiREVIEW on verifiability and ROUGE-1 (100-point scale).

Rubric Grading	OmniThink		WikiREVIEW	
	Average	≥ 5 Rates	Average	≥ 5 Rates
Relevance	5.1	53.3%	**5.8**	**70%**
Breadth	5.2	66.7%	**5.7**	**70%**
Interest Level	4.9	53.3%	**5.1**	**60%**
Organization	5.2	60%	**5.6**	**73.3%**

Table 4: Human evaluation results on 20 pairs of articles generated by WikiREVIEW and OmniThink.

compared the articles generated by WikiREVIEW after the multi-perspective review with the initial articles to evaluate whether the revised sentences were reasonable and reliable. The statistics of the evaluation results are shown in Figure 5.

The results in Table 4 indicate that WikiREVIEW outperforms OmniThink across all evaluation dimensions, with particularly notable improvements in relevance and organization, which increased by 16.7% and 13.3%, respectively. Notably, a discrepancy is observed between the automated and human evaluations with respect to the Interest Level indicator. Human evaluators generally considered the generated articles to be less engaging than suggested by the automated scores. This deviation indicates that current automated evaluation methods may not fully capture human preference. Consequently, developing fine-grained evaluation approaches that better align with human preference is a promising direction in future research. Figure 5 shows that the average approval rate among the three evaluators is approximately 85%, indicating that the sentence generated by WikiREVIEW using the multi-perspective review mechanism is considered reliable and of higher quality in most cases. Furthermore, the evaluators noted that the remaining sentences were not unreasonable, but rather showed no significant differences between the modified and unmodified versions. More human evaluation details are provided in Appendix E.

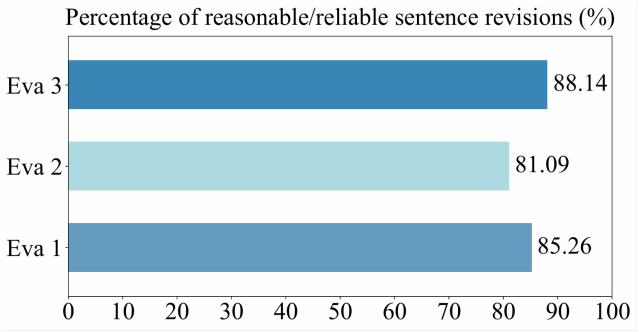

Figure 5: Results of human reliability assessment on the revised sentences.

5.4 Necessity Analysis of Multi-perspective Review Mechanism

As demonstrated by the ablation results in §5.2, the multi-perspective review mechanism plays a critical role in the overall article generation process. To further verify its necessity from another perspective, we randomly selected four distinct topic types in FreshWiki and then counted the defective chapters and sentences in the articles generated by OmniThink according to the post-generation evaluation criteria (see Appendix D). As seen in the Figure 6, each article contains defective chapters, and all four include chapters with up to five defective sentences. Notably, the article *Silicon Valley Bank* has as many as six defective chapters. Furthermore, analyzing all topics in the FreshWiki dataset, we found that each article contains an average of 5.1 defective chapters, with approximately 2.6 sentences per defective chapter requiring improvement. These findings further underscore the necessity of multi-perspective review mechanism after initial article generation, as it constitutes a critical step in producing high-quality wiki-style articles.

5.5 Hallucination Analysis of WikiREVIEW

Although WikiREVIEW integrates LLMs and RAG frameworks to automatically generate long wiki-style articles and mitigate hallucinations, LLMs still pose risks when producing extended text. These risks include generating fictitious information not grounded in retrieved sources and deviating from accurate content due to failure to strictly adhere to retrieval results. To empirically evaluate WikiREVIEW's effectiveness in mitigating these risks and verify whether the multi-view review mechanism arbitrarily modifies content, we adopt RAGAs (Es et al. 2024), a reference-free evaluation framework designed to assess RAG-based systems. RAGAs evaluates performance across three dimensions: faithfulness (i.e., whether the generated content is strictly grounded in the retrieved context), answer relevance (i.e. whether the answer appropriately addresses the actual question), and context relevance (i.e., whether the retrieved content is sufficiently focused and pertinent).

The results of the RAGAs evaluation are presented in Table 5. WikiREVIEW outperforms the OmniThink across all three dimensions, achieving a high faithfulness score of 91.71. This performance is primarily attributed to WikiREVIEW's multi-perspective review mechanism, which incorporates the article credibility perspective to review the credibility of the content after initial generation, thereby effectively reducing hallucinations during the generation process. In addition, both WikiREVIEW and OmniThink exhibit lower performance in context relevance compared to the other evaluation dimensions, highlighting the need to improve the quality of retrieved content in future research.

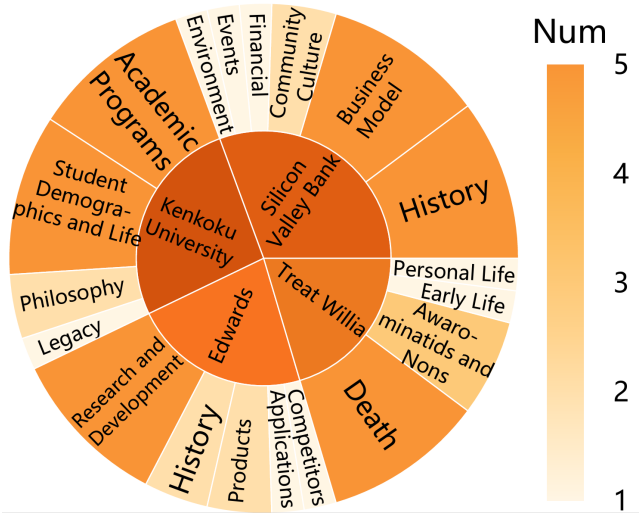

Figure 6: Statistics of defective chapters and sentences in four randomly selected articles

Methods	Faithfulness	Answer Relevance	Context Relevance
OmniThink	88.52	80.19	78.28
WikiREVIEW	**91.71**	**83.65**	**78.96**

Table 5: Results of RAGAs Hallucination Evaluation.

6 Conclusion and Future Work

In this paper, we propose a novel multi-perspective review framework for automatic wiki-style article generation, named WikiREVIEW. Unlike previous "one-shot generation" approaches, WikiREVIEW simulates the human cognitive writing process and introduces a multi-perspective review mechanism that conducts chapter and paragraph-level reviews following the initial article generation, continuously revising and refining the article content to generate high-quality wiki-style articles. Experiments on the FreshWiki and ChineseWiki datasets demonstrate that WikiREVIEW significantly outperforms existing state-of-the-art methods in both automatic evaluation metrics and human evaluation.

In the future, we aim to explore more advanced RAG frameworks and retrieval methods to improve the accuracy of information retrieval during the writing process. Furthermore, we plan to investigate multimodal wiki-style articles generation and to incorporate real-time review and dynamic iterative revision methods into the generation process.

Acknowledgments

This work was supported by the National Key Research and Development Program of China (No. 2024YFF0908200), the National Natural Science Foundation of China (No. 62402043, 62172039, 62572056, 62302040, U21B2009, and 62276110), the China Postdoctoral Science Foundation (No. 2022TQ0033), and the Beijing Institute of Technology Research Fund Program for Young Scholars.

References

Achiam, J.; Adler, S.; Agarwal, S.; Ahmad, L.; Akkaya, I.; Aleman, F. L.; Almeida, D.; Altenschmidt, J.; Altman, S.; Anadkat, S.; et al. 2023. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*.

Chen, J.; Lin, H.; Han, X.; and Sun, L. 2024. Benchmarking large language models in retrieval-augmented generation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, 17754–17762.

Dubey, A.; Jauhri, A.; Pandey, A.; Kadian, A.; Al-Dahle, A.; Letman, A.; Mathur, A.; Schelten, A.; Yang, A.; Fan, A.; et al. 2024. The llama 3 herd of models. *arXiv e-prints*, arXiv–2407.

Es, S.; James, J.; Anke, L. E.; and Schockaert, S. 2024. Ragas: Automated evaluation of retrieval augmented generation. In *Proceedings of the 18th Conference of the European Chapter of the Association for Computational Linguistics: System Demonstrations*, 150–158.

Fan, A.; and Gardent, C. 2022. Generating Biographies on Wikipedia: The Impact of Gender Bias on the Retrieval-Based Generation of Women Biographies. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics*, 8561–8576.

Flower, L.; and Hayes, J. R. 1981. A cognitive process theory of writing. *College Composition & Communication*, 32(4): 365–387.

Frisoni, G.; Cocchieri, A.; Presepi, A.; Moro, G.; and Meng, Z. 2024. To generate or to retrieve? on the effectiveness of artificial contexts for medical open-domain question answering. *arXiv preprint arXiv:2403.01924*.

Gao, T.; Yen, H.; Yu, J.; and Chen, D. 2023. Enabling large language models to generate text with citations. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, 6465–6488.

Guo, D.; Yang, D.; Zhang, H.; Song, J.; Zhang, R.; Xu, R.; Zhu, Q.; Ma, S.; Wang, P.; Bi, X.; et al. 2025. Deepseek-r1: Incentivizing reasoning capability in llms via reinforcement learning. *arXiv preprint arXiv:2501.12948*.

Hurst, A.; Lerer, A.; Goucher, A. P.; Perelman, A.; Ramesh, A.; Clark, A.; Ostrow, A.; Welihinda, A.; Hayes, A.; Radford, A.; et al. 2024. Gpt-4o system card. *arXiv preprint arXiv:2410.21276*.

Jiang, A. Q.; Sablayrolles, A.; Mensch, A.; Bamford, C.; Chaplot, D. S.; Casas, D. d. l.; Bressand, F.; Lengyel, G.; Lample, G.; Saulnier, L.; et al. 2023. Mistral 7B. *arXiv preprint arXiv:2310.06825*.

Jiang, Y.; Shao, Y.; Ma, D.; Semnani, S. J.; and Lam, M. S. 2024. Into the unknown unknowns: Engaged human learning through participation in language model agent conversations. *arXiv preprint arXiv:2408.15232*.

Khattab, O.; Singhvi, A.; Maheshwari, P.; Zhang, Z.; Santhanam, K.; Vardhamanan, S.; Haq, S.; Sharma, A.; Joshi, T. T.; Moazam, H.; et al. 2023. Dspy: Compiling declarative language model calls into self-improving pipelines. *arXiv preprint arXiv:2310.03714*.

Kim, J.; Yu, S.; Detrick, R.; and Li, N. 2025. Exploring students' perspectives on generative AI-assisted academic writing. *Education and Information Technologies*, 30(1): 1265–1300.

Kim, S.; Suk, J.; Longpre, S.; Lin, B. Y.; Shin, J.; Welleck, S.; Neubig, G.; Lee, M.; Lee, K.; and Seo, M. 2024. Prometheus 2: An open source language model specialized in evaluating other language models. *arXiv preprint arXiv:2405.01535*.

Lan, T.; Mao, X.-L.; Wei, W.; Gao, X.; and Huang, H. 2020. Pone: A novel automatic evaluation metric for open-domain generative dialogue systems. *ACM Transactions on Information Systems (TOIS)*, 39(1): 1–37.

Lan, T.; Xu, C.; Lv, Z.; Dong, Q.; Zhang, J.; Huang, H.; Yang, M.; and Hu, B. 2025. Bridging the Gap Between Data Distribution and Model: Dynamic Data Distribution Optimization for Improving Critique Capabilities of Large Language Models. *Expert Systems with Applications*, 129878.

Lan, T.; Zhang, W.; Lyu, C.; Li, S.; Xu, C.; Huang, H.; Lin, D.; Mao, X.-L.; and Chen, K. 2024a. Training language models to critique with multi-agent feedback. *arXiv preprint arXiv:2410.15287*.

Lan, T.; Zhang, W.; Xu, C.; Huang, H.; Lin, D.; Chen, K.; and Mao, X.-l. 2024b. CriticEval: Evaluating large language model as critic. *arXiv preprint arXiv:2402.13764*.

Li, H.; Zhang, R.; and Chaturvedi, S. 2025. Improving Fairness of Large Language Models in Multi-document Summarization. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, 1143–1154.

Li, I.; Fabbri, A. R.; Kawamura, R.; Liu, Y.; Tang, X.; Tae, J.; Shen, C.; Ma, S.; Mizutani, T.; and Radev, D. 2022. Surfer100: Generating Surveys From Web Resources, Wikipedia-style. In *LREC*.

Lin, C.-Y. 2004. ROUGE: A package for automatic evaluation of summaries. In *Text Summarization Branches Out*, 74–81.

Liu, A.; Feng, B.; Xue, B.; Wang, B.; Wu, B.; Lu, C.; Zhao, C.; Deng, C.; Zhang, C.; Ruan, C.; et al. 2024. Deepseek-v3 technical report. *arXiv preprint arXiv:2412.19437*.

Liu, Y.; Wang, Z.; and Yuan, R. 2024. Querysum: A multi-document query-focused summarization dataset augmented with similar query clusters. In *Proceedings of the AAAI Conference on Artificial Intelligence*, 18725–18732.

Minguillón, J.; Lerga, M.; Aibar, E.; Lladós-Masllorens, J.; and Meseguer-Artola, A. 2017. Semi-automatic generation of a corpus of Wikipedia articles on science and technology. *Profesional de la información*, 26(5): 995–1005.

Nandy, A.; and Bandyopadhyay, S. 2025. Language models of code are few-shot planners and reasoners for multi-document summarization with attribution. In *Proceedings of the AAAI Conference on Artificial Intelligence*, 24930–24938.

Ram, A. 1991. A theory of questions and question asking. *Journal of the Learning Sciences*, 1(3-4): 273–318.

Reimers, N.; and Gurevych, I. 2019. Sentence-BERT: Sentence embeddings using siamese BERT-Networks. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing*, 3982–3992.

Sauper, C.; and Barzilay, R. 2009. Automatically generating wikipedia articles: A structure-aware approach. In *Proceedings of the Joint Conference of the 47th Annual Meeting of the ACL and the 4th International Joint Conference on Natural Language Processing of the AFNLP*, 208–216.

Shao, Y.; Jiang, Y.; Kanell, T.; Xu, P.; Khattab, O.; and Lam, M. 2024. Assisting in writing Wikipedia-like articles from scratch with large language models. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics*, 6252–6278.

Team, G.; Anil, R.; Borgeaud, S.; Alayrac, J.-B.; Yu, J.; Soricut, R.; Schalkwyk, J.; Dai, A. M.; Hauth, A.; Millican, K.; et al. 2023. Gemini: A family of highly capable multimodal models. *arXiv preprint arXiv:2312.11805*.

Team, Q. 2024. Qwen2.5: A Party of Foundation Models.

Tian, L.; Ziao, M.; Yanghao, Z.; Chen, X.; and Xianling, M. 2024. A Survey of Automatic Evaluation on the Quality of Generated Text. In *Proceedings of the 23rd Chinese National Conference on Computational Linguistics (Volume 2: Frontier Forum)*, 169–196.

Wang, Y.; Zhang, H.; Pang, L.; Guo, B.; Zheng, H.; and Zheng, Z. 2025. MaFeRw: Query rewriting with multi-aspect feedbacks for retrieval-augmented large language models. In *Proceedings of the AAAI Conference on Artificial Intelligence*, 25434–25442.

Xi, Z.; Yin, W.; Fang, J.; Wu, J.; Fang, R.; Zhang, N.; Yong, J.; Xie, P.; Huang, F.; and Chen, H. 2025. OmniThink: Expanding knowledge boundaries in machine writing through thinking. *arXiv preprint arXiv:2501.09751*.

Xia, Y.; Zhou, J.; Shi, Z.; Chen, J.; and Huang, H. 2025. Improving retrieval augmented language model with self-reasoning. In *Proceedings of the AAAI conference on artificial intelligence*, 25534–25542.

Yang, X.; Sun, K.; Xin, H.; Sun, Y.; Bhalla, N.; Chen, X.; Choudhary, S.; Gui, R.; Jiang, Z.; Jiang, Z.; et al. 2024. Crag-comprehensive rag benchmark. *Advances in Neural Information Processing Systems*, 37: 10470–10490.

Yang, Z.; Chen, J.; Xu, D.; Fei, J.; Shen, X.; Zhao, L.; Feng, C.-M.; and Elhoseiny, M. 2025. WikiAutoGen: Towards Multi-Modal Wikipedia-Style Article Generation. *arXiv preprint arXiv:2503.19065*.

Yuan, D.; Zhou, S.; Chen, X.; Wang, D.; Liang, K.; Liu, X.; and Huang, J. 2025. Knowledge graph completion with relation-aware anchor enhancement. In *Proceedings of the AAAI Conference on Artificial Intelligence*, 15239–15247.

Zeng, A.; Liu, X.; Du, Z.; Wang, Z.; Lai, H.; Ding, M.; Yang, Z.; Xu, Y.; Zheng, W.; Xia, X.; et al. 2022. Glm-130b: An open bilingual pre-trained model. *arXiv preprint arXiv:2210.02414*.

Zeng, Z.; Sha, L.; Li, Y.; Yang, K.; Gašević, D.; and Chen, G. 2024. Towards automatic boundary detection for human-ai collaborative hybrid essay in education. In *Proceedings of the AAAI Conference on Artificial Intelligence*, 22502–22510.