WikiMAG: A Multi-Agent Guided Framework for Generating Structured Wikipedia-like Articles

Xiuli Kang[1,2], Yinlong Xiao[2], Minghao Hu[2*], Yuan Huang[1*], Bin Mao[2], Ming Wang[3], Fang Wang[4] Zhunchen Luo[2], Wei Luo[2], Guotong Geng[2]

[1]Hebei University of Engineering, Handan, China
[2]Center of Information Research, AMS, Beijing, China
[3]North China University of Technology, Beijing, China
[4]Peking University, Beijing, China
humh573@163.com, kangxiuli214@gmail.com

Abstract

Wikipedia serves as the world's largest and most popular online reference encyclopedia, rich in structured knowledge and authoritative citations. Recently, numerous works have leveraged large language models to automatically generate Wikipedia-like articles. However, existing approaches primarily focus on producing singular narrative-type content, overlooking higher information-density structured elements such as *timeline* and *table*. To address these limitations, we propose **WikiMAG**, a multi-agent guided framework for generating structured Wikipedia-like articles. This framework employs a collaborative multi-agent mechanism to orchestrate the creation process, featuring three synergistic core components: **Progressive planner** first constructs the coarse-grained outline framework and then annotate fine-grained types for outline units, encompassing *narrative*, *timeline*, and *table* formats; **Reflective inspector** dynamically curates high-quality references via multi-round interactive feedback, thereby enhancing the authority and relevance of citations; **Versatile writer** integrates fine-grained outline details and high-quality reference information to generate information-rich articles, incorporating the three annotated formats. We evaluate WikiMAG on two public datasets, FreshWiki and WikiGenBen, across *outline*, *writing*, and *verifiability* dimensions. Compared with the best baseline method, our method achieves an average improvement of 6.73 points and 4.39 points in Heading Soft Recall and the METEOR metric (a machine translation and text generation evaluation metric) respectively, and an average increase of 16.84 percentage points in Citation Rate.

Introduction

As the world's most influential multilingual collaborative knowledge platform, Wikipedia stands as a paragon of knowledge structuring through its content organization paradigm. It takes a standardized chapter system as the framework, extracts key attributes via infoboxes, and constructs chronological contexts in conjunction with timeline. These elements integrate to form a multi-dimensional knowledge network, which not only demands content to be authoritative, comprehensive, and accurate but also emphasizes the hierarchy and logical consistency of knowledge presentation. Therefore, developing technical solutions capable of automatically generating high-quality articles that meet Wikipedia's standards has become a critical issue awaiting breakthrough in the field of knowledge engineering.

Method	Plan	Retr.	MR-Int	Refl.	Struct.
AgentWrite	✓	×	×	×	×
PP	×	✓	×	×	×
RPP	✓	✓	×	×	×
CO-STORM	✓	✓	✓	×	×
Ours	✓	✓	✓	✓	✓

Table 1: Comparison of functional support of different methods in key stages including Plan, Retrieval(Retr.), Multi-round Interaction(MR-Int), Reflection(Refl.), and Timeline and Table(Struct.).

With the emergence of advanced large language models such as GPT-4(Achiam et al. 2023), LLaMA(Touvron et al. 2023), and Gemini(Team et al. 2023), significant progress has been made in the field of long-text generation, demonstrating strong application potential in scenarios such as news writing, novel creation, and academic abstracts. Specifically, AgentWrite(Bai et al. 2024) simulates human thinking processes through a segmented writing strategy, effectively breaking through the context window limitations of traditional models; WikiGenBen(Zhang et al. 2024) compares two text generation strategies, the PP (Retrieve-then-Read) strategy generates the complete text in one go by chunking the input reference, while the RPP (Plan-Retrieve-Read) strategy, building on PP, incorporates a planning mechanism, which generates content incrementally after completing the global planning of the text structure; CO-STORM(Jiang et al. 2024) continuously revises and expands information through multi-round dialogue to optimize text quality; Document(Liang et al. 2024) deeply embeds planning capabilities into the model architecture, achieving high-quality single-round long-text generation.

*Corresponding author.

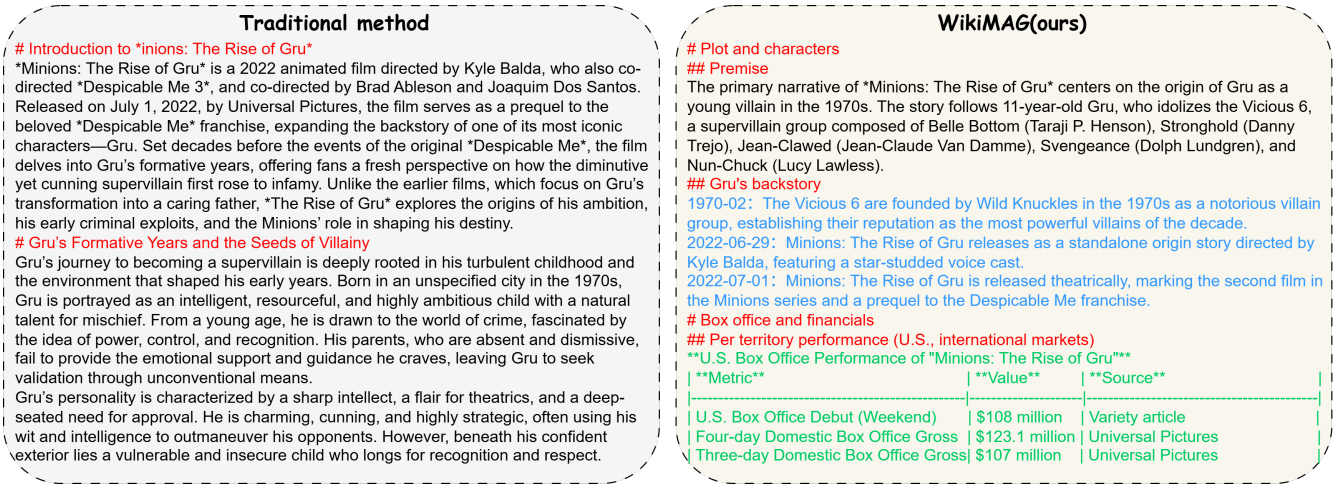

Figure 1: Comparison between our method and existing methods. The WikiMAG multi-agent framework enables three types of structured content in text generation: narrative, timeline, and table, where red represents the outline, black represents the narrative type, blue represents the timeline type, and green represents the table type.

However, existing technologies pay insufficient attention to the core structural needs of encyclopedia-style articles, particularly neglecting the standardized presentation of structured information such as timeline and table, and thus fail to conduct targeted optimizations for Wikipedia's unique formatting requirements.

To generate structured information that meets the required standards, it is necessary not only to properly handle the coordination between different types of generated information but also to address the challenge of ensuring the reliability of information sources. To this end, this paper proposes **WikiMAG** framework, with Table 1 illustrating the core differences between this framework and other methods. WikiMAG constructs a multi-agent collaborative working model, where a planner coordinates the generation of different content types and a labeling mechanism distinguishes between these types. Meanwhile, an inspector accurately eliminates redundant and irrelevant content while fully considering the reliability of information sources. Additionally, a writer, aided by the three content type labels, composes the corresponding content using reference materials. This method can specifically meet the structural requirements of Wikipedia and generate high-quality content that complies with relevant norms. As shown in Figure 1, it clearly presents a comparison between the texts generated by the method in this paper and those generated by existing methods.

Our contributions are summarized as follows:

- We propose **WikiMAG**, a novel multi-agent guided framework specifically designed to generate structured Wikipedia-like articles by effectively integrating higher information-density structured elements (timeline and table).
- We introduce three synergistic core components: a **progressive planner** for constructing coarse-grained outlines and annotating fine-grained types (narrative, timeline, and table); a **reflective inspector** for dynamically curating high-quality references via multi-round feedback; and a **versatile writer** for generating the final content in the three annotated formats based on outline details and curated references.
- Extensive experiments on the FreshWiki and WikiGen-Ben datasets, evaluated across outline, writing, and verifiability dimensions, show that compared with the best baseline, WikiMAG achieves an average improvement of 6.73 points in Heading Soft Recall, 4.39 points in METEOR, and 16.84 percentage points in Citation Rate.

Related Works

Retrieval-Augmented Text Generation Retrieval mechanisms can effectively compensate for the inherent limitations of large language models (LLMs) in knowledge retention and real-time updates. Early methods such as REALM(Guu et al. 2020) and DrQA(Chen et al. 2017) relied on pre-built knowledge bases for auxiliary generation but suffered from the problem of delayed knowledge updates. As research progressed, the retrieval-augmented generation (RAG) paradigm, which mines information from training data, gradually highlighted its value. In terms of technological evolution, the early AgentWrite(Bai et al. 2024) adopted an architecture that simply concatenated retrieval systems with generation models, while dynamic retrieval methods like DRAGON(Toro et al. 2024) significantly reduced factual errors by introducing external information to assist generation. Since then, more deeply integrated methods have continued to emerge: knowledge graph-based models such as DynaGRAG(Radford et al. 2018) and MindSearch(Chen et al. 2024b) dynamically acquire knowledge through entity relationships to generate logically rigorous texts; ConTReGen(Roy et al. 2024) realizes the structured output of multi-source information through a tree-like framework; Selfmem(Cheng et al. 2023)

and RQ-RAG(Chan et al. 2024) further improve model performance through self-memory mechanisms and query optimization strategies, respectively. To address the issue that current retrieved information may be irrelevant to the target task, this paper proposes a solution based on agent interaction.

Automatic Wikipedia Generation The core of the Wikipedia generation task lies in creating high-quality content that meets encyclopedic standards. Researchers have effectively overcome the limitations of model context windows through phased generation strategies(Bai et al. 2024; Wan et al. 2025), while combining instruction fine-tuning(Liu et al. 2023) quality and accuracy of generated texts. At the domain adaptation level, models pre-trained on Wikipedia corpora(Lewis et al. 2019) and structured generation methods(Slobodkin et al. 2024) have jointly optimized the professionalism of text styles; retrieval-augmented technologies(Shao et al. 2024) and interactive generation(Jiang et al. 2024) have further enriched content dimensions and enhanced reliability. Although the knowledge integration framework of RAPID(Gu et al. 2025) has made progress in improving style consistency and content authority, related issues remain unresolved. Unlike existing methods, we incorporate timeline and table elements by means of content type tagging to enhance the structural diversity of generated Wikipedia content, which provides an important direction for future research in this field.

WikiMAG

To address the issue of diverse content types in Wikipedia article creation, WikiMAG proposes a structured generation mechanism based on multi-agent collaboration. Specifically, as shown in Figure 2, the framework relies on the Progressive Planner to generate outlines and mark content types, uses the Reflective Inspector to eliminate redundant and low-relevance reference, and leverages the Versatile Writer to complete content composition according to different content label types, thereby achieving accurate generation of multi-type structured content.

Progressive Planner

When a single model simultaneously undertakes the dual tasks of long-range knowledge planning and step-by-step content generation, it often struggles due to excessive cognitive load, which easily leads to fragmented content logic or inconsistent quality. This issue is particularly prominent in the generation of long texts such as Wikipedia entries. Meanwhile, the current text composition mode using multi-model collaboration generally has the drawback of lacking refined consideration for diverse content types. To address these issues, we adopt a progressive approach to construct outlines and label content types.

Coarse-grained Outline Creation Based on Wikipedia's well-established section system, we adopt a coarse-grained construction method to generate hierarchical multi-level outlines, providing a structural framework for the creation of encyclopedia entries. Specifically, for the input entry t, we obtain relevant information through coarse-grained web retrieval, and then use the BGE model (Chen et al. 2024a) to perform semantic similarity calculation and sorting on the retrieval results, obtaining an ordered dataset $D = \{D_1, D_2, \cdots, D_n\}$. Subsequently, we analyze the information related to t in the dataset, sort out the thematic units therein, and further construct the outline $O = \{O_1, O_2, \cdots, O_n\}$.

$$O = \text{Coarse}(t, \text{Sort}_{\text{BGE}}(\text{Retrieve}(t))) \tag{1}$$

where $\text{Retrieve}(t)$ refers to the coarse-grained web retrieval for entry t and the output of original retrieved information; $\text{Sort}_{\text{BGE}}(\cdot)$ denotes the output of the ordered dataset D sorted by the BGE model; and $\text{Coarse}(t, \cdot)$ is a function that generates the outline O based on entry$t$ and dataset D.

Fine-grained Outline Labeling To address the issue of a single content type, we perform fine-grained processing on the outline O and divide the custom labeling system $T = \{T_1, T_2, T_3\}$ into three categories: T_1 corresponds to the narrative type, T_2 denotes the timeline type, and T_3 represents the table type, with optimization achieved through classification and labeling. Taking O as the basic content unit, we rely on 3 voters to respectively conduct semantic association and matching between each outline unit O_k and the labeling system T. Through dynamic decision-making, a probability mapping relationship between O_k and T_j ($j = 1, 2, 3$) is constructed. Finally, by counting the votes $\text{Count}(T_j \mid O_k)$, the category with the highest number of votes is determined as the labeling result:

$$\text{Label}(O_k) = \arg \max_{j=1,2,3} \text{Count}(T_j \mid O_k) \tag{2}$$

Reflective Inspector

In the process of knowledge integration and content generation, valid and accurate reference is an indispensable reference support. However, online retrieval is generally plagued by issues such as information noise interference, missing key details, and uneven quality of reference, which greatly affect the reliability of reference materials. To address this problem, we achieve dynamic optimization of the reference set through the synergistic effect of retrieval optimization and reference feedback, thereby enabling the accurate identification of high-value reference resources.

Hunting References This module is responsible for analyzing feedback information, dynamically generating and optimizing retrieval strategies, and systematically constructing reference resources. It invokes the retrieval program to perform initial retrieval based on the topic t and outline O, while each subsequent round of retrieval generates a new retrieval instruction Q through the retrieval reconstruction function based on the feedback information.

In the operation of the multi-round iterative feedback mechanism, it continuously improves the candidate reference set $\mathcal{C}$ through dynamic accumulation and optimization of reference resources:

$$\mathcal{C} = \mathcal{C} \cup \mathcal{C}'_m \tag{3}$$

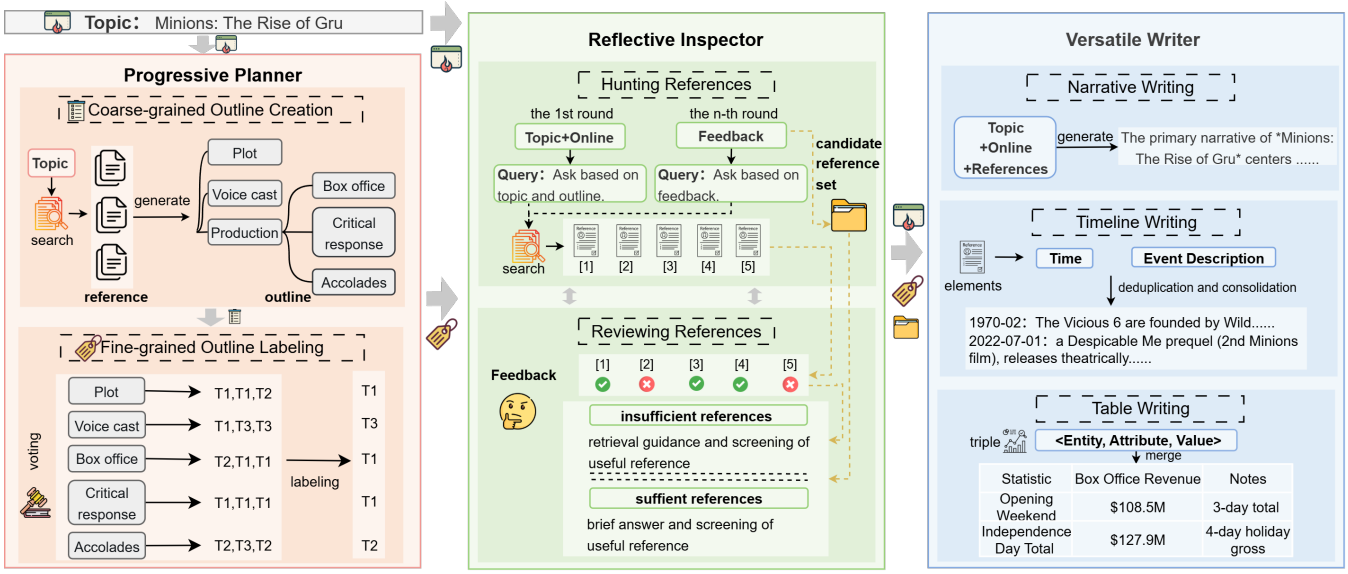

Figure 2: Overview of the WikiMAG Framework. The framework adopts a collaborative workflow involving three core components: a) Progressive Planner generates an outline with type labelings based on the input topic; b) Reflective Inspector retrieves and filters high-quality reference according to the topic and outline; c)Versatile Writer integrates the reference and topic outline to produce the final text. Meanwhile, the outline nodes use three standard labelings: narrative type (T_1) for continuous text, timeline type (T_2) for handling chronologically ordered events, and table type (T_3) for organizing structured data.

where $\mathcal{C}'_m = \{C'_1, C'_2, \cdots, C'_n\}$ represents the new references to be incorporated after feedback evaluation in the m-th iteration. By continuously updating $\mathcal{C} = \{C_1, C_2, \cdots, C_n\}$, the dynamic management of references is achieved.

Reviewing References A reflection mechanism is introduced to review retrieved reference and eliminate irrelevant documents. It conducts reference information evaluation for the query instruction Q based on the reference R obtained from each round of web retrieval and the saved candidate dataset $\mathcal{C}$. When the information is insufficient to answer the query, it will output retrieval-guiding feedback $\mathcal{F}_m$, while screening and retaining the reference $\mathcal{C}'_m$ for subsequent steps:

$$\mathcal{F}_m, \mathcal{C}'_m = \text{Insufficiency}(\mathcal{C}, Q) \tag{4}$$

On the contrary, if the current reference information is sufficient to answer the query, it will generate an answer $\mathcal{A}$ and attach a reference $\mathcal{R}^{\text{ref}} = \{R_1, R_2, \cdots, R_n\}$:

$$\mathcal{A}, \mathcal{R}^{\text{ref}} = \text{Sufficiency}(\mathcal{C}, Q) \tag{5}$$

where $\text{Insufficiency}(\cdot)$ and $\text{Sufficiency}(\cdot)$ are information review functions executed by the reflection mechanism. The two functions perform knowledge evaluation and output through the interaction between the reference set and the query instruction, corresponding to the judgments of insufficient information and sufficient information, respectively.

Versatile Writer

Based on the structured guidance provided by the labelings of outline node types (T_1, T_2, T_3), the Versatile Writer takes the candidate reference set $\mathcal{C}_n$ and core topic t as the foundation, and collaborates with various content processing modules to carry out coordinated operations, achieving the organic integration and standardized output of different content types.

Narrative Writing For the outline node O_k labeled as T_1 (narrative type), it is necessary to rely on high-quality reference resources in $\mathcal{C}_n$ verified through multiple rounds of iteration, combine with the topic t, and use the text generation function $\text{Generate}(\cdot)$ to conduct semantic-level organic integration and logical reconstruction of scattered knowledge units. By in-depth analysis of the conceptual connections and argumentative logic in the references, structured knowledge is transformed into a coherent text $\mathcal{T}$ that conforms to academic norms and expression habits. Its generation relationship can be expressed as:

$$\mathcal{T} = \text{Generate}(O_k, \mathcal{C}_n, t) \tag{6}$$

Timeline Writing For the outline node labeled as T_2 (timeline type) O_k, its core objective is to perform fine-grained information extraction. Specifically, timeline elements are first extracted sequentially from each document in the candidate reference set $\mathcal{C}_n$. Then, for the same time point, redundant event descriptions are eliminated and consolidated using the cosine similarity function Sim. This process can be expressed as:

$$\mathcal{T} = \text{Sort}\left(\text{Remove}_{\text{Sim}(e_i,e_j)<\theta}(E)\right) \tag{7}$$

where $E = \{E_1, E_2, \cdots, E_n\}$ denotes the set of raw events, θ is the similarity threshold, Remove represents the

Algorithm 1: WikiMAG Framework

Require: Topic t, Content types $T = \{T_1, T_2, T_3\}$, Max iterations M
Ensure: Structured Wikipedia article $\mathcal{D}$

Initialization: $\mathcal{C} \leftarrow \emptyset$, $\mathcal{D} \leftarrow \emptyset$
$O \leftarrow \text{Coarse}(t, \text{Sort}_{\text{BGE}}(\text{Retrieve}(t)))$
for each O_k in O **do**
  $\text{Label}(O_k) \leftarrow \arg\max_{j=1,2,3} \text{Count}(T_j \mid O_k)$
  for $m \leftarrow 1$ to M **do**
    $Q \leftarrow \text{GenerateQuery}(t, O, \mathcal{F}_{m-1})$
    $\mathcal{C}_m \leftarrow \text{Retrieve}(Q)$
    if $\text{NeedMoreRef}(\mathcal{F}_m)$ **then**
      $\mathcal{F}_m, \mathcal{C}'_m \leftarrow \text{Insufficiency}(\mathcal{C}_m, Q)$
      $\mathcal{C} \leftarrow \mathcal{C} \cup \mathcal{C}'_m$
    else
      $\mathcal{A}, \mathcal{R}^{\text{ref}} \leftarrow \text{Sufficiency}(\mathcal{C}_m, Q)$
      $\mathcal{C} \leftarrow \mathcal{C} \cup \mathcal{R}^{\text{ref}}$
      break
    end if
  end for
  if $\text{Label}(O_k) = T_1$ **then**
    $\mathcal{T} \leftarrow \text{Generate}(O_k, \mathcal{C}, t)$
  else if $\text{Label}(O_k) = T_2$ **then**
    $E \leftarrow \text{ExtractTimeEvents}(O_k, \mathcal{C}, t)$
    $\mathcal{T} \leftarrow \text{Sort}\left(\text{Remove}_{\text{Sim}(e_i,e_j)<\theta}(E)\right)$
  else if $\text{Label}(O_k) = T_3$ **then**
    $\text{H} \leftarrow \text{ExtractTriples}(O_k, \mathcal{C}, t)$
    $\mathcal{T} \leftarrow \text{Merge}(\text{H}, \tau)$
  end if
  $\mathcal{D} \leftarrow \mathcal{D} \cup \mathcal{T}$
end for

return $\mathcal{D}$

deduplication and consolidation function, and Sort refers to the temporal ordering operation. Finally, the overall result is sorted to generate the timeline $\mathcal{T}$.

Table Writing For the outline node O_k labeled as table type (T_3), it is necessary to perform triple extraction $H = \{H_1, H_2, \cdots, H_n\}$ (where $H_i = (e_i, r_i, a_i)$, with e_i representing an entity, r_i a relationship, and a_i an attribute) based on the candidate reference set to accurately capture core semantic elements such as entities, relationships, and attributes contained within. On this basis, redundancy elimination is conducted on the triple data from different document paragraphs, where τ serves as a critical threshold for similarity between triples to quantify the judgment criteria. Finally, the processed structured information is converted into academic-standard Markdown table data $\mathcal{T}$ through the $\text{Merge}(\cdot)$ function, achieving systematic aggregation and intuitive presentation of scattered knowledge.

$$\mathcal{T} = \text{Merge}(H, \tau) \tag{8}$$

Experiment

Datasets and Baselines

This study employs two datasets: FreshWiki (Shao et al. 2024), a curated set of 100 high-quality structured entries, and WikiGenBen (Zhang et al. 2024), a collection of 309 emerging events used to assess complex and dynamic text generation. To comprehensively evaluate the performance of WikiMAG in generating Wikipedia articles, we select four representative baseline models for comparison: AgentWrite (Bai et al. 2024), STORM (Shao et al. 2024), CO-STORM (Jiang et al. 2024), and the model from WikiGenBen (Zhang et al. 2024). Under the WikiGenBen framework, we compare its two distinct methods, PP and RPP, which address the challenges through different technical approaches.

Evaluation Metrics

In the construction of the evaluation system, the evaluation indicators for STORM are optimized: two core indicators, the heading soft recall and the heading entity recall(Shao et al. 2024), are introduced for outline quality evaluation; the full-text content quality evaluation adopts three n-gram measurement standards, Rouge-L, Rouge-1(Chin-Yew 2004), and METEOR(Banerjee and Lavie 2005), and the content quality assessment is carried out with a 5-point scoring system across four dimensions of "interest level", "coherence and organization", "relevance and focus", and "coverage", with automated scoring implemented by an LLM driven by Prometheus(Kim et al. 2023); the verifiability evaluation designs three indicators of citation recall rate, citation precision, and citation rate(Liu, Zhang, and Liang 2023), and the evaluation is carried out relying on an NLI model TRUE(Honovich et al. 2022).

Implementation Details

This study selects the open-source Qwen3-32B model(Yang et al. 2025) as the core generation model and adopts the vllm inference framework to improve inference efficiency. The experimental environment is built on an NVIDIA A800 80 GB GPU cluster. In the information retrieval phase, the Google Serper search tool API is called to provide rich and high-quality data support for model input. Regarding the configuration of model parameters: the temperature coefficient is set to 0.6, and the top-p parameter is set to 0.95; for the Reflective Inspector, the maximum number of iterations is limited to 3 to balance retrieval efficiency and information completeness. To accurately obtain the model performance under the basic configuration, the thinking mode mechanism is not enabled in this experiment, and the resulting benchmark data can provide a reliable reference for subsequent research.

Results and Analysis

Main Results

Experimental results show that the WikiMAG method based on the Qwen3-32B model demonstrates significant advantages in multiple dimensions, the results are shown in Table 2. In terms of outline generation quality, WikiMAG exhibits

Dataset	Methods	Outline			Writing						Verifiability		
		SRec	ERec	MET	R-1	R-L	Int.	Org.	Rel.	Cov.	Rec.	Prec.	Rte.
FreshWiki	AgentWrite	55.26	19.38	24.64	22.16	17.56	3.37	3.96	3.10	3.14	—	—	—
	PP	72.23	33.84	13.64	22.05	18.10	3.82	2.86	2.73	3.9	33.60	35.32	34.63
	RPP	75.67	33.17	21.83	21.32	17.52	3.33	3.19	3.27	3.05	40.35	42.97	41.28
	STORM	79.53	36.39	24.43	23.41	19.35	4.26	4.18	3.66	3.94	50.92	50.60	49.27
	CO-STORM	80.74	35.26	23.12	24.96	19.37	3.92	3.61	3.65	3.97	51.13	50.87	50.34
	Ours	**87.64**	**44.15**	**29.27**	**29.95**	**22.03**	**4.40**	**4.40**	**3.98**	**4.06**	**68.79**	**70.27**	**61.24**
WikiGenBen	AgentWrite	57.46	8.46	23.53	20.04	18.09	3.65	3.34	3.69	3.47	—	—	—
	PP	74.23	32.79	13.70	20.72	18.83	2.83	3.40	2.87	2.61	36.15	38.51	36.34
	RPP	76.86	31.48	22.95	19.98	21.66	3.57	3.72	3.98	3.14	36.01	36.40	34.57
	STORM	74.13	35.95	24.40	22.90	21.19	4.02	3.97	4.03	3.86	21.64	21.94	26.73
	CO-STORM	74.39	36.13	25.62	22.33	20.14	3.97	4.28	3.71	3.90	40.93	43.86	34.74
	Ours	**83.43**	**40.27**	**29.76**	**29.57**	**24.27**	**4.06**	**4.31**	**4.10**	**3.92**	**65.57**	**67.62**	**59.13**

Table 2: Comparison results of the Qwen3-32B model under different methods. SRec refers to Heading Soft Recall, ERec to Heading Entity Recall, MET to the METEOR metric, R-1 to the ROUGE-1 metric, R-L to the ROUGE-L metric, Int. to Interest Level, Org. to Coherence/Organization, Rel. to Relevance/Focus, and Cov. to Coverage. Rec., Prec., and Rte. denote Citation Recall, Precision, and Rate respectively. LLM scores (ranging 0-5, italicized) contrast with other metrics (scaled 0-100). Best results appear in bold, with second-best underlined.

remarkable superiority on the FreshWiki dataset. Its heading soft recall (SRec) reaches 87.64, which is 6.9 percentage points higher than the second-best method CO-STORM (80.74); the heading entity recall (ERec) is 44.15, far exceeding STORM's 36.39. This advantage is equally evident on the WikiGenBen dataset, where SRec and ERec lead other methods by 6.57 and 4.14 percentage points respectively.

Further analysis of writing quality reveals that WikiMAG performs particularly prominently. The METEOR on the FreshWiki and WikiGenBen datasets reach 29.27 and 29.76 respectively, which are more than 4 points higher than the second-best method. In terms of ROUGE metrics, its R-1 and R-L are significantly ahead, especially on the FreshWiki dataset, where the R-1 is nearly 5 points higher than that of CO-STORM.

The large language model evaluation dimension shows that WikiMAG also maintains a leading position in indicators such as interest level (Int.) and organization (Org.). It is worth noting that on the FreshWiki dataset, WikiMAG achieves the highest scores in both Org. and Int. , and also outperforms other methods in the relevance(Rel.) and coverage (Cov.) indicators.

It is particularly noteworthy that WikiMAG also performs excellently in terms of content verifiability. By virtue of dynamically retrieving and filtering reference, its citation recall rate (Rec.) on the FreshWiki dataset reaches 68.79, which is 17.66 percentage points higher than that of CO-STORM ; the citation precision (Prec.) of 70.27 is also outstanding. This advantage continues on the WikiGenBen dataset, where its citation recall rate is nearly 30 percentage points higher than the second-best method PP.

The above results fully verify the effectiveness of WikiMAG's design idea that combines dynamic retrieval mechanism with three content type processing methods, and the text quality has been improved in the task of Wikipedia content generation.

Ablation Study

We conducted an ablation study based on the FreshWiki dataset, aiming to quantitatively evaluate the independent contributions and synergistic effects of the Progressive Planner, Reflective Inspector, and Versatile Writer. Specifically, we performed three sets of controlled experiments by disabling each of the aforementioned components individually, and the results are summarized in Table 3.

As indicated by the experimental data, the Progressive Planner has a significant impact on the quality of the articles. This suggests that it plays a fundamental role in constructing content outlines and planning core information; the absence of this component directly leads to loose content structure and reduced thematic focus. The Reflective Inspector is crucial for ensuring the logical coherence of the text and is also indispensable in optimizing the accuracy of information filtering, as it can effectively filter redundant information to enhance the relevance of the content. The core value of the Versatile Writer lies in standardizing content expression (such as the presentation of narrative, timeline, and table) and strengthening the effectiveness of citations, thus providing important support for the credibility of the content.Overall, these results fully demonstrate that all three components are essential parts of the framework. They form a complete workflow through collaborative interaction, collectively ensuring the generation of high-quality content.

Reference Retrieval and Citation Efficiency

Although WikiGenBen employs a retrieval mechanism, this mechanism is static. Therefore, this experiment only compares STORM, CO-STORM, and WikiMAG. The results in Table 4 show that, compared with the other two methods, WikiMAG retrieves a slightly smaller number of references. This is not a limitation of its retrieval capability,

Methods	Writing							Verifiability		
	MET	R-1	R-L	Int.	Org.	Rel.	Cov.	Recall	Prec.	Rate
w/o Progressive Planner	26.41	24.32	15.47	3.58	3.47	3.31	3.11	58.62	59.34	57.15
w/o Reflective Inspector	28.33	25.86	18.12	4.34	4.25	3.64	3.25	68.14	68.24	69.35
w/o Versatile Writer	28.34	28.85	20.63	3.96	4.36	3.52	3.08	63.73	66.16	58.48
Ours	**29.27**	**29.95**	**22.03**	**4.40**	**4.40**	**3.98**	**4.06**	**68.79**	**70.27**	**61.24**

Table 3: The results of the ablation study on our framework under the FreshWiki dataset.

	Rtd.	Ctd.	C/R.
FreshWiki			
STORM	41.01	24.81	60%
CO-STORM	**42.18**	26.73	63%
Ours	38.17	**27.33**	**72%**
WikiGenBen			
STORM	40.95	22.33	55%
CO-STORM	**42.39**	24.25	57%
Ours	36.82	**24.71**	**67%**

Table 4: Comparison of the performance of STORM, CO-STORM, and WikiMAG on related citation metrics. Covering the Dimensions of References Retrieved(Rtd.), Cited References(Ctd.), and Cited/Retrieved Reference Ratio(C/R.)

but rather the result of its active filtering of references irrelevant to the topic and outline—whereas other methods do not perform such filtering and often retain a large amount of irrelevant content. Even so, WikiMAG still significantly outperforms the baseline methods in both the actual number of cited references after retrieval and the citation-retrieval ratio (the ratio of citations to retrievals). The core reason lies in the fact that WikiMAG aims for precise screening, focusing on acquiring high-quality references that are highly consistent with the thematic logic, thereby improving the usability of references, rather than blindly pursuing the quantity of retrievals. This result fully demonstrates WikiMAG's ability to finely control the quality of references in highly structured text generation.

Distribution and Citation of Content Types

As shown in the experimental results in Figure 3, there are significant differences between the FreshWiki and WikiGenBen datasets in terms of content type labeling and citation behavior, with timeline and table information exhibiting stronger citation dependence. Although narrative occupies an absolute advantage in quantity, the citation proportion of timeline and table content shows a significant upward trend, and this feature is particularly prominent in the FreshWiki dataset. Compared with narrative, timeline with clear structural characteristics and refined table content are more inclined to cite external sources. This citation preference may stem from higher requirements for information accuracy and authority. It is worth noting that the high citation tendency

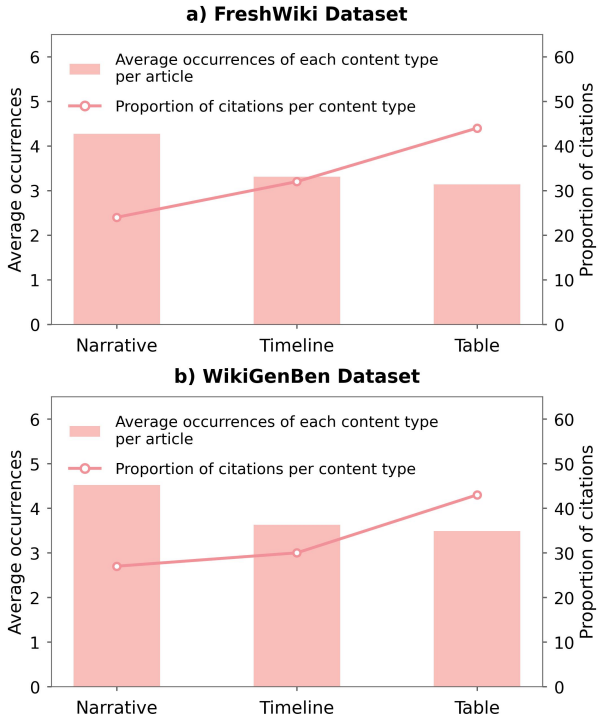

Figure 3: Average occurrences of different content types per article and proportion of citation counts contained in different content types per article.

of structured information essentially reflects the special emphasis on verifiability for such content, which also implies that it may play a more critical quality control role in the construction of knowledge systems.

Conclusion

This paper proposes WikiMAG, a novel multi-agent framework for generating structured Wikipedia-like articles. The framework integrates three core components to filter low-quality references and process narrative, timeline, and table content types. Its collaborative mechanism ensures module coordination while adhering to Wikipedia's standards. Experiments show WikiMAG outperforms baselines in outline quality, text generation, and citation accuracy, demonstrating the value of multi-agent and multi-content integration for structured article generation.

Acknowledgments

This work was supported by the National Natural Science Foundation of China (No. 62476283).

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*.

Bai, Y.; Zhang, J.; Lv, X.; Zheng, L.; Zhu, S.; Hou, L.; Dong, Y.; Tang, J.; and Li, J. 2024. Longwriter: Unleashing 10,000+ word generation from long context llms. *arXiv preprint arXiv:2408.07055*.

Banerjee, S.; and Lavie, A. 2005. METEOR: An automatic metric for MT evaluation with improved correlation with human judgments. In *Proceedings of the acl workshop on intrinsic and extrinsic evaluation measures for machine translation and/or summarization*, 65–72.

Chan, C.-M.; Xu, C.; Yuan, R.; Luo, H.; Xue, W.; Guo, Y.; and Fu, J. 2024. Rq-rag: Learning to refine queries for retrieval augmented generation. *arXiv preprint arXiv:2404.00610*.

Chen, D.; Fisch, A.; Weston, J.; and Bordes, A. 2017. Reading wikipedia to answer open-domain questions. *arXiv preprint arXiv:1704.00051*.

Chen, J.; Xiao, S.; Zhang, P.; Luo, K.; Lian, D.; and Liu, Z. 2024a. Bge m3-embedding: Multilingual, multi-functionality, multi-granularity text embeddings through self-knowledge distillation. *arXiv preprint arXiv:2402.03216*.

Chen, Z.; Liu, K.; Wang, Q.; Liu, J.; Zhang, W.; Chen, K.; and Zhao, F. 2024b. Mindsearch: Mimicking human minds elicits deep ai searcher. *arXiv preprint arXiv:2407.20183*.

Cheng, X.; Luo, D.; Chen, X.; Liu, L.; Zhao, D.; and Yan, R. 2023. Lift yourself up: Retrieval-augmented text generation with self-memory. *Advances in Neural Information Processing Systems*, 36: 43780–43799.

Chin-Yew, L. 2004. Rouge: A package for automatic evaluation of summaries. In *Proceedings of the Workshop on Text Summarization Branches Out, 2004*.

Gu, H.; Li, D.; Dong, K.; Zhang, H.; Lv, H.; Wang, H.; Lian, D.; Liu, Y.; and Chen, E. 2025. Rapid: Efficient retrieval-augmented long text generation with writing planning and information discovery. *arXiv preprint arXiv:2503.00751*.

Guu, K.; Lee, K.; Tung, Z.; Pasupat, P.; and Chang, M. 2020. Retrieval augmented language model pre-training. In *International conference on machine learning*, 3929–3938. PMLR.

Honovich, O.; Aharoni, R.; Herzig, J.; Taitelbaum, H.; Kukliansy, D.; Cohen, V.; Scialom, T.; Szpektor, I.; Hassidim, A.; and Matias, Y. 2022. TRUE: Re-evaluating factual consistency evaluation. *arXiv preprint arXiv:2204.04991*.

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*.

Kim, S.; Shin, J.; Cho, Y.; Jang, J.; Longpre, S.; Lee, H.; Yun, S.; Shin, S.; Kim, S.; Thorne, J.; et al. 2023. Prometheus: Inducing fine-grained evaluation capability in language models. In *The Twelfth International Conference on Learning Representations*.

Lewis, M.; Liu, Y.; Goyal, N.; Ghazvininejad, M.; Mohamed, A.; Levy, O.; Stoyanov, V.; and Zettlemoyer, L. 2019. Bart: Denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension. *arXiv preprint arXiv:1910.13461*.

Liang, Y.; Wu, Y.; Zhuang, H.; Chen, L.; Shen, J.; Jia, Y.; Qin, Z.; Sanghai, S.; Wang, X.; Yang, C.; et al. 2024. Integrating Planning into Single-Turn Long-Form Text Generation. *arXiv preprint arXiv:2410.06203*.

Liu, N. F.; Zhang, T.; and Liang, P. 2023. Evaluating verifiability in generative search engines. *arXiv preprint arXiv:2304.09848*.

Liu, Y.; Su, Y.; Shareghi, E.; and Collier, N. 2023. Instruct-SCTG: Guiding Sequential Controlled Text Generation through Instructions. *arXiv preprint arXiv:2312.12299*.

Radford, A.; Narasimhan, K.; Salimans, T.; Sutskever, I.; et al. 2018. Improving language understanding by generative pre-training.

Roy, K. K.; Akash, P. S.; Chang, K. C.-C.; and Popa, L. 2024. ConTReGen: Context-driven Tree-structured Retrieval for Open-domain Long-form Text Generation. *arXiv preprint arXiv:2410.15511*.

Shao, Y.; Jiang, Y.; Kanell, T. A.; Xu, P.; Khattab, O.; and Lam, M. S. 2024. Assisting in writing wikipedia-like articles from scratch with large language models. *arXiv preprint arXiv:2402.14207*.

Slobodkin, A.; Hirsch, E.; Cattan, A.; Schuster, T.; and Dagan, I. 2024. Attribute first, then generate: Locally-attributable grounded text generation. *arXiv preprint arXiv:2403.17104*.

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*.

Toro, S.; Anagnostopoulos, A. V.; Bello, S. M.; Blumberg, K.; Cameron, R.; Carmody, L.; Diehl, A. D.; Dooley, D. M.; Duncan, W. D.; Fey, P.; et al. 2024. Dynamic retrieval augmented generation of ontologies using artificial intelligence (DRAGON-AI). *Journal of Biomedical Semantics*, 15(1): 19.

Touvron, H.; Lavril, T.; Izacard, G.; Martinet, X.; Lachaux, M.-A.; Lacroix, T.; Rozière, B.; Goyal, N.; Hambro, E.; Azhar, F.; et al. 2023. Llama: Open and efficient foundation language models. *arXiv preprint arXiv:2302.13971*.

Wan, K.; Mu, H.; Hao, R.; Luo, H.; Gu, T.; and Chen, X. 2025. A cognitive writing perspective for constrained long-form text generation. *arXiv preprint arXiv:2502.12568*.

Yang, A.; Li, A.; Yang, B.; Zhang, B.; Hui, B.; Zheng, B.; Yu, B.; Gao, C.; Huang, C.; Lv, C.; et al. 2025. Qwen3 technical report. *arXiv preprint arXiv:2505.09388*.

Zhang, J.; Yu, E. J.; Chen, Q.; Xiong, C.; Zhu, D.; Qian, H.; Song, M.; Xiong, W.; Li, X.; Liu, Q.; et al. 2024. WIKIGEN-BENCH: Exploring Full-length Wikipedia Generation under Real-World Scenario. *arXiv preprint arXiv:2402.18264*.