DeepWriter: A Multi-Agent Collaboration Framework for Information-rich Ultra-long Book Writing

Ming Wang[1,2], Minghao Hu[2*], Xiuli Kang[3], Li He[1], Yu Tian[1], Chunming Liu[3]
Han Shi[1], Zhunchen Luo[2], Wei Luo[2], Guotong Geng[2]

[1]North China University of Technology, Beijing, China
[2]Center of Information Research, AMS, Beijing, China
[3]Hebei University of Engineering, Hebei, China

Abstract

Long-form books are among the most information-rich and structurally complex forms of written content, often exceeding 100,000 words. While recent methods have enabled basic long-text generation, they remain limited in two key aspects: the inability to generate ultra-long content at book scale, and the lack of mechanisms for integrating rich factual information. To address these limitations, we propose **DeepWriter**, a multi-agent collaborative framework that follows a structured planning-then-generation paradigm. It first constructs a detailed book outline with narrative arcs and chapter semantics, then incrementally generates content conditioned on retrieved knowledge and contextual signals. DeepWriter supports controllable generation of full-length books exceeding 100,000 words, enriched with citations, trivia and images. To support evaluation beyond surface-level fluency, we introduce **DeepWriter-Bench**, a bilingual benchmark of 18 annotated books designed to assess book-scale coherence, richness, and factual grounding. Additionally, we propose **BookScore**, a unified 100-point metric for quantifying book maturity. Experimental results show that DeepWriter achieves a state-of-the-art BookScore of 80.92, consistently outperforming strong baselines.

Introduction

Recent advances in large language models (LLMs) including GPT series (OpenAI 2023), Gemini (Google 2023), and LLaMA (Touvron et al. 2023) have significantly improved automatic text generation (Kasneci et al. 2023; Bubeck et al. 2023; Chen et al. 2024; Kim et al. 2025a). Despite these strides, generating ultra-long, information-rich content at book scale remains a formidable challenge. Most existing approaches fail in two crucial aspects: maintaining structural coherence across extensive narratives and densely integrating factual knowledge into the generated text.

Although recent multi-agent writing frameworks attempt to improve long-form generation, they still suffer from coarse task decomposition and lack mechanisms for rich information integration—limiting the diversity and depth of generated content. LongAgent (Zhao et al. 2024) introduces

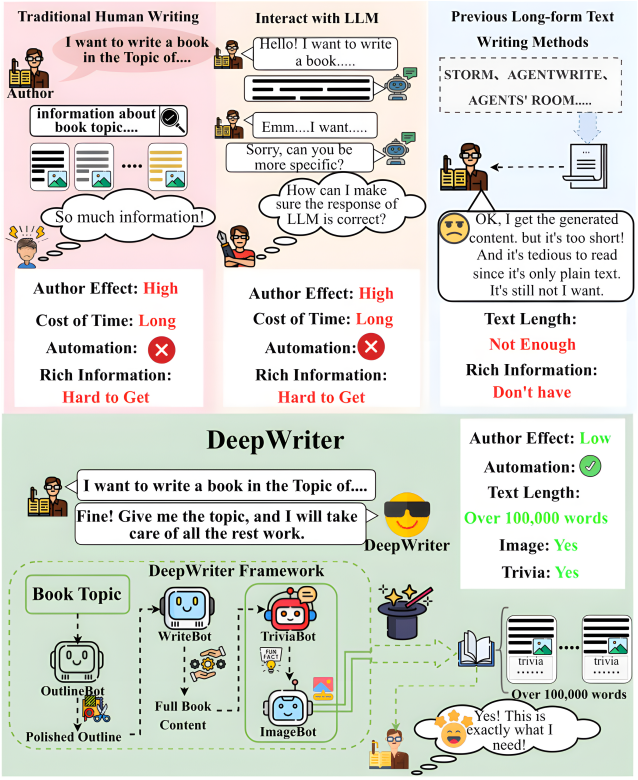

Figure 1: Limitations of prior long-form generation frameworks: human-in-the-loop methods lack scalability; prior long-form writing methods suffer from rigidity and repetition. **DeepWriter** employs specialized agent collaboration to enable scalable, coherent, and multimodal long-form generation.

role-based collaboration to maintain consistency, but is constrained to structured formats like news or academic articles, lacking flexibility for open-ended book writing. Chain of Agents (CoA) (Zhang et al. 2024) assigns subtasks to multiple agents to enhance text quality, but lacks dynamic inter-agent coordination, making it less effective for adaptive storytelling. StoryWriter (Xia et al. 2025) leverages memory for narrative coherence, yet still struggles with redundancy

*Corresponding authors, email: humh573@163.com

Method	Generation Max Length	Retrieval Granularity	Trivia	Image
AGENTS'R	2k words	None	No	No
STORM	4k words	Coarse	No	No
AgentWrite	20k words	None	No	No
SuperWriter	100k words	None	No	No
DeepWriter	**over 100k words**	**Fine**	**Yes**	**Yes**

Table 1: Overview of long-form text generation methods with respect to length scalability, retrieval strategy, and support for rich information. *AGENTS'R* refers to AGENTS' ROOM (Huot et al. 2024).

and fails to support multimodal enrichment.

In parallel, existing long-form generation benchmarks, such as LongBench (Bai et al. 2024a) and LongGen-Bench (Liu et al. 2024), focus largely on sentence-level fluency or span-level answering, without capturing essential qualities of book-scale generation: hierarchical structure, narrative coherence, factual alignment, and content richness. This lack of holistic evaluation hinders progress toward generating mature, publishable long-form texts.

As summarized in Table 1, existing long-form generation methods struggle with three fundamental limitations: the inability to scale generation to book-length content, insufficient factual grounding through retrieval, and a lack of support for injecting rich information such as trivia, citations, and images. These limitations prevent current models from achieving the structural and informational depth required for book-scale writing.

To address these challenges, we propose **DeepWriter**, a multi-agent collaborative framework that decomposes the book writing process into three coordinated stages: (1) structured outline planning, where simulated expert deliberation and retrieval-augmented critique produce a pedagogically sound and evidence-grounded outline; (2) literature-grounded drafting, where the framework retrieves multilingual scholarly sources and generates semantically aligned text with citation-level traceability; and (3) rich information insertion, where contextual trivia and semantically matched images are injected using dedicated agents to enhance reader engagement and factual richness. By aligning agents along these semantically coherent phases with interleaved refinement, DeepWriter supports controllable, modular generation of full-length books exceeding **100,000** words, while maintaining structural coherence and knowledge density throughout the manuscript.

In addition, we introduce **DeepWriter-Bench**, a bilingual benchmark designed for comprehensive evaluation of information-rich, book-scale generation. It comprises 18 annotated books and enables structured assessment through **BookScore**, a unified 100-point metric of book maturity.
Our main contributions are as follows:

- We present **DeepWriter**, a modular multi-agent collaboration framework for generating ultra-long, information-rich books. It decomposes the writing process into three coordinated stages, enables DeepWriter to generate books exceeding **100,000 words** with high structural coherence, factual grounding, and content density.
- We introduce **DeepWriter-Bench**, a bilingual benchmark of 18 manually annotated books across diverse topics, designed to comprehensively evaluate book-scale generation across coherence, factuality, and richness. We further propose **BookScore**, a unified 100-point metric that quantifies book maturity based on structural, lexical, and evidential dimensions.
- We conduct extensive experiments spanning multiple genres and languages. DeepWriter consistently outperforms strong long-form generation baselines across all major evaluation dimensions, demonstrating its effectiveness in controllable generation and rich-content integration.

Related Work

Long-Form Text Generation

Generating long-form content with large language models (LLMs) remains a core challenge due to the need for sustained discourse coherence, structured planning, and contextual fidelity. Prior work has attempted to improve structural control: Align-to-Structure (Kim et al. 2025b) enhances discourse coherence by aligning output with human-like outlines, but lacks fine-grained content planning. CogWriter (Wan et al. 2025) introduces a cognitive-inspired architecture with separate planning and writing agents, though its static decomposition limits flexibility in dynamic writing.

Recent work pushes the boundary of ultra-long generation. SuperWriter-Agent (Wu et al. 2025) proposes a three-stage reflection-driven pipeline (planning–writing–refining) and achieves high coherence, but does not incorporate retrieval or rich information like visuals or trivia. AGENTS' ROOM (Huot et al. 2024) organizes agents into expert roles to simulate collaborative discourse, improving fluency but lacking fine-grained control and scalability. LongWriter (Bai et al. 2024b) enables LLMs to generate over 10,000 words per document, yet struggles with cross-chapter consistency and long-range dependency.

Multi-Agent Frameworks

Multi-agent collaboration has emerged as a promising approach for enhancing reasoning, coordination, and creativity in complex tasks. Systems such as AutoGen (Wu et al. 2023) and HybriDialogue (Nakamura et al. 2022) demonstrate that coordinated agents can outperform single-agent models in planning and dialogue tasks. In the context of text generation, CO-STORM (Jiang et al. 2024) proposes a round-based conversational framework where multiple agents collaboratively compose articles, while Chain-of-Agents (CoA) (Zhang et al. 2024) decomposes writing into subtasks distributed across LLM agents to improve long-range coherence.

AGENTS' ROOM (Huot et al. 2024) further simulates structured expert discussions among role-based agents, but remains constrained in adapting to open-domain narrative needs. MAS-GPT (Ye et al. 2025) explores LLMs trained to

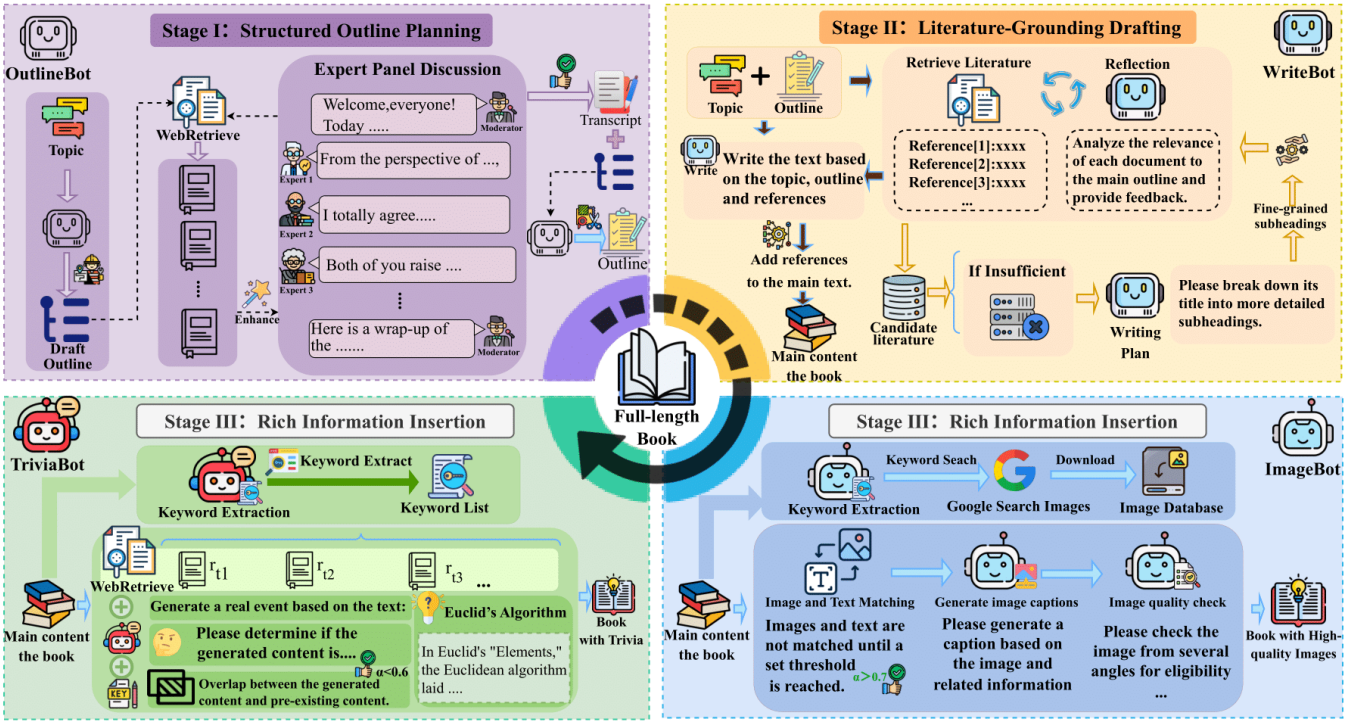

Figure 2: DeepWriter framework architecture. Each agent operates in a specific phase and interacts with others to share context, feedback, and retrieved knowledge.

construct and manage agent collectives, highlighting opportunities and coordination challenges in large-scale generation.

Methodology

DeepWriter employs a collaborative multi-agent paradigm inspired by cognitive and collective intelligence theories. Rather than isolating functionality into rigid agent-based sections, DeepWriter aligns agent contributions along a natural framework of long-form rich information book creation.

The entire framework is divided into three modular stages—structured outline planning, literature-grounded drafting, and rich-content enrichment—each governed by specialized agents. The output from one stage forms the structured input for the next, ensuring consistent semantic flow and modular extensibility. A high-level illustration of agent interactions and data flow is shown in Figure 2.

Stage I: Structured Outline Planning

DeepWriter's pipeline begins with comprehensive outline generation, orchestrated by *OutlineBot* within a multi-agent deliberation framework. This stage produces a hierarchical, pedagogically coherent skeleton that guides all downstream writing steps.

Initial Drafting. Given a research topic T and optional meta-constraints L (e.g. preferred writing style, target section depth), OutlineBot issues a parameterized prompt to the base LLM:

$$O_d = \texttt{LLM}\big(\texttt{Prompt}(T, L)\big) \tag{1}$$

The resulting draft outline O_d captures major thematic headings and subheadings, serving as the structural scaffold for refinement.

Expert-Agent Panel Simulation. To inject diverse perspectives and simulate epistemic disagreement, OutlineBot simulates a virtual panel comprising M domain experts $\{E_1, \ldots, E_M\}$ and a coordinating moderator H. Each expert E_i is dynamically instantiated based on the topic T and the current draft outline O_d, embodying a distinct disciplinary or functional lens (e.g. theory, application, evaluation). In asynchronous, free-form exchanges, agents may:

- Propose additions or reorganizations of sections based on their expertise.
- Question or rebut other agents' suggestions, facilitating adversarial refinement through multi-agent deliberation.
- Surface latent connections between topics (e.g. cross-sectional themes).

Figure 3 illustrates a sample of this dynamic. This simulated "academic workshop" yields a rich dialog transcript rather than a rigid turn-taking log.

Retrieval-Augmented Critique. During the debate, each agent performs retrieval-augmented grounding to maximize evidence utility for its structural proposals. At each turn, the moderator extracts a set of task-specific keywords:

$$K = \texttt{KeywordExtract}(\texttt{O}_\texttt{d}, \texttt{D}) \tag{2}$$

Here, D denotes the multi-agent dialog transcript at the current deliberation turn. Agents submit RAG queries with

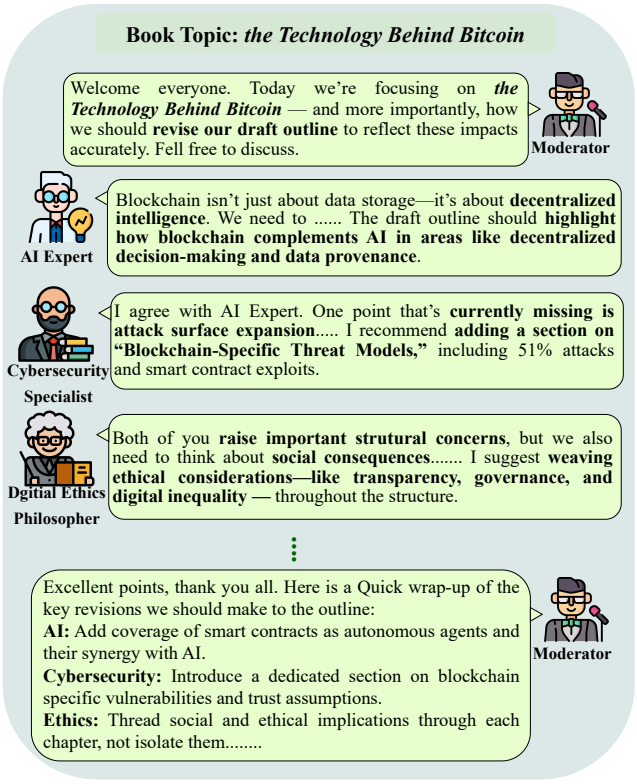

Figure 3: A sample expert panel discussion within OutlineBot, illustrating how domain-specialized agents collaboratively critique and refine an initial outline draft.

K, retrieve up-to-date evidence (papers, textbooks, benchmarks), and incorporate concise citations into their comments. This grounding step:

- Ensures factual support for structural suggestions.
- Highlights emerging trends or recent findings relevant to each heading.
- Mitigates hallucination by tying debates to real sources.

The resulting conversation is stored as a structured transcript:

$$\texttt{Transcript} = \{(H \to E_i, Q_i),\ (E_i \to H, A_i)\}_{i=1}^{M} \quad (3)$$

where Q_i is the query from moderator H, and A_i is the corresponding response from expert E_i. These interactions may include cross-agent replies and rebuttals, allowing multi-threaded discourse.

Synthesis and Hierarchical Refinement. After the panel concludes, OutlineBot aggregates expert feedback F and issues a consolidation prompt:

$$O_f = \texttt{LLM}\big(\texttt{RefinePrompt}(O_d, F)\big) \quad (4)$$

The refined outline O_f features:

- **Deeper granularity:** finer subheadings where debate exposed complexity.
- **Logical transitions:** narrative connectors that guide reader flow.
- **Pedagogical alignment:** ordering of sections optimized for learning or argumentation.

Compared to linear or single-pass methods, this deliberative approach yields an outline that is not only structurally sound but also dynamically informed by multiple expert perspectives and up-to-date evidence.

Outline Deployment and Agent Handoff. The final outline O_f is committed to DeepWriter's shared workspace. Compared to prior outline-based methods which rely on single-pass or rigid templates, our deliberative outline generation integrates multi-agent epistemic diversity, retrieval-augmented debate, and adaptive structural refinement—yielding significantly richer and pedagogically aligned outlines.

Stage II: Literature-Grounded Drafting

The second stage of **DeepWriter** focuses on synthesizing coherent and academically grounded content, guided by *WriteBot*. This module transforms the refined outline O_f into well-structured text by leveraging multi-source semantic retrieval, iterative quality screening, and citation-aligned generation.

Keyword Extraction and Cross-Lingual Web Retrieval. Given the global topic T and its corresponding structured outline O_f, WriteBot begins by extracting a set of representative keywords:

$$K = \texttt{KeywordExtract}(\texttt{O}_\texttt{f}, \texttt{T}) \quad (5)$$

These keywords guide a dual-channel retrieval process across both Chinese and English. To address cross-lingual retrieval asymmetries, we introduce a convex score fusion mechanism combining language-specific retrieval scores and global semantic similarity. Let S_1 and S_2 denote the relevance scores returned by Chinese and English retrieval engines, with $W_1 + W_2 = 1$ as their respective weights. The fused score is further re-ranked by semantic similarity $\texttt{Sim}$, computed via the BGE-m3 model (Chen et al. 2024), resulting in the initial candidate literature set R:

$$R = (W_1 \cdot S_1 + W_2 \cdot S_2) \cdot \texttt{Sim}(T) \quad (6)$$

Reflective Screening. To refine the quality of retrieved references beyond superficial lexical similarity, WriteBot performs reflective screening through an iterative mechanism. In each iteration t, WriteBot evaluates the relevance and contextual fit of the current reference set using a multi-dimensional scoring function that outputs a fit score F_t. The candidate set is then updated as follows:

$$R_t = \Phi(R_{t-1}, F_t) \quad (7)$$

If no satisfactory references are identified after $t_{\texttt{max}}$ iterations, WriteBot triggers an adaptive refinement process: it decomposes a coarse-grained outline heading o_i into finer sub-headings o_i', followed by re-extraction of a more specific keyword set K_i':

$$o_i' = \Psi(o_i), \quad K_i' = \texttt{KeywordExtract}(\texttt{o}_\texttt{i}', \texttt{T}) \quad (8)$$

This decomposition increases retrieval resolution and improves the chances of locating semantically aligned sources.

Grounded Content Generation. Once the refined literature set R_t is finalized, WriteBot proceeds to content generation. The generation process is guided by a multi-factor composition function that incorporates outline depth $D(o_i)$, logical context L, and generation history H:

$$C = \Gamma\left(R_t, D(o_i), L, H\right) \tag{9}$$

The resulting text is coherent, topically consistent, and adheres to academic writing conventions. Additionally, WriteBot automatically inserts domain-specific terminology and formats in-text citations according to standard scholarly style.

Citation Alignment. To ensure academic traceability, WriteBot conducts fine-grained citation alignment by computing semantic similarity between each generated content fragment $c_p \in C$ and the retrieved references $r_q \in R_t$. A mapping function Θ aligns each fragment to its most semantically similar reference, based on a BGE-based similarity threshold:

$$C' = \Theta\left(\left\{\texttt{Sim}_{\texttt{BGE}}(c_p, s_q) \;\middle|\; c_p \in C,\ r_q \in R_t\right\}\right) \tag{10}$$

A cosine similarity threshold of 0.75 is empirically set based on dev set performance. The final output C' consists of fully annotated content, where each citation is semantically aligned and empirically traceable. Compared to existing long-form generation approaches that rely on shallow retrieval or static references, our drafting process ensuring not only factuality and coverage, but also traceability and academic rigor in ultra-long text generation.

Stage III: Rich Information Insertion

In this stage, DeepWriter enriches the drafted text $C' = \{C'_1, \ldots, C'_n\}$ with contextual trivia and semantically aligned visuals, driven by *TriviaBot* and *ImageBot* respectively. Both agents operate on the shared draft, inserting supplemental content without disrupting the primary narrative flow.

Contextual Trivia Agent for Semantic Enrichment and Reader Engagement. For each revised paragraph $C'_i \in C' = \{C'_1, C'_2, \ldots, C'_n\}$, TriviaBot first identifies a set of concept keywords $K = \{k_1, k_2, \ldots, k_m\}$ based on both paragraph semantics and global topic structure T:

$$K = \texttt{KeywordExtract}(\texttt{C}'_\texttt{i}, \texttt{T}) \tag{11}$$

These keywords serve as input for a retrieval-augmented generation (RAG) pipeline to build a knowledge reference set $R = \{r_1, r_2, \ldots, r_m\}$, from which it constructs a candidate trivia pool via prompt-driven synthesis:

$$F_i = \texttt{TriviaGen}(\texttt{K}_\texttt{i}, \texttt{R}_\texttt{i}) \tag{12}$$

Each generated trivia $f_{ij} \in F_i$ is paired with a micro-title f^t_{ij}, enhancing reader skimmability and topical salience. Trivia candidates are selected based on informativeness and lexical orthogonality to the source content, ensuring novelty without redundancy. All candidates undergo GPT-4–based semantic validation and credibility filtering, and are deduplicated via cosine similarity:

$$\texttt{Sim}(F_i, R_i, C'_i) < \alpha \ (\alpha = 0.7) \tag{13}$$

Selected trivia is inserted immediately following its source paragraph, enhancing engagement and semantic depth.

Semantic-Guided Visual Insertion via Cross-Modal Retrieval. Concurrent to trivia injection, for each paragraph C'_i from C', ImageBot first extracts core keywords:

$$K = \texttt{KeywordExtract}(\texttt{C}'_\texttt{i}, \texttt{T}) \tag{14}$$

Using these keywords, ImageBot retrieves candidate images via web search API along with their source webpage URLs. For each URL, it crawls the page to obtain the full text and then identifies the segment D most semantically aligned with the image by comparing CLIP embeddings:

$$D = \texttt{FactExtract}(\texttt{URL}, \texttt{K}, \texttt{C}'_\texttt{i}, \texttt{T}) \tag{15}$$

Once the best-matching text segment D is found, ImageBot combines image and D to generate a descriptive caption. Each image and paragraph C'_i are then embedded via CLIP–ViT–H/14 (Radford et al. 2021), retaining only those where:

$$\cos\big(\texttt{CLIP}_{\texttt{img}}(I_j),\ \texttt{CLIP}_{\texttt{text}}(C'_i)\big) > \theta \tag{16}$$

The similarity threshold θ is empirically tuned on a held-out validation set to balance semantic relevance and visual diversity in image selection. Duplicates and top-level overload are avoided by limiting insertion to mid-level sections. Captions are center-aligned and scaled to 4 inches, formatted as "Figure x-y: [generated description]" for consistent academic presentation.

Both trivia and images are committed to the shared workspace in tandem, ensuring that DeepWriter's final manuscript is semantically rich, visually intuitive, and cohesively integrated.

DeepWriter-Bench

To thoroughly evaluate the performance of **DeepWriter** in multilingual long-form text generation, we construct **DeepWriter-Bench**—a meticulously curated benchmark of 18 full-length books, equally split between Chinese and English.

Language	Average Chapter Num	Average Words	Average Section Num
Chinese	7.32	107241	18.23
English	7.85	113254	17.41

Table 2: Basic Statistics of DeepWriter-Bench Dataset

High Fidelity Manual Annotation. Each book in this benchmark is sourced from professionally published high-resolution PDFs and annotated through meticulous manual effort. This involves page-by-page extraction and structural segmentation into hierarchical components:

- **Title1 (Book Title)**: The name of the book.
- **Title2 (Chapter Title)**: Major thematic divisions.
- **Title3 (Section Title)**: Subdivisions within each chapter.
- **Original Content**: The full narrative content per section.

This manual pipeline eliminates common artifacts in auto-curated datasets—such as broken boundaries and inconsistent formatting. DeepWriter-Bench serves as the foundation for all experiments in this paper, offering realistic, structurally sound ground truth for book-scale generation.

We further propose **BookScore**, a unified 100-point metric for evaluating overall book maturity. It comprises 20 fine-grained indicators across four modules—outline, main text, trivia, and images—each assessed under five weighted dimensions: structural integrity, textual coherence, citation verifiability, trivia informativeness, and image-text alignment. Each indicator contributes up to 5 points, summing to a composite score out of 100.

Experiment

Implementation Details

In the paper, we use Qwen3-32B (Yang et al. 2025) as the core driving model, and the model parameters for all generation tasks are uniformly configured with a temperature of 0.6 and top_p of 0.9. For ImageBot, Qwen2.5-VL-32B (Bai et al. 2025) is employed as the vision-language foundation model, which completes the processing by virtue of its powerful multimodal understanding and generation capabilities.

To further enhance the authenticity and timeliness of the content, the experiment obtains real-time web search results (top 10 for each query) via Google SERPER API. The retrieved passages are parsed and injected into the model prompts as context, enhancing factuality and timeliness.

Evaluation Metrics

Outline quality includes heading soft recall, heading entity recall, and LLM ratings on logical coherence, hierarchical structure , and relevance. Main text quality combines ROUGE-1, ROUGE-L (Lin 2004), entity recall (via FLAIR), and four LLM rating scores covering interest, organization, relevance, and coverage. Verifiability relies on NLI-based citation metrics (precision, recall, rate). Trivia is judged by factuality and interest. Image-text alignment uses BLIP-2 (Li et al. 2023) scores for semantic consistency, caption quality, and visual grounding.

LLM-based evaluations are conducted using model Prometheus (Kim et al. 2023), a 13B model trained to produce 0–5 scores across multiple quality aspects. Objective metrics are normalized before aggregation.

Baselines

To evaluate the contribution of each agent in *DeepWriter*, we design three hierarchical baselines by progressively integrating its core modules:

- **Direct_Gen**: A single-pass baseline that jointly generates outline and main text without retrieval or agent collaboration. The outline is the draft outline in stage I, and the text is produced using *WriteBot*'s prompt but lacks prior planning.
- **RAG_Gen**: Adds structured planning via *OutlineBot* and retrieval-augmented generation via *WriteBot*, covering DeepWriter's core writing pipeline without enrichment.

Baseline	Out.	Con.	Trivia	Image	B.S.
Direct_Gen	14.82	17.11	-	-	31.93
AgentWrite	14.91	17.95	-	-	32.86
STORM	16.57	27.88	-	-	44.45
Co-STORM	17.88	29.35	-	-	47.23
RAG_Gen	21.30	37.05	-	-	58.35
D.W._Trivia	21.30	37.05	9.19	-	67.54
DeepWriter	21.30	37.05	9.19	13.38	80.92

Table 3: Performance comparison of all baselines. Each row represents the average BookScore obtained for each book under a specific baseline. *Out.* refers to outline, *Con.* refers to content, *B.S.* refers to BookScore, *D.W._Trivia* refers to DeepWriter_Trivia baseline.

- **DeepWriter_Trivia**: Further incorporates *TriviaBot* to inject contextual trivia, isolating the enrichment stage's effect.

All models are evaluated on identical topics and content length. For fairness, retrieval excludes evaluation books. We also compare with recent systems—AgentWrite (Bai et al. 2024b), STORM (Shao et al. 2024), and CO-STORM (Jiang et al. 2024)—using *BookScore* as the primary metric.

Results and Analysis

Holistic Book-level Performance

Table 3 summarizes the average end-to-end BookScore (out of 100) across seven baselines. Each score aggregates performance across outline quality, chapter writing, trivia generation, and image alignment. *DeepWriter* achieves the highest score of 80.92, significantly outperforming earlier variants and prior methods like Co-STORM and AgentWrite.

Notably, the addition of trivia and image yields a substantial BookScore boost, underscoring the importance of rich content enrichment. This validates the effectiveness of rich information and contextual grounding in ultra-long book generation.

Outline Planning Performance

As shown in Table 5, *OutlineBot* outperforms all baselines across objective and subjective metrics. It achieves the highest heading soft recall (89.37%) and heading entity recall (62.12%), capturing both structure and content. LLM ratings further confirm its strengths in clarity (4.56), coherence (4.47), and relevance (4.69). In contrast, baselines like *Direct_Gen* and *AgentWrite* show weaker structural and topical quality, highlighting the value of agent-driven iterative planning.

Writing and Verifiability Performance

Table 4 shows that *WriteBot* surpasses all baselines. It achieves ROUGE-1 of 60.34, ROUGE-L of 45.13, and entity recall of 33.89. Subjective scores are also high in interest (4.53), organization (4.61), relevance (4.82), and coverage (4.46). Verifiability metrics lead across the board, with citation recall (67.35), precision (68.24), and rate (67.53). These

Methods	Writing							Verifiability		
	R-1	R-L	E-R	Int.	Org.	Rel.	Cov.	Rec.	Prec.	Rte.
AgentWrite	41.22	26.15	15.43	3.13	3.33	3.41	3.94	—	—	—
STORM	43.35	27.13	16.41	3.21	3.45	3.56	3.92	63.28	64.19	60.31
CO-STORM	45.32	28.64	16.63	3.74	3.95	3.41	3.77	65.73	68.83	64.35
Direct_Gen	41.30	26.64	15.32	3.16	3.33	3.14	3.32	—	—	—
WriteBot	**60.34**	**45.13**	**33.89**	**4.53**	**4.61**	**4.82**	**4.46**	**77.35**	**78.24**	**77.53**

Table 4: Evaluation results for writing quality and verifiability across different methods. **Writing** metrics include ROUGE-1 (R-1), ROUGE-L (R-L), Entity-Recall(E-R), Interest (Int.), Organization (Org.), Relevance (Rel.), and Coverage (Cov.). **Verifiability** metrics include Citation Recall (Rec.), Citation Precision (Prec.), and Citation Rate (Rte.). WriteBot achieves the best performance across all metrics.

Baseline	SoftR.	EntR.	LoC.	HiS.	Rele.
AgentWrite	73.13	32.15	3.15	3.42	3.07
STORM	79.22	40.18	3.57	3.41	3.62
Co-STORM	79.65	43.26	3.78	3.91	4.05
Direct_Gen	64.32	31.26	3.18	3.41	3.45
OutlineBot	**89.37**	**62.12**	**4.56**	**4.47**	**4.69**

Table 5: Evaluation for outline generation. Metrics include Heading Soft Recall (SoftR.), Heading Entity Recall (EntR.), Logical Coherence (LogC.), Hierarchical Structure (HieS.), and Rele.. OutlineBot significantly outperforms baselines in all dimensions.

Baseline	Factual Accuracy	Interest Level
Trivia_Direct	3.41	3.25
Trivia_RAG	4.03	3.91
TriviaBot	**4.63**	**4.56**

Table 6: Evaluation results for trivia generation. Metrics include Factual Accuracy and Interest Level. TriviaBot demonstrates both higher factual accuracy and greater user interest compared to baselines.

results validate the two-stage agents collaboration combining outline conditioning and retrieval-augmented generation.

Information Enrichment Performance

We further evaluate the effectiveness of *TriviaBot* and *ImageBot* in enhancing information richness and engagement. To isolate the effectiveness of *TrivialBot*, we designed two baselines: *Trivia_Direct* and *Trivia_RAG*. *Trivia_Direct* prompts the LLM to directly generate trivia without external retrieval or contextual grounding; while *Trivia_RAG* uses traditional retrieval-augmented generation (RAG) to introduce relevant knowledge before trivia generation. As shown in Table 6, *TriviaBot* outperforms retrieval-free and traditional RAG methods in both factual accuracy (4.63) and interest level (4.56). These results demonstrate the value of domain-aware, context-sensitive trivia generation over generic fact insertion.

In image generation, we compared *ImageBot* against two baselines: *Image_TextMatch*: uses the same 1000+ academic image repository as *ImageBot*. It extracts core paragraph keywords via LLMs (e.g., algorithm optimization) and filters images through exact matches with filenames/tags, using original image captions directly. *Image_FuzzyCrawl*: builds its repository via unstructured topic-based online crawling (no LLM keyword extraction). Post-construction, its matching (deduplication, insertion) and captioning follow *ImageBot*'s logic. Results (Table 7) show *ImageBot* outperforms both baselines across metrics: semantic consistency (4.32), caption quality (4.41), visual grounding (4.65). This validates the need for LLM-guided visual concept extraction and structured reasoning to align images with text.

Baseline	SemanC.	CapQ.	VisG.
Image_TextMatch	3.21	3.15	3.32
Image_FuzzyCrawl	3.45	3.21	3.42
ImageBot	**4.32**	**4.41**	**4.65**

Table 7: Evaluation results for image insertion quality. Metrics include Semantic Consistency (SemanC.), Caption Quality (CapQ.), and Visual Grounding (VisG.). ImageBot shows superior performance in aligning images with textual content and generating meaningful captions.

Conclusion

We introduce **DeepWriter**, a multi-agent collaboration framework for information-rich, ultra-long book generation. By integrating structured planning, retrieval-grounded drafting, and contextual enrichment, DeepWriter addresses key limitations of existing approaches. Experimental results on DeepWriter-Bench show that it consistently outperforms strong baselines across structure, coherence, factuality, and visual-textual synergy, offering a scalable and controllable solution for diverse narrative generation tasks.

Limitations. While DeepWriter performs robustly across genres, its dependence on prompt engineering and heuristic task decomposition may limit narrative diversity in highly creative domains. Future work will explore reinforcement learning for agent coordination, cross-section memory mechanisms, and story flow control to address these limitations.

Acknowledgments

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

References

Bai, S.; Chen, K.; Liu, X.; Wang, J.; Ge, W.; Song, S.; Dang, K.; Wang, P.; Wang, S.; Tang, J.; Zhong, H.; Zhu, Y.; Yang, M.; Li, Z.; Wan, J.; Wang, P.; Ding, W.; Fu, Z.; Xu, Y.; Ye, J.; Zhang, X.; Xie, T.; Cheng, Z.; Zhang, H.; Yang, Z.; Xu, H.; and Lin, J. 2025. Qwen2.5-VL Technical Report. arXiv:2502.13923.

Bai, Y.; Tu, S.; Zhang, J.; Peng, H.; Wang, X.; Lv, X.; Cao, S.; Xu, J.; Hou, L.; Dong, Y.; et al. 2024a. LongBench v2: Towards deeper understanding and reasoning on realistic long-context multitasks. *arXiv preprint arXiv:2412.15204*.

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

Bubeck, S.; Chandrasekaran, V.; Eldan, R.; Gehrke, J.; Horvitz, E.; Kamar, E.; Lee, P.; Lee, Y. T.; Li, Y.; Lundberg, S.; et al. 2023. Sparks of artificial general intelligence: Early experiments with gpt-4. *arXiv preprint arXiv:2303.12712*.

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

Google, G. T. 2023. Gemini: A Family of Highly Capable Multimodal Models. *arXiv preprint arXiv:2312.11805*.

Huot, F.; Amplayo, R. K.; Palomaki, J.; Jakobovits, A. S.; Clark, E.; and Lapata, M. 2024. Agents' Room: Narrative Generation through Multi-step Collaboration. *arXiv preprint arXiv:2410.02603*.

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*.

Kasneci, E.; Seßler, K.; Küchemann, S.; Bannert, M.; Dementieva, D.; Fischer, F.; Gasser, U.; Groh, G.; Günnemann, S.; Hüllermeier, E.; et al. 2023. ChatGPT for good? On opportunities and challenges of large language models for education. *Learning and individual differences*, 103: 102274.

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 *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, 1–15. Singapore: Association for Computational Linguistics.

Kim, Z. M.; Ramachandran, A.; Tavazoee, F.; Kim, J.-K.; Rokhlenko, O.; and Kang, D. 2025a. Align to Structure: Aligning Large Language Models with Structural Information. *arXiv preprint arXiv:2504.03622*.

Kim, Z. M.; Ramachandran, A.; Tavazoee, F.; Kim, J.-K.; Rokhlenko, O.; and Kang, D. 2025b. Align to Structure: Aligning Large Language Models with Structural Information. arXiv:2504.03622.

Li, J.; Li, D.; Savarese, S.; and Hoi, S. 2023. BLIP-2: Bootstrapping Language-Image Pre-training with Frozen Image Encoders and Large Language Models. arXiv:2301.12597.

Lin, C.-Y. 2004. ROUGE: A Package for Automatic Evaluation of Summaries. In *Text Summarization Branches Out*, 74–81. Association for Computational Linguistics.

Liu, X.; Dong, P.; Hu, X.; and Chu, X. 2024. Longgenbench: Long-context generation benchmark. *arXiv preprint arXiv:2410.04199*.

Nakamura, K.; Levy, S.; Tuan, Y.-L.; Chen, W.; and Wang, W. Y. 2022. HybriDialogue: An Information-Seeking Dialogue Dataset Grounded on Tabular and Textual Data. In Muresan, S.; Nakov, P.; and Villavicencio, A., eds., *Findings of the Association for Computational Linguistics: ACL 2022*, 481–492. Dublin, Ireland: Association for Computational Linguistics.

OpenAI. 2023. GPT-4 Technical Report. *arXiv preprint arXiv:2303.08774*.

Radford, A.; Kim, J. W.; Hallacy, C.; Ramesh, A.; Goh, G.; Agarwal, S.; Sastry, G.; Askell, A.; Mishkin, P.; Clark, J.; et al. 2021. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, 8748–8763. PmLR.

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*.

Touvron, H.; Lavril, T.; Izacard, G.; Martinet, X.; Lachaux, M.-A.; Lacroix, T.; Rozière, B.; Goyal, N.; Hambro, E.; Azhar, F.; Rodriguez, A.; Joulin, A.; Grave, E.; and Lample, G. 2023. LLaMA: Open and Efficient Foundation Language Models. 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:2502.12568.

Wu, Q.; Bansal, G.; Zhang, J.; Wu, Y.; Li, B.; Zhu, E.; Jiang, L.; Zhang, X.; Zhang, S.; Liu, J.; Awadallah, A. H.; White, R. W.; Burger, D.; and Wang, C. 2023. AutoGen: Enabling Next-Gen LLM Applications via Multi-Agent Conversation. arXiv:2308.08155.

Wu, Y.; Bai, Y.; Hu, Z.; Li, J.; and Lee, R. K.-W. 2025. SuperWriter: Reflection-Driven Long-Form Generation with Large Language Models. arXiv:2506.04180.

Xia, H.; Peng, H.; Qi, Y.; Wang, X.; Xu, B.; Hou, L.; and Li, J. 2025. StoryWriter: A Multi-Agent Framework for Long Story Generation. arXiv:2506.16445.

Yang, A.; Li, A.; Yang, B.; Zhang, B.; Hui, B.; Zheng, B.; Yu, B.; Gao, C.; Huang, C.; Lv, C.; Zheng, C.; Liu, D.; Zhou, F.; Huang, F.; Hu, F.; Ge, H.; Wei, H.; Lin, H.; Tang, J.; Yang, J.; Tu, J.; Zhang, J.; Yang, J.; Yang, J.; Zhou, J.; Zhou, J.; Lin, J.; Dang, K.; Bao, K.; Yang, K.; Yu, L.; Deng, L.; Li, M.; Xue, M.; Li, M.; Zhang, P.; Wang, P.; Zhu, Q.; Men, R.; Gao, R.; Liu, S.; Luo, S.; Li, T.; Tang, T.; Yin, W.; Ren, X.; Wang, X.; Zhang, X.; Ren, X.; Fan, Y.; Su, Y.; Zhang, Y.; Zhang, Y.; Wan, Y.; Liu, Y.; Wang, Z.; Cui, Z.; Zhang, Z.; Zhou, Z.; and Qiu, Z. 2025. Qwen3 Technical Report. arXiv:2505.09388.

Ye, R.; Tang, S.; Ge, R.; Du, Y.; Yin, Z.; Chen, S.; and Shao, J. 2025. MAS-GPT: Training LLMs to Build LLM-based Multi-Agent Systems. arXiv:2503.03686.

Zhang, Y.; Sun, R.; Chen, Y.; Pfister, T.; Zhang, R.; and Arik, S. 2024. Chain of agents: Large language models collaborating on long-context tasks. *Advances in Neural Information Processing Systems*, 37: 132208–132237.

Zhao, J.; Zu, C.; Xu, H.; Lu, Y.; He, W.; Ding, Y.; Gui, T.; Zhang, Q.; and Huang, X. 2024. Longagent: scaling language models to 128k context through multi-agent collaboration. *arXiv preprint arXiv:2402.11550*.