NEXUSSUM: Hierarchical LLM Agents for Long-Form Narrative Summarization

Hyuntak Kim* **Byung-Hak Kim*** †

CJ Corporation

Abstract

Summarizing long-form narratives—such as books, movies, and TV scripts—requires capturing intricate plotlines, character interactions, and thematic coherence, a task that remains challenging for existing LLMs. We introduce NEXUSSUM, a multi-agent LLM framework for narrative summarization that processes long-form text through a structured, sequential pipeline—without requiring fine-tuning. Our approach introduces two key innovations: **(1) Dialogue-to-Description Transformation**: A narrative-specific preprocessing method that standardizes character dialogue and descriptive text into a unified format, improving coherence. **(2) Hierarchical Multi-LLM Summarization**: A structured summarization pipeline that optimizes chunk processing and controls output length for accurate, high-quality summaries. Our method establishes a new state-of-the-art in narrative summarization, achieving up to **a 30.0% improvement in BERTScore (F1)** across books, movies, and TV scripts. These results demonstrate the effectiveness of multi-agent LLMs in handling long-form content, offering a scalable approach for structured summarization in diverse storytelling domains.

1 Introduction

Summarizing long-form narratives, such as books, movies, and TV scripts, remains an open challenge in NLP. Unlike news or document summarization, narratives require capturing intricate plotlines, evolving character relationships, and thematic coherence over tens of thousands of tokens (Zhao et al., 2022). The hybrid structure of narratives, which combines descriptive prose with multi-speaker dialogues, implicit inference, and dynamic topic shifts (see Figure 1), adds further complexity (Khalifa et al., 2021; Zou et al., 2021; Chen et al., 2022b; Saxena and Keller, 2024a), demanding an approach that preserves contextual integrity while condensing information effectively. Furthermore, the sheer length of narrative texts, typically ranging from 40K to 160K tokens (Kryscinski et al., 2022; Saxena and Keller, 2024b,a), poses significant challenges for standard summarization models.

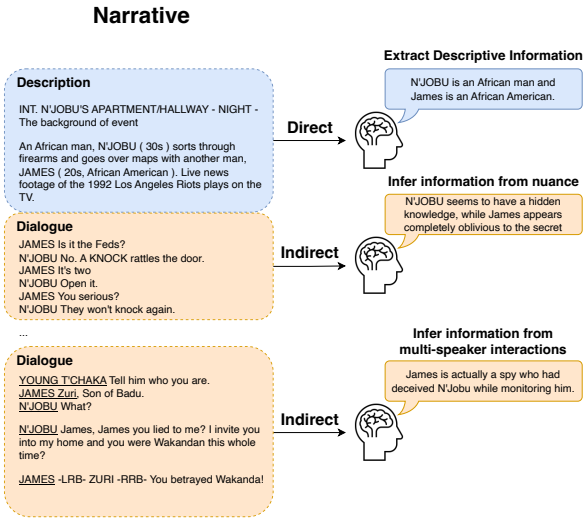

Figure 1: Illustration of narrative structure, showcasing the interplay between descriptive text and multi-speaker dialogues. NEXUSSUM enhances coherence by converting dialogues into structured prose for improved long-form summarization.

Despite advances in large language models (LLMs) for abstractive summarization (Pu et al., 2023), existing methods struggle with long-form narratives for three key reasons. First, context window limitations in LLMs (even with 200K-token capacities (Anthropic, 2024; OpenAI, 2023; Mistral AI, 2024)) lead to information loss when processing extended narratives (Liu et al., 2024). Second, extractive-to-abstractive pipelines (Ladhak et al., 2020; Liu et al., 2022; Saxena and Keller, 2024b) mitigate input constraints by selecting salient sections, but risk omitting critical details, disrupting

*Equal contribution.
†Corresponding author: bhak.kim@cj.net

Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), pages 10120–10157

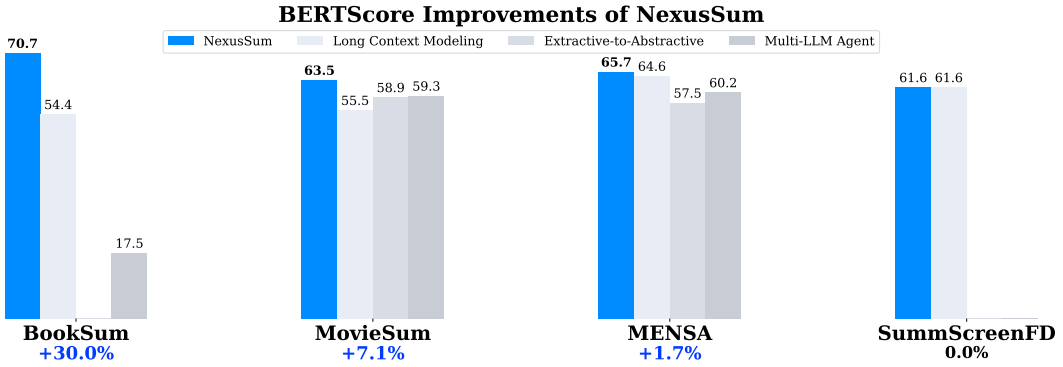

Figure 2: Performance comparison of NEXUSSUM with state-of-the-art summarization models using BERTScore (F1) across multiple benchmarks. NEXUSSUM achieves up to a 30.0% improvement, particularly excelling in BookSum, where hierarchical processing mitigates context truncation, demonstrating its advantage in long-form narrative summarization. Bold values indicate the new state-of-the-art score.

narrative coherence. Third, zero-shot LLM approaches perform poorly compared to fine-tuned models in narrative summarization (Saxena and Keller, 2024b; Saxena et al., 2025), indicating the need for task-specific adaptations beyond prompt engineering.

Recent Multi-LLM agent frameworks (Guo et al., 2024; Zhang et al., 2024; Chang et al., 2024) have introduced strategies to handle long-context documents through segmented inference and hierarchical processing. However, these studies focus primarily on generic document summarization and lack domain-specific optimizations for narrative discourse, character-driven coherence, and length-controlled output generation.

In this work, our aim is to address these challenges by investigating:

- **RQ1** How can a Multi-LLM agent system be designed to summarize long-form narratives while preserving narrative structure and coherence?
- **RQ2** What impact does dialogue-to-description transformation have on improving summarization consistency and readability?
- **RQ3** How does iterative compression affect summary length control and content retention?

To address these challenges, we introduce NEXUSSUM, a hierarchical multi-agent LLM framework for long-form narrative summarization. NEXUSSUM employs a three-stage sequential pipeline to progressively refine summaries without fine-tuning:

- **Dialogue-to-Description Transformation** Preprocessor agent converts character dialogues into structured narrative prose, reducing fragmentation and improving coherence.
- **Hierarchical Summarization** Narrative Summarizer agent generates an initial comprehensive summary, preserving key plot points and character interactions.
- **Iterative Compression** Compressor agent dynamically reduces summary length through controlled compression, ensuring key information retention while enforcing length constraints.

By segmenting long inputs into manageable chunks and applying hierarchical processing across multiple LLM agents, NEXUSSUM ensures high-fidelity summarization with scalable length control. We evaluate NEXUSSUM on four long-form narrative benchmarks: BookSum (Kryscinski et al., 2022), MovieSum (Saxena and Keller, 2024a), MENSA (Saxena and Keller, 2024b), and SummScreenFD (Chen et al., 2022a). As shown in Figure 2, NEXUSSUM outperforms existing methods, achieving up to a 30.0% improvement in BERTScore (F1) (Zhang et al., 2020) over previous state-of-the-art models. Our work makes the following contributions:

- **Dialogue-to-Description Transformation** We introduce a novel LLM-based preprocessing step that improves narrative coherence by converting dialogue into structured prose, reducing ambiguity in multi-speaker interactions.

- **Hierarchical Multi-Agent Summarization** We design a structured LLM agent pipeline that refines summaries iteratively, mitigating information loss while preserving contextual dependencies.

- **Optimized Length Control and Chunk Processing** Our framework employs iterative compression and dynamically adjusts chunk sizes, ensuring factual consistency while improving summary conciseness.

- **State-of-the-Art Results** Our approach establishes new benchmarks for long-form narrative summarization, achieving higher accuracy, coherence, and length control than existing LLM-based summarization methods.

By advancing multi-LLM agent frameworks for domain-specific narrative summarization, NEXUSSUM provides a scalable, fine-tuning-free solution that enhances long-context understanding across diverse storytelling mediums.

2 Related Work

Narrative summarization differs from traditional document summarization, requiring specialized techniques to handle complex plots, evolving characters, and mixed prose-dialogue structures. This section reviews related work on narrative summarization, long-context summarization, and multi-agent LLMs, positioning NEXUSSUM within this research landscape.

2.1 Narrative Summarization

Benchmark datasets like BookSum, MENSA, MovieSum and SummScreenFD have advanced long-form narrative summarization research. Traditional extractive-to-abstractive pipelines (Ladhak et al., 2020; Liu et al., 2022) risk losing coherence by omitting character arcs and event dependencies. To address this, scene-based and discourse-aware techniques leverage graph-based models (Gorinski and Lapata, 2015) and transformer-based saliency classifiers (Saxena and Keller, 2024b). However, these methods struggle with full text processing, often truncating key content. Our approach overcomes this gap by introducing the dialogue-to-description transformation, allowing for a holistic narrative processing while preserving coherence.

2.2 Long-Context Summarization

Long-context summarization techniques typically fall into two categories:

Architectural Optimization Transformer models struggle with scalability due to the quadratic cost of self-attention. Solutions include sparse attention, memory-efficient encoding, and long-context finetuning (Zaheer et al., 2020; Beltagy et al., 2020; Kitaev et al., 2020; Guo et al., 2022; Wang et al., 2020a). Expanded context windows (up to 200K tokens) (Chen et al., 2023; OpenAI, 2023; Mistral AI, 2024) help but still degrade in multi-turn dependencies, entity tracking, and coherence (Liu et al., 2024).

Chunking-Based Method Chunking-based approaches like SLED (Ivgi et al., 2023) and Unlimiformer (Bertsch et al., 2023) segment text for hierarchical summarization, while CachED (Saxena et al., 2025) improves efficiency via gradient caching but requires finetuning.

Unlike prior methods, NEXUSSUM offers a training-free alternative leveraging Multi-LLM agents, allowing full text summarization without truncation.

2.3 Multi-Agent LLMs for Summarization

Recent multi-agent LLM frameworks, such as Chain of Agents (CoA) (Zhang et al., 2024) and BooookScore (Chang et al., 2024), improve document summarization through hierarchical merging and sequential refinement (HM-SR) (Jeong et al., 2025). However, they lack adaptations for narrative coherence, character interactions, and event dependencies. Retrieval-augmented generation (Lewis et al., 2020) improves factuality but struggles with long-form storytelling, often missing thematic continuity (Geng et al., 2022; Uthus and Ni, 2023). NEXUSSUM addresses these gaps by integrating the dialogue-to-description transformation and systematic length control, ensuring coherent and contextually faithful summaries.

3 NEXUSSUM Framework

To address the challenges of long-form narrative summarization, we introduce NEXUSSUM, a hierarchical multi-agent LLM framework that processes narratives through a three-stage pipeline: **Preprocessing**, **Narrative Summarization**, and **Iterative Compression**. The system is designed to preserve narrative coherence, optimize summary

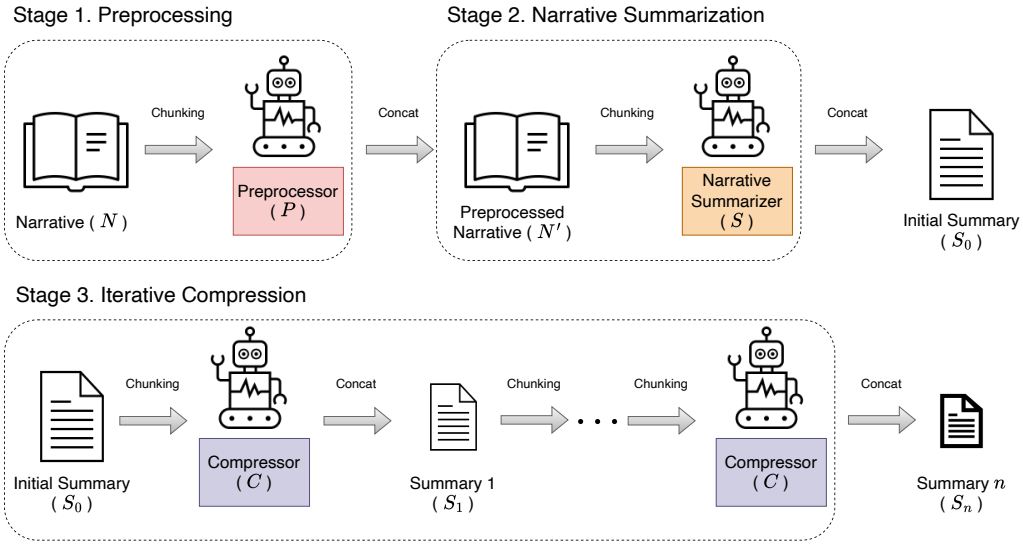

Figure 3: Overview of NEXUSSUM, a hierarchical multi-agent LLM framework for long-form narrative summarization. It follows a three-stage pipeline: (1) **Preprocessing** converts dialogues into descriptive prose, (2) **Narrative Summarization** generates an initial summary, and (3) **Iterative Compression** refines it for length control while preserving key details.

length, and ensure information retention without requiring fine-tuning. Figure 3 provides a schematic of the framework. Each stage of NEXUSSUM is optimized using a chunk-and-concat method, allowing for scalable summarization of narratives of arbitrary length while ensuring controlled compression. We detail the functionality of each stage below.[1]

3.1 Preprocessing Stage

Narrative texts combine dialogues and descriptions, often leading to fragmented summaries. The *Preprocessor* agent P enhances coherence by converting dialogues into structured third-person prose, simplifying input for summarization. Following (Xu et al., 2022), we prompt an LLM to reframe dialogues while preserving the intent of the speaker.

To manage long input lengths efficiently, P segments the input text into scene-based chunks, following recent studies (Saxena and Keller, 2024b; Jeong et al., 2025) that demonstrate the effectiveness scenes as semantic units for processing narratives:

$$N = n_1 \oplus n_2 \oplus \cdots \oplus n_k \quad (1)$$

where N represents the input narrative, segmented into k chunks. The number of chunks k is dynamically computed based on a fixed scene-based chunk size[2], and $\oplus$ denotes concatenation. Once processed, the output is a preformatted narrative text N', ready for summarization:

$$N' = P(n_1) \oplus P(n_2) \oplus \cdots \oplus P(n_k). \quad (2)$$

3.2 Narrative Summarization

The *Narrative Summarizer* agent S generates an initial abstract summary from the preprocessed text N'. To maintain coherence across long documents, N' is further chunked into scene-based units:

$$N' = n'_1 \oplus n'_2 \oplus \cdots \oplus n'_j \quad (3)$$

where j is the number of chunks[3]. The summarization process follows:

$$S_0 = S(n'_1) \oplus S(n'_2) \oplus \cdots \oplus S(n'_j) \quad (4)$$

where S_0 represents the initial summary. Unlike traditional single-pass models, NEXUSSUM applies hierarchical chunk processing, allowing long-range information retention.

3.3 Iterative Compression

While S_0 is an informative summary, it may exceed the desired length constraints. The *Compressor* agent C applies iterative compression to refine S_0 while preserving key narrative details. Our iterative compression method consists of two steps: sentence-based chunking followed by hierarchical compression.

[1]For a comprehensive breakdown of each stage, see Appendix A for the prompts used by each agent, and Appendix B for illustrative output samples.

[2]We set the chunk size to 8 scenes, balancing context retention and processing efficiency, resulting in $k = \text{total scenes}/8$.

[3]For simplicity, we set $j = k$, ensuring each chunk contains eight scenes.

Dataset	BookSum	MovieSum	MENSA	SummScreenFD
Domain	Novels	Movies	Movies	TV Shows
Eval Dataset Count	17	200	50	337
Avg. Input Length (Tokens)	158,645 (98.06%)	42,999 (24.08%)	39,808 (21.27%)	9,464 (38.91%)
Avg. Output Length (Tokens)	1,792 (46.43%)	902 (26.05%)	952 (17.02%)	151 (**76.16%**)

Table 1: Overview of four narrative summarization datasets, highlighting diverse text structures and summary styles. Input and output lengths are reported with Coefficient of Variation (CV), with SummScreenFD's high CV (76.16%) indicating significant variability, making it a challenging benchmark for consistency.

Sentence-Based Chunking Unlike the previous scene-based chunking, compression requires sentence-level granularity. We divide S_0 into smaller units:

$$S_0 = s_{0,1} \oplus s_{0,2} \oplus \cdots \oplus s_{0,l_0} \tag{5}$$

where l_0 is the number of chunks in the initial text, which is dynamically adjusted to maintain optimal compression ratios. Sentences are grouped into chunks up to a predetermined token size δ, allowing flexible compression rates. This δ plays a crucial role in controlling the compression ratio of our system's output, as the smaller size of input yields lower compression (see our empirical analysis in Appendix C).

Hierarchical Compression Following chunking, we apply hierarchical compression iteratively. In each iteration i, the Compressor agent C_i refines the previous compressed summary S_{i-1}, which is split into l_{i-1} chunks by Sentence-Based Chunking. The i-th Compressor agent C_i iteratively refines the summary:

$$S_i = C_i(s_{i-1,1}) \oplus C_i(s_{i-1,2}) \oplus \cdots \oplus C_i(s_{i-1,l_{i-1}}). \tag{6}$$

The process continues for n iterations, dynamically determined by a target word count θ[4]:

- If S_i exceeds θ, compression continues.
- If S_i falls below θ, the previous iteration's output is used.

To balance quality and computational efficiency, we limit compression to a maximum of 10 iterations.

4 Experimental Setup

This section describes four datasets of narrative summarization benchmarks, state-of-the-art baselines, implementation details, and evaluation metrics used in our study to evaluate NEXUSSUM.

4.1 Dataset

We conducted experiments on four diverse long-form narrative summarization datasets covering novels, movies, and TV scripts. Table 1 summarizes the key statistics of the dataset.

Dataset Descriptions

- BookSum: A novel-based summarization dataset with the longest input and output sequences requiring strong long-context comprehension.
- MovieSum: Contains summaries of 200 movies with moderate-length documents.
- MENSA: A script-based dataset combining ScriptBase (Gorinski and Lapata, 2015) and recent movie scripts, providing rich character interactions and scene-based storytelling.
- SummScreenFD: A dataset from TV shows with concise and highly variable summaries (CV[5] = 76.16%), testing the adaptability of NEXUSSUM to various writing styles.

4.2 Baselines

We compare NEXUSSUM with three main baseline categories, covering long context modeling, extractive-to-abstractive methods, and Multi-LLM agent frameworks.

Long Context Modeling Baselines These approaches modify model architectures to handle extended sequences:

- Zero-Shot through GPT-4o (OpenAI, 2023) and Mistral-Large (Mistral AI, 2024): Uses maximum context window expansion but struggles with truncation.
- SLED (Ivgi et al., 2023): Uses local attention with a sliding window mechanism.

[4]If the output is already below θ, the final summary may also be shorter than the target.

[5]CV measures dispersion calculated as the ratio of the standard deviation to the mean, expressed as a percentage.

- Unlimiformer (Bertsch et al., 2023): Extends transformers with unlimited retrieval-based attention.

- CachED (Saxena et al., 2025): A gradient caching approach for memory-efficient summarization.

Extractive-to-Abstractive Baselines These approaches extract salient segments before abstractive summarization:

- Description Only (Saxena and Keller, 2024a): Selects descriptive sections for summarization.

- Two-Stage Heuristics (Liu et al., 2022): Extracts character actions and key dialogues.

- Summ N (Zhang et al., 2022): Generates coarse summaries, then refines outputs iteratively.

- Select and Summ (Saxena and Keller, 2024b): Uses scene saliency classifiers to extract important moments.

Multi-LLM Agent Frameworks

- HM-SR (Jeong et al., 2025): Applies hierarchical chunk merging with refinement agents.

- CoA (Zhang et al., 2024): A multi-agent LLM pipeline, where each agent specializes in refining a specific summary aspect.

4.3 Implementation Details

We implement NEXUSSUM using Mistral-Large-Instruct-2407 (123B) (Mistral AI, 2024), with optimized inference via vLLM (Kwon et al., 2023) with temperature = 0.3, top-p = 1.0 and seed = 42. The model is run on four A100 GPUs. For Claude 3 Haiku (Anthropic, 2024), we set temperature = 0 to minimize randomness. For each benchmark, NEXUSSUM's configuration of δ and θ are detailed in the Appendices D and E.

To ensure that our LLM models were not exposed to evaluation datasets during training, we conducted an n-gram overlap analysis (see Appendix F). The results confirmed that the overlap remained below 2% on all benchmarks, indicating minimal data leakage and an unbiased evaluation.

4.4 Evaluation Metrics

We evaluate NEXUSSUM using a semantic similarity, length control metrics.

- **BERTScore (F1)** measures semantic similarity beyond n-gram overlap, aligning with human judgement. We use DeBERTa-XLarge-MNLI (He et al., 2021) as the base model, following established practices (Saxena et al., 2025; Saxena and Keller, 2024b). ROUGE (1/2/L) scores (Lin, 2004) are reported in Appendix G for comparability with prior work, though BERTScore better captures abstraction quality.

- **Length Adherence Rate (LAR)** measures the degree to which a summary matches the target word counts, defined as

$$\text{LAR} = 1 - |L_{\text{gen}} - L_{\text{target}}| \times L_{\text{target}}^{-1} \quad (7)$$

to quantify the effectiveness of iterative compression in controlling summary length.

5 Results and Analysis

We evaluate NEXUSSUM against state-of-the-art baselines on four narrative summarization benchmarks, assessing performance gains, ablation results, length control and agent adaptability. Additional analyses on factuality, document utilization and inference time complexity are provided in Appendices H, I and J.

5.1 Benchmark Performance

Table 2 summarizes the results, showing that NEXUSSUM outperforms all baselines across datasets, achieving state-of-the-art performance with substantial improvements over prior methods:

BookSum NEXUSSUM outperforms CachED by +30.0% BERTScore (F1), showcasing its effectiveness in processing extended narratives without context loss. Unlike CachED's static chunking approach, NEXUSSUM dynamically optimizes chunk sizes and applies iterative compression, preserving key information while enhancing coherence in long-form summarization. Additionally, NEXUSSUM surpasses CoA by +4.6% in ROUGE (geometric mean of 1/2/L), despite CoA leveraging Claude-3-Opus (Anthropic, 2024), a top-performing model for long-context summarization.

Method	BookSum	MovieSum	MENSA	SummScreenFD
Long Context Modeling				
Zero-Shot (Mistral Large, 123B)	46.42	55.50	54.80	57.23
Zero-Shot (GPT4o)	47.24	-	52.8	-
SLED (BART Large, 406M)	52.4	-	58.3	59.9
Unlimiformer (BART Base, 139M)	51.5	-	58.7	58.5
CachED (BART Large, 406M)	54.4	-	64.6	**61.59**
Extractive-to-Abstractive				
Description Only (LED-Large, 459M)	-	58.92	-	-
Two-Stage Heuristic (LED-Large, 459M)	-	58.54	56.34	-
Summ N (LED-Large, 459M)	-	-	40.87	-
Select and Summ (LED-Large, 459M)	-	-	57.46	-
Multi-LLM Agent				
HM-SR (GPT4o-mini)	-	59.32	60.22	-
CoA (Claude 3 Opus)	(17.47)	-	-	-
NEXUSSUM (Mistral Large, 123B)	**(18.27) / 70.70**	**63.53**	**65.73**	**61.59***

Table 2: Performance comparison of NEXUSSUM with state-of-the-art summarization models using BERTScore (F1) and ROUGE (geometric mean of 1/2/L) in parentheses. NEXUSSUM achieves its highest gains on BookSum (+30.0%) and MovieSum (+7.1%), outperforming Multi-LLM baselines like CoA and HM-SR. Baseline results are sourced from previous studies (Saxena and Keller, 2024a,b; Saxena et al., 2025; Jeong et al., 2025; Zhang et al., 2024). For open-source models, parameter sizes are shown in parentheses as (Model, Size).

Method	BERTScore (F1)	Improvement
Zero-Shot	54.81	-
P + Zero-Shot	57.26	+2.45
$P + S$	62.12	+4.86
$S + C$	63.90	+1.78
$P + S + C$ (NEXUSSUM)	**65.73**	+1.83

Table 3: Ablation analysis on the MENSA dataset, showing contributions of each LLM agent stage to a final BERTScore (F1) of 65.73.

Target Length	600	900	1200	1500
Length (Word)				
Zero-Shot	453	540	571	592
Ours	**670**	**891**	**1385**	**1621**
LAR				
Zero-Shot	0.245	0.400	0.524	0.605
Ours	**0.883**	**0.990**	**0.988**	**0.914**
BERTScore (F1)				
Zero-Shot	56.85	58.18	58.55	57.75
Ours	**63.59**	**65.73**	**65.21**	**62.86**

Table 4: Comparison of NEXUSSUM and Zero-Shot summarization on length control, measured by word count deviation, LAR and BERTScore (F1). NEXUSSUM achieves a higher BERTScore while maintaining an LAR close to 1.0, demonstrating precise adherence to target length constraints.

Method	BERTScore (F1)
NEXUSSUM $_{\text{base}}$	56.61
NEXUSSUM $_{\text{CoT}}$	58.61
NEXUSSUM $_{\text{CoT+FewShot}}$	**61.59**

Table 5: Effect of prompt engineering on NEXUSSUM performance in SummScreenFD. Incorporating CoT and Few-Shot learning results in a 5.0-point BERTScore improvement, highlights NEXUSSUM's adaptability to diverse summarization styles without parameter updates.

MovieSum NEXUSSUM outperforms HM-SR by +7.1%, leveraging structured length control to maintain consistency in multi-scene script summaries. While HM-SR effectively merges hierarchical summaries, its lack of precise length constraints results in variable-length outputs.

MENSA NEXUSSUM achieves a +1.7% gain over CachED, surpassing long-context models in screenplay summarization. It also outperforms Select and Summ by +14.4%, demonstrating superior abstraction for character-driven plots, where even extractive-to-abstractive methods struggle with maintaining narrative depth beyond scene selection.

SummScreenFD NEXUSSUM matches the performance of CachED while ensuring better length control through iterative compression, reducing output variability compared to zero-shot baselines.

5.2 Contribution of LLM Agents

Table 3 quantifies the contribution of each NEXUSSUM component through an ablation study. Zero-Shot summarization serves as the baseline, achieving BERTScore of 54.81. Introducing P improves coherence by converting dialogues into narrative text, raising BERTScore to 57.26 (+2.45). This confirms that P is insufficient alone for high-quality summarization. The addition of S further improves BERTScore to 62.12 (+4.86). Finally, C refines summary length while retaining critical details, producing the highest performance of 65.73. These results validate that each component of NEXUSSUM contributes to performance improvements, with the multi-agent LLM framework being essential for

long-form narrative coherence and retention.

5.3 Length Control with Quality Preservation

We evaluated NEXUSSUM's length control capabilities on the MENSA dataset. As a baseline, we use a Zero-Shot model with explicit length constraints applied via prompt instructions (*"Write in [target length]"*). As shown in Table 4, NEXUSSUM effectively balances semantic quality (BERTScore (F1)) and length adherence (LAR) in all target lengths. This shows that NEXUSSUM not only generates more semantically accurate summaries, but also enforces structured length control effectively than conventional prompting strategies.

5.4 Adaptive Performance through Prompt Engineering

As a Multi-LLM agent framework, NEXUSSUM leverages prompt engineering to adapt to diverse summarization tasks without requiring parameter updates. We evaluate this adaptability on SummScreenFD, a challenging dataset characterized by spoken dialogue format and high variable summary styles (CV= 76.16%, see Table 1). To enhance adaptation, we incorporate Chain of Thoughts (CoT) reasoning (Kojima et al., 2022) in P and Few-Shot learning (Wang et al., 2020b) in S and C to refine output style.

Table 5 demonstrates that CoT alone improves BERTScore (F1) from 56.61 to 58.61 (+2.0 points), while adding Few-Shot learning further boosts performance to 61.59 (+2.98 points). These results highlight NEXUSSUM's ability to adapt to diverse summarization scenarios using simple prompt customization, ensuring robust generalization across narrative structures without additional training or fine-tuning.

5.5 Human Preference Analysis

Setup To explore human preference for generated summaries across different narrative styles and genres, we create three different K-Drama summaries that vary in genres (Fantasy-Romance, Korean History and Modern Romantic-Comedy). Summaries were generated using three different methodologies (Zero-Shot, NEXUSSUM and NEXUSSUM$_R$) to enable a comparative preference analysis.

A total of three K-Drama experts participate in the evaluation, with at least two evaluators assessing each output for a given work. They score the summaries on a 5-point Likert scale (1 = Not at all, 5 = Very much so) across four criteria:

- **Key Events**: Are the key events included?
- **Flow**: Is the contextual information demonstrated specifically?
- **Factuality**: Does the summary have high factual accuracy?
- **Readability**: Does the summary have high readability?

In addition, all three evaluators provided qualitative comments to explain the reasons behind their scores (See Appendix K).

Results First, we compare the Zero-Shot method with NEXUSSUM. Each method aims to generate summaries with a target length of 600 words. Zero-Shot is prompted to generate summaries with the target length of 600 as specified instruction in the prompt. NEXUSSUM generates summaries by halting the iteration process at a lower bound $\theta = 600$. As shown in Table 6, NEXUSSUM demonstrates superior summary length control, achieving an average summary length of 609 words, compared to Zero-Shot, which produces summaries with an average length of 219 words. NEXUSSUM outperforms Zero-Shot in capturing key events (4.17), maintaining narrative flow (3.34), and ensuring factual accuracy (4). However, Zero-Shot demonstrates superior readability (4.17).

To further enhance readability, we introduce a third method, NEXUSSUM$_R$. This approach incorporates an additional LLM agent that rewrites the original NEXUSSUM summary to emulate the concise and fluent style characteristic of the Zero-Shot method. The refining agent smooths sentence transitions, adjusts verbosity, and enhances fluency while preserving key narrative details. In Table 6, NEXUSSUM$_R$ improves readability by +1.5 points compared to NEXUSSUM, bridging the gap between structured factual summarization and human-preferred fluency.

6 Conclusion

We introduce NEXUSSUM, a hierarchical multi-agent LLM framework that advances long-form narrative summarization by improving coherence (RQ2), long-context processing (RQ1), and length control (RQ3). Our results demonstrate that structured multi-agent collaboration enhances information retention while maintaining coherence, laying

	Zero-Shot	NEXUSSUM	NEXUSSUM$_R$
Key Events	3.5	**4.17**	**4.17**
Flow	2.83	3.34	3
Factuality	3.5	4	3.67
Readability	**4.17**	2.17	3.67
Avg. Output Len	219	609	234

Table 6: Expert evaluation of Zero-Shot, NEXUSSUM, and NEXUSSUM$_R$ on K-Drama summaries using a target length (θ) of 600 words. Scores reflect performance across four criteria. NEXUSSUM$_R$ introduces a reflection step to enhance readability while maintaining high content retention.

the groundwork for scalable, adaptive AI summarization systems.

Beyond state-of-the-art performance, NEXUSSUM has broader implications for AI-driven storytelling, personalized summarization, and conversational AI. Our findings on Chain-of-Thought-driven self-planning suggest a path toward autonomous, context-aware LLM agents capable of refinement without retraining. However, human evaluation highlights a readability gap compared to Zero-Shot baselines. Future work should explore a fluency-enhancing summarization framework while preserving factual consistency and optimizing multi-agent collaboration efficiency.

7 Limitations

While NEXUSSUM introduces significant advancements in long-form narrative summarization, certain limitations remain, particularly in evaluation paradigms, readability, and adaptability. This section outlines key challenges and directions for future improvements.

Limitation of Automated Metrics Automated evaluation metrics such as BERTScore and ROUGE provide useful approximations of summary quality but fail to capture readability, coherence, and user preference, which are critical for long-form narrative summarization. To address these gaps, we conducted an expert evaluation on three K-drama summaries from distinct genres (historical, fantasy, and slice-of-life) to assess readability, coherence, and factual accuracy (Section 5.5).

As shown in Table 6, NEXUSSUM produces summaries closer to the target length (609 words) than Zero-Shot (219 words). However, despite NEXUSSUM achieving higher BERTScore and ROUGE, experts rated Zero-Shot outputs as more readable (4.17 vs. 2.17). This discrepancy suggests Zero-Shot favors fluency and stylistic variation at the cost of factual accuracy, whereas NEXUSSUM focuses on key event retention, leading to denser summaries that may feel less natural to human readers. These findings highlight a crucial limitation of current summarization evaluation paradigms—higher automated scores do not necessarily align with human preference.

Future Directions Human feedback (Section 5.5, Appendix K) suggests that NEXUSSUM$_R$ reduces rigid phrasing and improves narrative flow, making summaries more natural while retaining essential content. This demonstrates that an additional reflection step can significantly enhance human preference alignment, opening the door to adaptive post-processing techniques for long-form summarization to offer customizable and more engaging user experiences.

Acknowledgments

We thank our colleagues at the AI R&D Division for their insightful discussions that helped shape the direction of this work. We also thank the team at CJ ENM for their support with human evaluation and for providing valuable feedback throughout the project.

References

Anthropic. 2024. The claude 3 model family: Opus, sonnet, haiku. https://www-cdn.anthropic.com/de8ba9b01c9ab7cbabf5c33b80b7bbc618857627/Model_Card_Claude_3.pdf. Online; accessed March 2024.

Iz Beltagy, Matthew E. Peters, and Arman Cohan. 2020. Longformer: The long-document transformer. *Preprint*, arXiv:2004.05150.

Amanda Bertsch, Uri Alon, Graham Neubig, and Matthew R. Gormley. 2023. Unlimiformer: Long-range transformers with unlimited length input. In *Thirty-seventh Conference on Neural Information Processing Systems: NeurIPS 2023*.

Yapei Chang, Kyle Lo, Tanya Goyal, and Mohit Iyyer. 2024. Booookscore: A systematic exploration of book-length summarization in the era of llms. In *The Twelfth International Conference on Learning Representations, ICLR 2024, Vienna, Austria, May 7-11, 2024*. OpenReview.net.

Mingda Chen, Zewei Chu, Sam Wiseman, and Kevin Gimpel. 2022a. SummScreen: A dataset for abstractive screenplay summarization. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), ACL 2022*, pages 8602–8615, Dublin, Ireland. Association for Computational Linguistics.

Shouyuan Chen, Sherman Wong, Liangjian Chen, and Yuandong Tian. 2023. Extending context window of large language models via positional interpolation. *Preprint*, arXiv:2306.15595.

Xinyun Chen, Renat Aksitov, Uri Alon, Jie Ren, Kefan Xiao, Pengcheng Yin, Sushant Prakash, Charles Sutton, Xuezhi Wang, and Denny Zhou. 2024. Universal self-consistency for large language models. In *ICML 2024 Workshop on In-Context Learning*.

Yulong Chen, Naihao Deng, Yang Liu, and Yue Zhang. 2022b. DialogSum challenge: Results of the dialogue summarization shared task. In *Proceedings of the 15th International Conference on Natural Language Generation: Generation Challenges*, pages 94–103, Waterville, Maine, USA and virtual meeting. Association for Computational Linguistics.

Zhichao Geng, Ming Zhong, Zhangyue Yin, Xipeng Qiu, and Xuanjing Huang. 2022. Improving abstractive dialogue summarization with speaker-aware supervised contrastive learning. In *Proceedings of the 29th International Conference on Computational Linguistics, COLING 2022*, pages 6540–6546, Gyeongju, Republic of Korea. International Committee on Computational Linguistics.

Philip John Gorinski and Mirella Lapata. 2015. Movie script summarization as graph-based scene extraction. In *Proceedings of the 2015 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies : NAACL HLT 2015*, pages 1066–1076, Denver, Colorado. Association for Computational Linguistics.

Mandy Guo, Joshua Ainslie, David Uthus, Santiago Ontanon, Jianmo Ni, Yun-Hsuan Sung, and Yinfei Yang. 2022. LongT5: Efficient text-to-text transformer for long sequences. In *Findings of the Association for Computational Linguistics, NAACL 2022*, pages 724–736, Seattle, United States. Association for Computational Linguistics.

Taicheng Guo, Xiuying Chen, Yaqi Wang, Ruidi Chang, Shichao Pei, Nitesh V. Chawla, Olaf Wiest, and Xiangliang Zhang. 2024. Large language model based multi-agents: A survey of progress and challenges. *Preprint*, arXiv:2402.01680.

Pengcheng He, Xiaodong Liu, Jianfeng Gao, and Weizhu Chen. 2021. Deberta: decoding-enhanced bert with disentangled attention. In *9th International Conference on Learning Representations, ICLR 2021, Virtual Event, Austria, May 3-7, 2021*. OpenReview.net.

Maor Ivgi, Uri Shaham, and Jonathan Berant. 2023. Efficient long-text understanding with short-text models. *Transactions of the Association for Computational Linguistics*, 11:284–299.

Yeonseok Jeong, Minsoo Kim, Seung won Hwang, and Byung-Hak Kim. 2025. Agent-as-judge for factual summarization of long narratives. *Preprint*, arXiv:2501.09993.

Muhammad Khalifa, Miguel Ballesteros, and Kathleen McKeown. 2021. A bag of tricks for dialogue summarization. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing, EMNLP 2021*, pages 8014–8022, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Nikita Kitaev, Łukasz Kaiser, and Anselm Levskaya. 2020. Reformer: The efficient transformer. *Preprint*, arXiv:2001.04451.

Takeshi Kojima, Shixiang Shane Gu, Machel Reid, Yutaka Matsuo, and Yusuke Iwasawa. 2022. Large language models are zero-shot reasoners. In *Advances in Neural Information Processing Systems 35: Annual Conference on Neural Information Processing Systems 2022, NeurIPS 2022*.

Wojciech Kryscinski, Nazneen Rajani, Divyansh Agarwal, Caiming Xiong, and Dragomir Radev. 2022. BOOKSUM: A collection of datasets for long-form narrative summarization. In *Findings of the Association for Computational Linguistics, EMNLP 2022*, pages 6536–6558, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Woosuk Kwon, Zhuohan Li, Siyuan Zhuang, Ying Sheng, Lianmin Zheng, Cody Hao Yu, Joseph Gonzalez, Hao Zhang, and Ion Stoica. 2023. Efficient memory management for large language model serving with pagedattention. In *Proceedings of the 29th Symposium on Operating Systems Principles*, SOSP '23, page 611–626, New York, NY, USA. Association for Computing Machinery.

Faisal Ladhak, Bryan Li, Yaser Al-Onaizan, and Kathleen McKeown. 2020. Exploring content selection in summarization of novel chapters. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, ACL 2020*, pages 5043–5054, Online. Association for Computational Linguistics.

Patrick Lewis, Ethan Perez, Aleksandra Piktus, Fabio Petroni, Vladimir Karpukhin, Naman Goyal, Heinrich Küttler, Mike Lewis, Wen-tau Yih, Tim Rocktäschel, Sebastian Riedel, and Douwe Kiela. 2020. Advances in neural information processing systems 33, neurips 2020. In *Advances in Neural Information Processing Systems*, volume 33, pages 9459–9474. Curran Associates, Inc.

Chin-Yew Lin. 2004. ROUGE: A package for automatic evaluation of summaries. In *Text Summarization Branches Out*, pages 74–81, Barcelona, Spain. Association for Computational Linguistics, ACL 2004.

Dongqi Liu, Xudong Hong, Pin-Jie Lin, Ernie Chang, and Vera Demberg. 2022. Two-stage movie script summarization: An efficient method for low-resource long document summarization. In *Proceedings of The Workshop on Automatic Summarization for Creative Writing*, pages 57–66, Gyeongju, Republic of Korea. Association for Computational Linguistics.