SeqUAL: Sequel story Unfolding and Augmentation with LLMs

Anonymous ACL submission

Abstract

Generating sequels for an original story is a challenging task that requires maintaining both consistency and diversity. Existing story generation studies are limited in that they typically measure either diversity or consistency, but not both, even though there exists a trade-off between them during the sequel generation. In this study, we propose the SeqUAL (Sequel Unfolding and Augmentation with LLM) framework, which aims to enhance the diversity of events and participating characters in sequels while preserving the original story's worldview; SeqUAL leverages a History Extraction module and a Sequel Character Augmentation module. We evaluated its performance in terms of diversity and consistency through comprehensive analyses. Experimental results demonstrate that SeqUAL achieves significant improvements in both diversity and consistency, impressively striking a balance between the two. Additionally, we identify points within a story where the SeqUAL framework can effectively enhance diversity. This work represents the first practical approach for generating more diverse and consistent sequels, contributing to the advancement of automated story generation.

1 Introduction

With the recent advancement of Large Language Models (LLMs), research on automatic story generation has been actively progressing (Venkatraman et al., 2025; Dzhubaeva et al., 2025; Xie et al., 2023; Wen et al., 2023). Mainly, studies on story generation have focused on generating consistent stories while overlooking the diversity of stories (Wang et al., 2025a; Mu and Li, 2024; Ma et al., 2023; Yi et al., 2025). So, recently, researchers have begun to generate diverse derivative works based on original stories (Xu et al., 2025; Uduehi et al., 2024). In a similar vein, we focus on a type of derivative work that has received little attention in prior research: sequels. Sequels have different properties compared to other derivative works; they should maintain a worldview *consistent* with the original story while producing distinct and *diverse* story plots. As these *diversity* and *consistency* have a trade-off relationship, sequels could be an ideal testbed for evaluating both.

Existing research, however, has not simultaneously considered both *diversity* and *consistency*. Therefore, it remains questionable whether those methods can be applied to sequels. Regarding diversity, the current story generation modules produce similar story plots because they are tightly-aligned with the events in the original work. Existing studies on derivative works have focused on re-generating stories by following similar patterns and structures of the original, paying little attention to story diversity (Han et al., 2025; Rashkin et al., 2020). In a sequel, however, its plot should be different from the original since it is a new story of other characters after the last event of the original. The introduction of new characters also brings changes in social relationships, societal structures, and events, making the plot different from the original. Specifying such changes is challenging since changes may harm the consistency.

Regarding consistency, the current story generation methods have less examined whether the generated stories are consistent with the original historical flow. New characters and events increasing diversity should be consistent with the original world. However, existing research has only examined whether original characters fit their background (Ran et al., 2025; Park et al., 2025), without ensuring that their historical or societal relationship is preserved. So, adding new characters into those generation systems may produce an inconsistent story. Moreover, when the methods are good at preserving the original worldview, high consistency with the original world may harm the diversity, as they consistently generate a similar plot.

We attempt to address the diversity-consistency

trade-off that occurs when generating sequels. We propose SeqUAL (Sequel Unfolding and Augmentation with LLM), which contains two key components. First, we design a Sequel Character Augmentation module that automatically generates new characters from personality traits, and thus diversifies the plot. Second, we design a History Extraction module that makes new characters and plots grounded in the worldview of the original story. Using these two simple approaches, our SeqUAL can successfully generate a wide range of diverse stories, while maintaining consistency with the original world.

Through experiments, we validate effectiveness of SeqUAL. We collect ten recent Sci-Fi novels and ask SeqUAL to generate 20 sequels for each novel. As a result, this paper has following contributions:

- We present SeqUAL, which is the first work handling the diversity-consistency trade-off.
- Regarding diversity, SeqUAL generated diverse content. Specifically, near a major setback, the diversity becomes high.
- Regarding consistency, SeqUAL increased the consistency by successfully recovering societal structures from the original story.

2 The SeqUAL Framework

As shown in Figure 1, SeqUAL takes an original story as an input and unfolds diverse sequels in an end-to-end manner. It consists of three LLM-based modules. First, History Extraction module extracts the past history of original characters from the original story. Second, Sequel Character Augmentation module creates personality candidates for new characters and augments the new characters by combining them with the extracted history. Third, Sequel Unfolding module then unfolds the sequel based on this constructed new character.

2.1 Three Elements of A Persona

Before presenting each module, we first define three key concepts that will be used throughout the section. These three concepts shape the character's past, present, and future within a narrative. After we consulted with professional actors and writers with over five years of experience, we decided on three key elements regarding the order of events, as follows. We refer to these three the *persona* of a character, which can contribute to making a distinct and consistent story.

History refers to the past actions performed by original characters. We assume that new characters are already aware of this history.

Relations refer to the current relationship between new and original characters at the beginning of a sequel. After introducing a new character, the relationships between characters can change.

Guideline refers to the future tendency of how a character behaves within the sequel. The character's personality traits, history, and relations can decide this guideline about the future.

2.2 History Extraction

The first step of sequel generation is extracting history from the story. We extracted the history with two distinct levels: world-level and character-level. World-level history pertains to information about general events or the worldview, while character-level history corresponds to the individual actions and the internal psychology of each individual character. Specifically, we extracted those histories after chunking, as the length of the original story may exceed the context length of the LLM. We split the original story into chunks of 100 sentences each. From each chunk, the LLM first extracts world-level history, and then further refine it into character-level histories for each original character. Subsequently, we aggregate two histories from all chunks to construct whole world-level and character-level histories.

2.3 Sequel Character Augmentation

Next, SeqUAL generates new characters to increase the diversity of a sequel. We augment history by creating a new and unique character with distinct personalities for each sequel. Even with the same original story, we believe that the injection of different personalities might lead to vastly different sequel plots. Specifically, we generate future guidelines of actions and present relations between the new character and each original character.

Constructing future guidelines requires specifying personality traits. We automatically generate a short personality description based on Big Five theory (John et al., 1999). A user can set strength of five personality traits using three levels: High, Low, or Unspecified. It is sufficient to use these three, because characters in stories often have exaggerated traits. As a result, the user can set 242 ($= 3^5 - 1$) possible combinations, excluding the case when all traits are left unspecified. Next, SeqUAL fed this

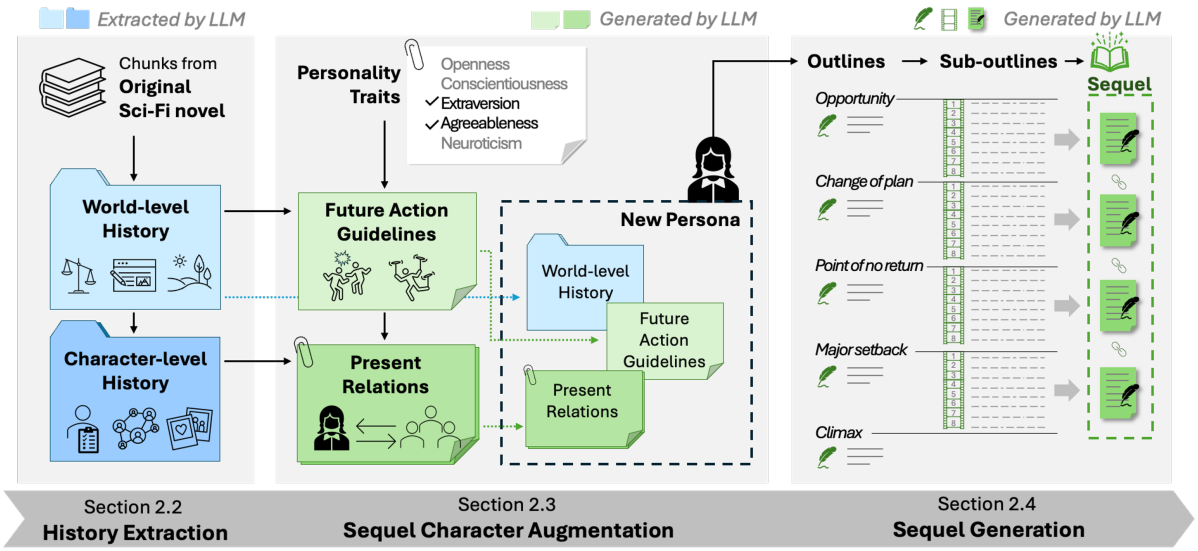

Figure 1: Overview of the SeqUAL Framework

specification and world-level history into an LLM and generated future guidelines.

The generated guideline is then combined with the character-level history to generate relations. As behavioral patterns of the new characters can affect interactions between the new character and the original one, we let an LLM predict potential relations. Specifically, we provide guidelines for the new one and character-level histories for all characters as input, and LLM outputs a list of possible interactions between the new and each original character.

2.4 Consistent Sequel Generation

As a result of character augmentation, we obtain a *persona* of a newly generated character to unfold a sequel. Using the persona, we unfold a sequel. Inspired by previous research (Qi et al., 2024), we adopt a three-step procedure to generate sequels.

First, to construct the events, we generate an outline of a sequel. Following Tian et al. (2024), we prompt an LLM to generate a five-stage outline based on the new persona. This five-stage outline represents the progression of a sequel, as follows: ① opportunity, ② change of plans, ③ point of no return, ④ major setback, and ⑤ climax. Taking the new persona as input, the LLM then constructs an outline of events for each stage of this progression.

Second, to facilitate more detailed story generation, we flesh out the outline with detailed sub-outlines. For each transition between two consecutive outline stages, we use an LLM to generate eight sub-outlines. These represent the intervening events that ensure a natural sequel flow. As there are five main stages, this process yields a total of 32 ($= 4 \times 8$) sub-outlines. We then transform each sub-outline with a single-paragraph story using an LLM. Here, we conditioned the LLM on the immediately preceding paragraph when writing the next one to ensure coherent continuation of the story.

Third, to construct the final story, we concatenate all paragraphs. As LLMs occasionally re-produce unnecessary words from input prompts (e.g., step or outline), we removed such superfluous terms with a post-processing.

3 Experiments

We conduct an experiment to examine whether a new persona can be introduced and augmented to generate a sequel by unfolding. Specifically, we evaluate both *diversity* and *consistency* of generated sequels. In following paragraphs, we describe datasets, analysis, models, and implementations.

3.1 Dataset

We manually constructed a dataset of Korean Sci-Fi novels for two reasons. First, we wanted to exclude the effect of parametric knowledge. As popular English novels are likely to have been seen by language models during their pretraining, LLMs might utilize parametric knowledge to generate a more nuanced sequel. As such, utilization may lead to an incorrect interpretation of the result on consistency, we decided to use novels where LLMs may not have been trained. So, we chose novels published after 2022 that were originally written in Korean. Second, we wanted to make a clear distinc-

tion between consistent sequel from inconsistent ones during the evaluation. As Sci-Fi novels introduce distinctive and unconventional worlds, it is easy to check whether the sequel follows the original worldview or not. So, we decided to use Sci-Fi novels instead of other genres.

Specifically, we selected 10 works awarded the Korean Science Fiction Awards, which is similar to Jang and Jung (2024). We used awarded works as superior-quality novels are more likely to contain a rich worldview and well-defined characters. We used published versions without any modifications[1], and selected works are listed in Appendix A. As a result, the dataset contains 10 novels whose average number of sentences is 733.5 (min 316, max 1,201). The number of syllables per work was 26,796.8 on average (min 15,878, max 49,171).

3.2 Analysis method

To assess whether SeqUAL can address diversity-consistency trade-off, we measured both diversity and consistency. For each story, we generated 20 sequels using SeqUAL, and we averaged results for each measure. Additionally, we measured diversity between consecutive outline stages to identify where diversity of a story emerges in general.

Diversity. We employ two metrics that measure overlap among the 20 sequels generated from the same original story: Event Content Diversity (ECD) and Event Participant Diversity (EPD). We need key event $k_{o,i}$ and their participants $p_{o,i}$ for each sequel $s_i(o)$ (i=1, ... 20) of original story o. Using GPT-4.1-mini[2], we extracted $k_{o,i}$ and $p_{o,i}$ for each $s_i(o)$. And, we obtained sentence embedding using 'all-MiniLM-L6-v2[3]' and computed cosine similarity $\cos(k_{o,i}, k_{o,j})$ and $\cos(p_{o,i}, p_{o,j})$ for each pair of sequels originating from the same story; That is,

$$\begin{aligned} \text{ECD} &= \mathbb{E}_{o,\, i\neq j}[1-\cos(k_{o,i}, k_{o,j})], \\ \text{EPD} &= \mathbb{E}_{o,\, i\neq j}[1-\cos(p_{o,i}, p_{o,j})]. \end{aligned}$$

Note that higher ECD and EPD indicates high diversity. To obtain stability of measurement, we averaged 10 runs of diversity measurement.

Furthermore, we segmented the sequels according to outline stages and measured ECD and EPD for each segment. As mentioned earlier, we used five stages: ① opportunity, ② change of plans, ③ point of no return, ④ major setback, and ⑤ climax.

Consistency. We employ two metrics that measure the alignment between worldview and sequel: Consistency on Societal Singularity (CSS) and Consistency on Key Events (CKE). Here, a societal singularity refers to an institutional characteristic that differentiates the social setting of the original novel from reality, while a key event denotes a pivotal incident that drives the plot in the original novel. For each original story o, we extracted a set of societal singularities Σ_o and a key event K_o using GPT-4.1-mini. We then queried GPT-4.1-mini to classify whether each generated sequel $s_i(o)$ is consistent with $\sigma \in \Sigma_o$ and $k \in K_o$ into three categories: *consistent*, *inconsistent*, or *undecidable*. Since not all elements of the original story necessarily appear in the sequels, we computed CSS and CKE on consistent or undecidable categories.

$$\begin{aligned} \text{CSS}(x) &= P(\text{GPT}(\sigma, s_i(o)) = x), \\ \text{CKE}(x) &= P(\text{GPT}(k, s_i(o)) = x), \end{aligned}$$

where $\text{GPT}(\cdot,\cdot)$ indicates the classification result and x is either consistent or undecidable. Note that CSS(con) and CKE(con) count cases where the worldview is explicitly described, whereas CSS(und) and CKE(und) count cases without any violation or explicit mention. We used $|\Sigma_o| = |K_o| = 10$ for each original work o.

3.3 Models

We implemented SeqUAL using two LLMs: ExaOne-3.5 32B Instruct (Research et al., 2024) and Gemma-2 27B Instruct (Team et al., 2024). We selected them because they are open-weight models and showed state-of-the-art performance on KMMLU (Son et al., 2024), a benchmark for Korean language understanding; ExaOne and Gemma achieved scores of 57 and 53.8, respectively.

Also, as ablated models, we prepared two variants of SeqUAL. The first ablated model is SeqUAL without Sequel Character Augmentation (SeqUAL-C). This model generates outlines and a sequel directly from the extracted history, allowing us to observe the impact of character augmentation by comparing it with SeqUAL. The second ablated model is SeqUAL without History Extraction (SeqUAL-H). This model constructs new characters and generates sequels directly from the original story, enabling us to evaluate the effect of history extraction by comparing it with SeqUAL.

[1] Due to copyright restrictions, the exact contents of the novels cannot be disclosed.

[2] https://openai.com/index/gpt-4-1/

[3] https://huggingface.co/sentence-transformers/all-MiniLM-L6-v2

	Diversity		Consistency	
	ECD	EPD	CSS (con)	CKE (con)
SeqUAL (ExaOne)	0.383	0.203	0.233	0.155
w/o Character	0.319	0.126	0.318	0.306
w/o History	0.457	0.235	0.152	0.103
SeqUAL (Gemma)	0.375	0.170	0.316	0.121

Table 1: Results of diversity and consistency

	CSS		CKE	
	(con)	(und)	(con)	(und)
SeqUAL (ExaOne)	0.233	0.690	0.155	0.668
w/o Character	0.318	0.624	0.306	0.544
w/o History	0.152	0.772	0.103	0.752
SeqUAL (Gemma)	0.316	0.649	0.121	0.800

Table 2: Results of consistency (consistent/ undecidable)

3.4 Implementation Detail

All experiments were conducted on two workstations. One machine contains four A6000 GPUs and 128GB of RAM, and the other contains 6000 Ada GPUs and 128GB of RAM. We used Python 3.10 and 3.11 for running the experiments, with `accelerate` and `transformers` libraries. For all responses, the temperature was set to 1.0.

4 Result and Discussion

As a result, we identified three key findings: (1) SeqUAL addresses the diversity-consistency trade-off that arises during sequel generation; (2) Character Augmentation module contributes to maintaining diversity across the entire story flow; and (3) History Extraction module contributes to preserving the worldview of the original story.

4.1 Diversity-Consistency Trade-off

We observe the expected trade-off between diversity and consistency during the process of sequel generation in Table 1. Examining ECD diversity scores of the ExaOne model, we observe an increasing trend in the order of SeqUAL w/o Character (0.319), SeqUAL (0.383), and SeqUAL w/o History (0.457). Conversely, CSS(con) consistency scores increase in the order of SeqUAL w/o History (0.152), SeqUAL (0.233), and SeqUAL w/o Character (0.303). A similar trade-off phenomenon is observed between EPD and CKE(con) scores.

We attribute this phenomenon to the inherent characteristics of sequel generation. Efforts to enhance diversity in sequels generally require an introduction of new events or characters that do not present in the existing worldview, which may reduce consistency. Conversely, attempts to increase consistency often restrict the introduction of new events or characters outside the original worldview, making it challenging to achieve high diversity.

In the process of balancing two objectives of diversity and consistency, the two modules of SeqUAL perform a central role. We designed Sequel Character Augmentation module in SeqUAL to ensure diversity, while History Extraction module was designed to facilitate the maintenance of consistency. The experimental results demonstrate that each module operates as we expected. Specifically, in the ablated model without the Character Augmentation module, all diversity measures decrease, even though we allowed the ablated model to introduce new characters, which are specified in the prompt. This indicates that establishing relations between new and existing characters is an essential step for reliably securing diversity. Conversely, in the ablated model without the History Extraction module, all consistency measures decline, though we provided the entire source of the original story to the model while generating the sequel. This result suggests that extracting key information from the original story and utilizing it as memory is more effective for maintaining consistency than providing the entire source of the original story.

Additionally, we discovered two minor findings. First, EPD scores tended to be relatively lower than ECD scores as we added a new character to the short story with a small existing character pool. SeqUAL showed EPD of 0.203, and even without the History Extraction module, the score remained low at 0.236. Considering that EPD measures how well the extraction of characters participating in key events is, it is likely that the small character pool of the original short novel likely contributed to the lower diversity. Employing multiple characters may help in achieving higher diversity, but this must be balanced against the diversity-consistency tradeoff. Second, the SeqUAL framework demonstrated similar performance regardless of LLMs, confirming that the architecture is not dependent on a specific LLM. Here, SeqUAL (Gemma) exhibited comparable diversity scores to SeqUAL (ExaOne), with an ECD of 0.375 and an EPD of

	ECD				EPD			
Outline stages:	①-②	②-③	③-④	④-⑤	①-②	②-③	③-④	④-⑤
SeqUAL (ExaOne)	0.448	0.409	0.404	0.425	0.304	0.316	0.318	0.303
w/o Character	0.402	0.386	0.333	0.389	0.204	0.166	0.167	0.187
w/o History	0.488	0.477	0.459	0.472	0.277	0.266	0.261	0.280
SeqUAL (Gemma)	0.424	0.456	0.441	0.441	0.224	0.194	0.202	0.223

Table 3: Results of diversity according to the outline stages of the generated sequels.

Q. Social class and racial discrimination structure: The hierarchy is clear, including **Coolies** (seemingly former slaves or laborer class), nobles, and sheriffs, with the Coolie class facing restrictions in communication and legal rights.	
Original Story	"Doesn't a **Coolie** bear responsibility? A **Coolie** is an existence outside of responsibility. Rights and responsibilities always go hand in hand. Do you understand?"
SeqUAL	"They approach figures like Buford and York to obtain additional information. They maintain awareness of authority and social status in their interpersonal strategies, cooperating for practical benefits. Interactions with **Coolies** involve gathering information through temporary cooperation while avoiding intimacy."
w/o Character	"The **Coolies** were so highly guarded that direct conversation with the matron seemed impossible, yet Buford did not give up."
w/o History	NOWHERE IN THE TEXT does content such as "social class and racial discrimination structure" or "**Coolies**, nobles, sheriffs" appear.

Table 4: Examples of consistency in the generated sequels with ExaOne.

0.170. While Gemma's CSS(con) was higher at 0.316, its CKE(con) was lower at 0.121. We assume that the differences in consistency between two backbone LLMs arise from the difference between how they gave internal attention to social singularity versus specific events.

4.2 Diversity with New Characters

Beyond examining the overall diversity of events in the generated sequels, we investigated which outline stages in the sequels were most influenced by our proposed method. Table 3 shows the result. When we compare SeqUAL w/o Character and w/o History, the differences observed were 0.086 between ①–②, 0.091 between ②–③, 0.126 between ③–④, and 0.083 between ④–⑤. These differences indicate that the most significant difference mainly occurs during outline stages, where tension of a sequel escalates: between ③ and ④. In general, LLMs tend to converge to similar endings when generating a story(Xu et al., 2025; Wenger and Kenett, 2025). Thus, also when generating sequels based on the same original story, LLMs reveal such a tendency. However, as the Sequel Character Augmentation module within the SeqUAL framework is designed to consider diverse combinations of personality traits, employing the SeqUAL framework effectively prevents repetitive or similar events even in mid-to-late outline stages.

Similar patterns can also be observed in generated sequel samples. We examined some samples of stories generated by SeqUAL, SeqUAL w/o Character, and SeqUAL w/o History to investigate how events unfold differently in the same original story. An example is shown in Table 5. The results indicate that SeqUAL and w/o History generated diverse ways to resolve conflicts, whereas w/o Character primarily generated similar methods. For example, SeqUAL generated a story where the new character 'Emilia' leverages her knowledge

SeqUAL	In the process, Emilia, using her knowledge of constellations, suggested safe routes for night travel to the other members, teaching everyone how to find direction together. This mutual exchange further solidified Emilia's knowledge and leadership within the group, and **her special cooperation with the Coolie built a deep trust between them.**	Sensing that the Coolie had noticed something, Emilia grew increasingly restless. Seizing the moment, she suddenly raised her voice and insisted that they stop repairing the carriage. "Rather than waiting here any longer, I think it would be safer to head to the nearest town and ask for help immediately." **Her words heightened the tension within the team, and the other members glanced at one another in confusion at her sudden suggestion.**
w/o Character	"First, we need to move my sister-in-law to a safe place, and in the process, carefully observe the actions of the Coolies." Even in the midst of a new crisis, Buford never let go of the thread of hope — **protecting his sister-in-law while exploring the uncharted realms of time travel.**	Despite the complicated reality, **Buford resolved to protect his sister-in-law while also uncovering the secrets of the future.** When the Coolie suddenly appeared at the cattle theft site, Buford rushed to shield his sister-in-law, but the Coolie's gaze was already fixed on her.
w/o History	"It's important that we set our operating principles based on each person's ideas and experiences." As she felt the villagers' eyes converge into one, **Emilia once again took the lead in steering the atmosphere of the meeting.**	Usually quiet and supportive from behind the scenes, this time too she chose to observe the situation rather than step forward directly. **But as the tension within the team heightened and cracks in their cooperation widened, her heart grew increasingly heavy.**

Table 5: Examples of diversity in the generated sequels with ExaOne.

and leadership to engage in a special collaboration with 'Coolie'. Meanwhile, another example of SeqUAL shows the new character 'Emily' raising her voice to insist that an ongoing task be halted, creating confusion among the team members; the difference in these two demonstrates variation in how events unfold. Similarly, SeqUAL w/o History generated two distinct sequels. In one story, the new character 'Emily' guides the mood of a meeting, while in the other, she experiences a growing sense of personal burden as team tension escalates. But, on the other hand, SeqUAL w/o Character failed to generate diverse stories; instead, it generated two main types of sequels. One type of story focused on protecting the existing character 'Hyungsu' while maintaining hope, and the other focused on protecting 'Hyungsu' while attempting to uncover a secret. Though the intention is different, the plot reveals that the character aims to achieve the same overall goal. Overall, we concluded that the Sequel Character Augmentation module, which introduces characters with diverse personalities, can generate new events that create meaningful differences, particularly during moments of heightened tension.

4.3 Consistency with History

History Extraction module of SeqUAL helps to restore societal structures within the worldview of the original story. As shown in Table 2, the module contributes to improving the consistency of societal structures between the original story and generated sequels. Increasing the frequency of leveraging societal structures from the original story, CSS(con) score rises by 0.081, from 0.152 to 0.233. Furthermore, the CSS of undecidable cases decreases by 0.082, from 0.772 to 0.690. This phenomenon suggests that the History Extraction module organizes and conveys societal structures, thereby increasing both their mention frequency and usability within the sequel, even when such structures were not explicitly detailed in the original story.

Similar patterns can also be observed in generated sequel samples. We examined how the generated sequels described the societal structures mentioned in the original story. An example is shown in Table 4. In the examples, the word 'Coolies' refers to a class system that does not exist in reality. In sequels generated with the History Extraction module (SeqUAL and it w/o Character), Coolies are accurately mentioned. Meanwhile, in the ablated model without History Extraction, references to Coolies are omitted in the sequel. Although the original story was provided during the generation, the weight given to the word 'Coolies' was relatively diminished due to other content words in the original story, leading to the loss of these references. That is, the presence of the History Extraction module allows SeqUAL to generate new events that differ from the original story while maintaining consistency with events from the original story.

5 Related Work

Story generation studies have strived to generate coherent and plausible stories. However, approaches that prioritize plausibility and consistency tend to generate similar stories, thereby limiting their applicability to story generation tasks where diversity is crucial. For instance, (Qi et al., 2024) argued that training story generation via reinforcement learning could facilitate exploration to enhance the diversity of stories, while exploitation would help maintain internal consistency. Similarly, other studies focused on generating stories centered on plausibility and coherence (Tian et al., 2024; Han et al., 2025). Nonetheless, the resulting stories were generally similar, making these approaches less suitable for tasks that require highly diverse outputs.

Thus, recent studies have employed various mechanisms to ensure narrative diversity (Ran et al., 2025; Wang and Kreminski, 2024; Jiang et al., 2024). Some researchers attempted to secure diversity by generating story outlines. For example, Wang and Kreminski (2024) utilized Answer Set Programming (ASP) to produce as many diverse story outlines as possible based on a given topic, and then constructed varied stories from the generated outlines. While this approach contributed to generating diverse and plausible stories, it is not suitable for the sequel generation task targeted in this study, since an original story already exists, mechanisms are required to preserve the core characters and event structure of the source story.

To preserve the character composition of the original story, leveraging personas can serve as a potential solution. Prior studies have reported that language models are capable of persona-based dialogue simulation (Cohen et al., 2025; Wang et al., 2025b; Hu and Collier, 2024), suggesting that it is possible to maintain diversity by utilizing character personas. Accordingly, some researchers have explored story generation based on personas. For example, Ran et al. (2025) automatically constructed personas from a given original story and simulated dialogues to provide material for story generation. Han et al. (2025) demonstrated that it is possible to generate spin-off stories by reusing characters from the original story. While these approaches may assist in generating sequels, they do not account for the distinct characteristics of sequels as opposed to spin-offs, such as the introduction of new characters. To incorporate new personas that are consistent with the original story, it is necessary to consider existing character relationships and the social structure of the original story, as we demonstrated in the SeqUAL framework.

6 Conclusion

This study was conducted to address the problem of balancing diversity and consistency while generating sequels, which had not been sufficiently explored in prior research on story generation. In particular, when narratives were generated with large language models, there was a tendency for stories to converge toward similar plot developments and endings. Accordingly, this study proposed SeqUAL framework and introduced two core components: the History Extraction module and the Sequel Character Augmentation module. History Extraction module was designed to extract and summarize the worldview and original characters of the original story, thereby helping to maintain consistency between the original story and its sequels. Sequel Character Augmentation module was designed to enhance event and character diversity in the sequel generation process by considering various combinations of personality traits and unfolding events and interactions appropriate to those combinations.

To validate the effectiveness of the proposed framework, we analyzed the diversity and consistency of generated sequels. When we measure the diversity of generated sequels, we found that particularly in the mid-to-late stages of the sequels where tension escalated, there was a significant increase in diversity compared to models without the Character Augmentation module or without the History Extraction module. This indicates that repetitive plot developments of existing studies could be prevented even when we generate sequels based on the same original story. Also, when we measured the consistency of generated sequels, the results demonstrated that when the History Extraction module was included, the worldview consistency was better preserved. In addition, the model produced descriptions more frequently that were more closely aligned with the original worldview.

In conclusion, this study demonstrated that by incorporating the History Extraction module and the Seqeul Character Augmentation module, it was possible to identify a trade-off point in sequel generation that increased diversity while maintaining consistency. We believe that these findings provide a foundation for further applications and broader applicability of LLM-based story generation.

Limitations

While acknowledging the limitations of this study is essential, it reflects our commitment to transparency and honesty in our research. There are four limitations that we acknowledged. First, we evaluated generated stories primarily based on quantitative metrics such as ECD or EPD, highlighting the impact of character augmentation on sequel diversity. Although we did not thoroughly assess the qualitative aspects of the sequels, the quality of the generated sequels is good enough to read, as evidenced by some sample stories. Our confirmation that the SeqUAL framework can effectively balance diversity and consistency lays the groundwork for promising future research.

Second, it's worth noting that the diversity scores observed in the ablation experiment did not show drastic differences. This suggests that, while the SeqUAL framework is effective in balancing diversity and consistency, there may be more effective approaches for enhancing the diversity of the story. We see this as an opportunity for future exploration.

Third, we found that increases in diversity primarily occur during points of heightened story tension. By developing methods to boost diversity during climactic moments, we could pave the way for generating more varied conclusions in sequels.

Fourth, we constructed our dataset using Korean novels to minimize reliance on parametric knowledge. While our framework is not limited by language—making experiments with English novels feasible—we did not conduct such studies. This opens up a compelling possibility for future research to delve deeper into cross-linguistic applications and further enrich our understanding of narrative generation. Though there are limitations, we believe the insights gained from this study will contribute significantly to the community and inspire further advancements in this field.

Ethical Considerations

During the writing process of this paper, AI assistants supported our writing. We employed Grammarly for grammar checking. Also, we used GPT and Gemini for polishing our writing and enhancing the presentation clarity of our paper. These tools were used for editorial purposes only.

References

Myke C. Cohen, Zhe Su, Hsien-Te Kao, Daniel Nguyen, Spencer Lynch, Maarten Sap, and Svitlana Volkova. 2025. Exploring big five personality and ai capability effects in llm-simulated negotiation dialogues. *Preprint*, arXiv:2506.15928.

Nellia Dzhubaeva, Katharina Trinley, and Laura Pissani. 2025. Unstructured minds, predictable machines: A comparative study of narrative cohesion in human and LLM stream-of-consciousness writing. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 4: Student Research Workshop)*, pages 1079–1096, Vienna, Austria. Association for Computational Linguistics.

Xueran Han, Yuhan Liu, Mingzhe Li, Wei Liu, Sen Hu, Rui Yan, Zhiqiang Xu, and Xiuying Chen. 2025. Pastiche novel generation creating: Fan fiction you love in your favorite author's style. *Preprint*, arXiv:2502.15616.

Tiancheng Hu and Nigel Collier. 2024. Quantifying the persona effect in LLM simulations. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 10289–10307, Bangkok, Thailand. Association for Computational Linguistics.

Woori Jang and Seohyon Jung. 2024. Evaluating LLM performance in character analysis: A study of artificial beings in recent Korean science fiction. In *Proceedings of the 4th International Conference on Natural Language Processing for Digital Humanities*, pages 339–351, Miami, USA. Association for Computational Linguistics.

Hang Jiang, Xiajie Zhang, Xubo Cao, Cynthia Breazeal, Deb Roy, and Jad Kabbara. 2024. PersonaLLM: Investigating the ability of large language models to express personality traits. In *Findings of the Association for Computational Linguistics: NAACL 2024*, pages 3605–3627, Mexico City, Mexico. Association for Computational Linguistics.

Oliver P John, Sanjay Srivastava, and 1 others. 1999. The big-five trait taxonomy: History, measurement, and theoretical perspectives.

Congda Ma, Kotaro Funakoshi, Kiyoaki Shirai, and Manabu Okumura. 2023. Coherent story generation with structured knowledge. In *Proceedings of the 14th International Conference on Recent Advances in Natural Language Processing*, pages 681–690, Varna, Bulgaria. INCOMA Ltd., Shoumen, Bulgaria.

Feiteng Mu and Wenjie Li. 2024. Generating contrastive narratives using the brownian bridge process for narrative coherence learning. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 6538–6555, Bangkok, Thailand. Association for Computational Linguistics.

Kyeongman Park, Minbeom Kim, and Kyomin Jung. 2025. A character-centric creative story generation via imagination. In *Findings of the Association for Computational Linguistics: ACL 2025*, pages 1598–1645, Vienna, Austria. Association for Computational Linguistics.

Qianqian Qi, Lin Ni, Zhongsheng Wang, Libo Zhang, Jiamou Liu, and Michael Witbrock. 2024. Epic-level text generation with llm through auto-prompted reinforcement learning. In *2024 International Joint Conference on Neural Networks (IJCNN)*, pages 1–8.

Yiting Ran, Xintao Wang, Tian Qiu, Jiaqing Liang, Yanghua Xiao, and Deqing Yang. 2025. BOOKWORLD: From novels to interactive agent societies for story creation. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 15898–15912, Vienna, Austria. Association for Computational Linguistics.

Hannah Rashkin, Asli Celikyilmaz, Yejin Choi, and Jianfeng Gao. 2020. PlotMachines: Outline-conditioned generation with dynamic plot state tracking. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 4274–4295, Online. Association for Computational Linguistics.

LG AI Research, Soyoung An, Kyunghoon Bae, Eunbi Choi, Kibong Choi, Stanley Jungkyu Choi, Seokhee Hong, Junwon Hwang, Hyojin Jeon, Gerrard Jeongwon Jo, Hyunjik Jo, Jiyeon Jung, Yountae Jung, Hyosang Kim, Joonkee Kim, Seonghwan Kim, Soyeon Kim, Sunkyoung Kim, Yireun Kim, and 14 others. 2024. Exaone 3.5: Series of large language models for real-world use cases. *Preprint*, arXiv:2412.04862.

Guijin Son, Hanwool Lee, Sungdong Kim, Seungone Kim, Niklas Muennighoff, Taekyoon Choi, Cheonbok Park, Kang Min Yoo, and Stella Biderman. 2024. Kmmlu: Measuring massive multitask language understanding in korean. *Preprint*, arXiv:2402.11548.

Gemma Team, Morgane Riviere, Shreya Pathak, Pier Giuseppe Sessa, Cassidy Hardin, Surya Bhupatiraju, Léonard Hussenot, Thomas Mesnard, Bobak Shahriari, Alexandre Ramé, Johan Ferret, Peter Liu, Pouya Tafti, Abe Friesen, Michelle Casbon, Sabela Ramos, Ravin Kumar, Charline Le Lan, Sammy Jerome, and 179 others. 2024. Gemma 2: Improving open language models at a practical size. *Preprint*, arXiv:2408.00118.

Yufei Tian, Tenghao Huang, Miri Liu, Derek Jiang, Alexander Spangher, Muhao Chen, Jonathan May, and Nanyun Peng. 2024. Are large language models capable of generating human-level narratives? In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 17659–17681, Miami, Florida, USA. Association for Computational Linguistics.

Oseremen O Uduehi, Razvan C Bunescu, and Laurie L Lyda. 2024. Short story generation through autoregressive transfer of narrative continuations. In *International Conference on Computational Creativity*.

Saranya Venkatraman, Nafis Irtiza Tripto, and Dongwon Lee. 2025. CollabStory: Multi-LLM collaborative story generation and authorship analysis. In *Findings of the Association for Computational Linguistics: NAACL 2025*, pages 3665–3679, Albuquerque, New Mexico. Association for Computational Linguistics.

Phoebe J. Wang and Max Kreminski. 2024. Guiding and diversifying llm-based story generation via answer set programming. *Preprint*, arXiv:2406.00554.

Qianyue Wang, Jinwu Hu, Zhengping Li, Yufeng Wang, Daiyuan Li, Yu Hu, and Mingkui Tan. 2025a. Generating long-form story using dynamic hierarchical outlining with memory-enhancement. In *Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 1352–1391, Albuquerque, New Mexico. Association for Computational Linguistics.

Yilei Wang, Jiabao Zhao, Deniz S. Ones, Liang He, and Xin Xu. 2025b. Evaluating the ability of large language models to emulate personality. *Scientific Reports*, 15(1):519.

Zhihua Wen, Zhiliang Tian, Wei Wu, Yuxin Yang, Yanqi Shi, Zhen Huang, and Dongsheng Li. 2023. GROVE: A retrieval-augmented complex story generation framework with a forest of evidence. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 3980–3998, Singapore. Association for Computational Linguistics.

Emily Wenger and Yoed Kenett. 2025. We're different, we're the same: Creative homogeneity across llms. *Preprint*, arXiv:2501.19361.

Zhuohan Xie, Trevor Cohn, and Jey Han Lau. 2023. The next chapter: A study of large language models in storytelling. In *Proceedings of the 16th International Natural Language Generation Conference*, pages 323–351, Prague, Czechia. Association for Computational Linguistics.

Weijia Xu, Nebojsa Jojic, Sudha Rao, Chris Brockett, and Bill Dolan. 2025. Echoes in ai: Quantifying lack of plot diversity in llm outputs. *Proceedings of the National Academy of Sciences*, 122(35):e2504966122.

Qiang Yi, Yangfan He, Jianhui Wang, Xinyuan Song, Shiyao Qian, Xinhang Yuan, Li Sun, Yi Xin, Jingqun Tang, Keqin Li, and 1 others. 2025. Score: Story coherence and retrieval enhancement for ai narratives. *arXiv preprint arXiv:2503.23512*.

A Dataset

We selected ten Korean science fiction novels published between 2022 and 2024, all of which were awarded the Korean Science Fiction Awards. Table 6 shows the title of each novel along with its corresponding number of sentences and character counts (including spaces).

Title	Author	Sentences	Characters
옛날 옛적 판교에서는	김쿠만	566	22,476
루나	서윤빈	608	22,906
신께서는 아이들을	성수나	662	18,306
두 개의 세계	박민혁	1,201	40,407
삼사라	조서월	316	15,878
제니의 역	최이아	395	16,380
발세자르는 이 배에 올랐다	허달립	622	18,928
하늘의 공백	정현수	1,162	49,171
피폭	존벅	1,047	40,838
달은 차고 소는 비어간다	최우준	756	22,678

Table 6: Statistics of each novel used in our dataset

B Propmpt for SeqUAL framework

We used the following Korean prompts for sequel generation. To help readers understand, we also present English translation for each prompt.

B.1 History extraction

See Tables 7 and 8 for the prompts. Here, we want to extract history of world from the input chunk, which is represented as <chunk>. And then, the world-level history is used for extracting character-level histories.

World-level history	
Korean	다음 글에서 서술되는 세상에 관한 객관적인 정보를 추출하여라. 형식은 1. 2. 3. <chunk>
English	Extract objective information about the world described in the following text. Format: 1. 2. 3. <chunk>

Table 7: Instruction template for world-level history

B.2 Sequel character augmentation

See Tables 9 and 10 for the prompts. Here, we want to setup a new character based on the given information. First, we extracted future action guidelines based on past history and personality traits that we extracted. Second, we generated relationship between existing characters and new ones, by providing future action guidelines, character-level history, and personality traits.

Character-level history	
Korean	해당 글에 등장하는 등장인물들 각각에 대한 정보를 추출하여라. 형식은 1. 2. 3. <World_level_history>
English	Extract information about each character appearing in the given text. Format: 1. 2. 3. <World_level_history>

Table 8: Instruction template for character-level history

Future Action Guidelines	
Korean	선택된 성격 특성 2개는 **새로운** 인물의 성격 특성이다. 새로운 인물이 미래에 어떤 행동을 할 수 있을지 예측하여라. 주어지는 기존 등장인물들의 정보를 바탕으로 주어진 세계를 파악한다. <높은 성격 특성: {trait_1}, 낮은 성격 특성: {trait_2}> <World_level_history>
English	The two selected personality traits represent the **new** character's personality. Predict what actions this new character might take in the future, based on the given information about the existing characters and the described world. <High trait: {trait_1}, Low trait: {trait_2}> <World_level_history>

Table 9: Instruction template for future action guideline

B.3 Sequel generation

See Tables from 11 to 14. Here, we want to generate a new sequel based on the newly settled information. Table 11 shows how we provided the information about five stages of outline. Then, we generated outline based on persona information, as shown in Table 12. Using the outline generated, we further elaborated those outline with prompts stated in Table 13. Finally, we generated detailed story using sub-outlines, as shown in Table 14.

	Present Relations
Korean	다음은 소설에 등장하는 기존 등장인물들의 정보다. 선택된 성격 특성 2개는 **새로운** 인물의 성격 특성이다. 해당 성격 특성을 가지는 새로운 인물이 기존에 존재하는 등장인물과 미래에 어떤 상호작용을 할 수 있을지 예측하여라. <높은 성격 특성: {trait_1}, 낮은 성격 특성: {trait_2}> 기존 등장인물 정보: <Character_level_history> 예상되는 미래 행동: <future_actions>
English	The following provides information about the existing characters in the novel. The two selected personality traits represent the **new** character's personality. Predict how this new character, with the given traits, might interact with the existing characters in the future. <High trait: {trait_1}, Low trait: {trait_2}> Information about existing characters: <Character_level_history> Predicted future interactions: <future_actions>

Table 10: Instruction template for current relation prediction

	Outline Instructions
Korean	"opportunity": "주인공이 새로운 목표나 기회를 발견하는 단계이다. 인물의 동기와 기대가 드러나는 사건을 중심으로 작성하라." "change_of_plans": "예상치 못한 사건으로 계획이 바뀌는 단계이다. 주인공이 당황하거나 선택을 강요받는 전환점을 포함하라." "point_of_no_return": "돌아올 수 없는 지점에 이르는 단계이다. 주인공이 중요한 결정을 내리거나 위험을 감수하는 상황을 강조하라." "major_setback": "큰 좌절이나 위기를 맞이하는 단계이다. 주인공이 실패하거나 위기에 처해 극복이 필요함을 보여주어야 한다." "climax": "이야기의 절정으로, 주인공이 최대의 갈등을 마주하고 결정을 내리는 순간을 극적으로 표현하라."
English	"opportunity": "This stage represents the moment when the protagonist discovers a new goal or opportunity. Focus on events that reveal the character's motivations and expectations." "change_of_plans": "An unexpected event causes the original plan to change. Include a turning point where the protagonist feels confused or is forced to make a choice." "point_of_no_return": "The stage where the protagonist reaches a point of no return. Emphasize situations in which they make a crucial decision or take a significant risk." "major_setback": "A stage of major failure or crisis. Depict how the protagonist experiences a setback or faces danger that requires resilience and recovery." "climax": "The climax of the story, where the protagonist faces the greatest conflict and makes a decisive choice. Portray this moment with dramatic intensity."

Table 11: Instruction template for outline instructions

	Persona-to-Outline
Korean	**페르소나 → 아웃라인** 단계이다. 주어진 페르소나를 바탕으로 이야기를 구성할 것이다. 이야기는 총 5개의 큰 흐름(opportunity, change_of_plans, point_of_no_return, major_setback, climax)으로 이루어진다. 현재 단계: **<outline_type>** 다음 페르소나를 기반으로 이 단계에서 발생할 수 있는 **핵심 사건**을 설명하는 outline을 **한국어로 한 문장** 생성하라. future actions에 서술되어 있는 페르소나를 중심으로 작성하면 된다. [페르소나] <persona>
English	This is the Persona → Outline stage. The story will be constructed based on the given persona. The narrative consists of five main phases (opportunity, change_of_plans, point_of_no_return, major_setback, and climax). Current phase: ** <outline_type>** Generate a one-sentence outline in Korean that describes the key event likely to occur in this phase, based on the given persona. Focus on the persona described in the future actions. [Persona] <persona>

Table 12: Instruction template for Persona-to-Outline transformation

	Outline-to-Sub-outline
Korean	**아웃라인 → 세부 아웃라인** 단계이다. 다음 두 아웃라인 사이를 자연스럽게 연결할 수 있는 **세부 아웃라인을 한국어로 8문장 생성**하여라. 출력 형식은 1. 2. 3. 4. 5. 6. 7. 8. [이전 아웃라인]: <Previous Outline> [다음 아웃라인]: <Next Outline> 세부 상황:
English	This is the **Outline → Sub-Outline** stage. Generate 8 sentences in Korean that can naturally connect the two outlines below. Output format: 1. 2. 3. 4. 5. 6. 7. 8. [Previous Outline]: <Previous Outline> [Next Outline]: <Next Outline> Detailed situation:

Table 13: Instruction template for Outline-to-Sub-outline transformation

	Sub-outline-to-Story
Korean	**세부 아웃라인 → 단락 스토리** 단계이다. 다음 세부 아웃라인을 바탕으로 이야기 단락을 **한국어로 생성**하여라. 이야기는 이전 단락과 자연스럽게 이어져야 하며, 현재 아웃라인과 세부 아웃라인 내용이 자연스럽게 언급되어야 한다. 결과는 **문장으로만 구성**되어 있어야 하며, 부연 설명은 제외한다. [이전 단락]: <Previous Paragraph> [현재 아웃라인]: <Current Outline> [세부 아웃라인]: <Sub-outline Sentence>
English	This is the **Sub-outline → Story** stage. Generate a story paragraph in Korean based on the given sub-outline below. The story should connect naturally with the previous paragraph, and smoothly incorporate the current outline and sub-outline content. The result must consist only of sentences, without any additional explanation. [Previous Paragraph]: <Previous Paragraph> [Current Outline]: <Current Outline> [Sub-outline]: <Sub-outline Sentence>

Table 14: Instruction template for Sub-outline-to-Story transformation