SSP: Story-Space Prompting Improves the Reader Immersion in Long Story Generation

Anonymous ACL submission

Abstract

Generating long-form stories with neural network models, even the large language models (LLMs), e.g., GPT, has always been criticized for lacking interestingness and coherence, thus greatly diminishing the reader's sense of immersion. In this paper, we present a novel "story space" prompting (SSP) solution, which provides a coherent and consistent background to support long-term storytelling. Specifically, we first define the story space intricately connected to the given story premise. Then, Our framework systematically generates the story space by progressively constructing it from an abstract representation to a more informative and detailed one. Empirically, we implement our plug-in method upon an existing advanced story generation framework (Yang et al., 2023) and evaluate its impact on both interestingness and coherence. Our findings emphasize the significance of our SSP in enhancing reader enjoyment and immersion, contributing to advancements in long-form story generation.

1 Introduction

Advancements in natural language generation systems have sparked a growing interest in long-form text generation, where texts extend over thousands of words or more (Cho et al., 2019; Tan et al., 2021; Guo et al., 2022). Crafting long stories is a formidable task as it must ensure a coherent and consistent plot that extends over thousands of words, staying true to the initial premise. Additionally, it must maintain the narrative style throughout the entire story and prevent any factual contradictions, which become more challenging to manage over an extensive narrative horizon.

Recent efforts, notably the Re^3 method (Yang et al., 2022b), have pioneered the generation of extended narratives by iteratively refining and expanding stories through a recursive process. Subsequently, DOC (Yang et al., 2023) enhances the outlining phase within the Re^3 framework. These approaches offer a potential solution to improve complex long-form content. Despite advances, current state-of-the-art story generation methods, e.g., DOC, still underperform human capabilities in several aspects, notably in creating ***immersion***: the AI-generated stories often depict protagonists taking actions in a seemingly empty world, lacking the immersive elements that would fully engage the reader within the story world. An immersive story is always accompanied by a rich background (Gander, 1999), which is crucial for story development, influencing its tone and overall effective (Rozelle, 2005). Normally, rich backgrounds are built based on interesting and multi-dimensional story spaces. For example, J.R.R. Tolkien particularly constructed the captivating world of Middle-earth in The Lord of the Rings.

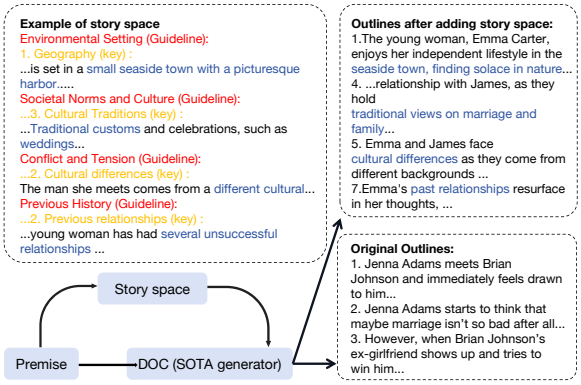

Figure 1: A case study of concatenating **SSP** with **DOC** system. The red words are the guidelines, The yellow words are the keys, and the black words are the value, respectively, in our story space

Motivated by above observations, we propose to improve the reader immersion of the AI-generated long story by introducing a plug-and-play "story space prompting (SSP)" mechanism upon the SOTA story generation framework (Yang et al., 2023). The philosophy of our proposed mechanism is to enhance the structure and continuity of the

story background conditioned on a given premise. To achieve this, we ❶ manually design the guidelines that define abstract dimensions spanning the story spaces (e.g., "*Environmental Setting*" in Figure 1), making them applicable to various story genres; ❷ automatically generate keywords that are highly correlated with the given premise (e.g., "*Geography*" in Figure 1); ❸ carefully develop an agent capable of real-time retrieval from extensive database to supplement the corresponding knowledge value for each keyword, and reinforcing the connection between keywords and premise (e.g., "*a small seaside town with a picturesque harbor and sandy beaches...*" in Figure 1).

To assess the effectiveness of our strategy, we integrate our SSP with the DOC during the generation of the story outline. By combining our SSP with DOC, we evaluate the impact of the story space on the overall quality and coherence of the generated narratives. Experimental results show that SSP achieves substantially higher plot coherence (17.9% absolute increase), interestingness (22.1%), and perceived immersion (22.8%) when judged by human annotators in pairwise evaluations (see Section 5).

In short, our main **contributions** are as follows:

- We find that the outline significantly influences story generation, particularly because LLMs primarily expand sentences. Therefore, crafting a detailed and informative outline to guide the LLM's generation process is crucial.
- Motivated by our finding, we define the story space serving as the story background in the task of long-text story generation, and we proposed our SSP framework to automatically generate story space according to the given premise.
- We implement our framework by concatenating with the previous long-text story generator. The experiment results show that our method significantly improves interestingness, coherence, and therefore immersion.

2 Related Work

Long Text Generation Long-form generation can be implemented in hierarchical structure (Yang et al., 2016; Miculicich et al., 2018; Guo et al., 2021). Fan et al. (2018) also utilize a hierarchical planning scheme for generating stories, starting with generating a premise, which is then expanded into a complete story. Other works similarly employ brief outlines or structured schemas to guide the story generation process (Yao et al., 2019; Goldfarb-Tarrant et al., 2020; Rashkin et al., 2020; Tian and Peng, 2022). Recently, Yang et al. (2022a) use a premise as the starting point and construct a much more detailed plan containing a setting, characters, and brief outline. DOC (Yang et al., 2023)'s detailed outliner constructs a natural language outline, which can easily be adjusted to increase the level of detail to match the desired scope of the final story.

Story Generation With External Knowledge There are two categories in story generation with external knowledge. The first category is transforming structured knowledge data like knowledge triples into natural language texts. Guan et al. (2020) targeted on generating a commonsense story given a starting sentence. Xu et al. (2020) proposed MEGATRON-CNTRL(Speer et al., 2018), which follows that transforming knowledge triples in Concept-Net into respective sentences by a template and selecting knowledge-conditioned sentences generated by GPT-2(Radford et al., 2019) into the story.The second category is building a knowledge reasoner to guide the generator. C2PO (Ammanabrolu et al., 2020) utilizes COMET (Bosselut et al., 2019) to generate intermediate events between any two consecutive plot points that infill the plot lines of a story are extracted. CAST (Peng et al., 2022) aims at generating character-centered stories where each sentence should contain one or two characters. It utilizes COMET as an inferential instructor to supervise the story generation process.

3 Immersion

Immersion in a story refers to the experience of being deeply engaged or absorbed in the narrative, characters, and world of the story. This level of engagement makes the reader or audience feel as though they are a part of the story themselves, often losing awareness of their actual surroundings. Effective immersion typically results from a combination of well-developed characters, a compelling plot, vivid descriptions, and a consistent, believable world (Weiland, 2011). Immersion is a important topic to investigate when we utilize the automatic story generation.

To explore this gap between human written and generated text. We did the experiments as follow:

1. We choose novel *Oliver Twist* by Charles Dickens as our ground true text.

2. We divided the book into 20 segments, each ranging from 700 to 1000 words. These segments were then summarized either into a single sentence or a 100-word paragraph using the LLM.

3. To ensure unbiased text generation, we replaced the original character names with randomly generated names. This step aims to prevent the language model from generating text specifically related to *Oliver Twist*.

4. Afterwards, Subsequently, these anonymized summaries were fed back into the model to generate new segments of the same length.

To evaluate the text pairs, we engaged human annotators to compare them against a provided standard. Simultaneously, we used this standard as a prompt to instruct the LLMs (GPT4 (OpenAI, 2023), Claude2) to perform comparative analysis. The results of these evaluations are presented in Table 1.

Immersion score	**Human**	**GPT4**	**Claude2**
Original	71.4	69.2	63.8
Generation-p	28.1	25.3	35.1
Original	94.7	86.5	91.4
Generation-s	7.8	14.3	9.0

Table 1: Pair-wise comparison of 20 segments of *Oliver Twist* with LLM generations. There is still a gap between these texts. 'Original' stands for *Oliver Twist*. 'Generation' stands for LLM generated text. 'p' stands for text generated from 100 words paragraph. 's' stands for text generated from one sentence.

Based on our findings, the texts generated using 100-word paragraphs exhibited a greater immersion score compared to the human-written texts, which serve as our ground truth. However, as shown in the Figure 2, the generated text tends to expand each sentence from the premise sequentially, whereas *Oliver Twist* uses longer, more descriptive passages that flow together to create a narrative. The generated text focuses more on expanding each sentence rather than weaving them into a cohesive narrative that engages the reader. We could find the omitted part is almost equal in the figure. The original novel provides richer details and paints a more vivid atmosphere, invoking emotions and sensory experiences through detailed descriptions of characters (e.g., the nurse with the green glass bottle).

Referring back to Table 1, it becomes evident that using only one sentence as the premise leads to unsatisfactory generated results. No matter whether a model or human annotator could easily recognize the difference between a masterpiece and generated text. To explore this further, we varied the length of input premises from 1 to 5 sentences and analyzed their impact on the immersion score of the generated texts. Due to the high Pearson coefficient, we only use model evaluation. The findings (Figure 3) indicate a positive correlation between input length and immersion score, suggesting that longer inputs provide more context and specificity for the language model to build upon. The immersion score remains stable after 4 sentences, which could indicate a potential plateau or saturation effect. We could briefly conclude that outline significantly influences story generation, particularly because LLMs primarily expand sentences. Therefore, crafting a detailed outline to guide the LLM's generation process is crucial.

4 Methodology

Based on our conclusions, we find that current state-of-the-art (SOTA) generators, such as DOC (Yang et al., 2023), still suffer from oversimplified outlines. We propose introducing a plug-and-play SSP mechanism for these SOTA frameworks to enhance reader immersion in model-generated long stories.

4.1 Setup

In this task, a given or generated premise p is necessary; it's normally a short paragraph of 30-50 words provided by the users. The premise is used to clarify the theme and main plot. Our ultimate goal is to build a story space $\mathcal{S}$ highly correlated to the premise p. This story space based on p is spanned by a set of independent guidelines g for one direction, like "temporal setting" and "societal norms etc. For each guideline, it's designed as a vector $(k_1 : d_1, k_2 : d_2, \ldots)$ consisting of some key-value pairs, where the key k_i stands for a specific implementation of the premise in this direction, and the value d_i is corresponding detailed descriptions. During the SSP generation, we use ChatGPT API (Ouyang et al., 2022; OpenAI, 2023). Figure 4 shows the example of the story space on the mentioned premise.

Premise: In a town's workhouse, Fletcher Burnett is born in poor health and struggles to breathe. With only a drunken pauper woman and a parish surgeon present, he survives naturally. His mother, weak and dying, asks to see him. After kissing him, she dies despite the surgeon's and nurse's attempts to revive her.

Generated Text:
The town's workhouse stood as a grim reminder of society's most destitute. ... [42 words] ... It was within this foreboding environment that Fletcher Burnett was born. ...[30 words] ... The midwife, a hunched and weathered pauper woman named Mary, had more whiskey than wisdom coursing through her veins. ... [54 words]... Beside her, the parish surgeon, a man of waning patience and thinning hair, hovered with a look of reluctant duty. ... [26 words] ... Fletcher's tiny chest heaved with difficulty, each breath a monumental effort. His mother, Eleanor Burnett, lay on the bed, her face pale and gaunt from the toll of childbirth and years of deprivation. She turned her head weakly toward the bundle of life that had just been placed in her arms. Her eyes, though clouded with exhaustion, shone with a fierce, fleeting light of maternal love.

Oliver Twist:
[30 words] ... there is one anciently common to most towns, great or small: to wit, a workhouse; ... [54 words] ... For a long time after it was ushered into this world of sorrow and trouble, by the parish surgeon, it remained a matter of considerable doubt whether the child would survive to bear any name at all; ... [68 words] ... There being nobody by, however, but a pauper old woman, who was rendered rather misty by an unwonted allowance of beer; and a parish surgeon who did such matters by contract; Oliver and Nature fought out the point between them. The result was, that, after a few struggles, Oliver breathed, sneezed, and proceeded to advertise to the inmates of the workhouse the fact of a new burden having been imposed upon the parish, ... [96 words] ... As Oliver gave this first proof of the free and proper action of his lungs, the patchwork coverlet which was carelessly flung over the iron bedstead, rustled; the pale face of a young woman was raised feebly from the pillow; and a faint voice imperfectly articulated the words, 'Let me see the child, and die.' ...[56 words]... 'Lor bless her dear heart, no!' interposed the nurse, hastily depositing in her pocket a green glass bottle, the contents of which she had been tasting in a corner with evident satisfaction. ... [46 words] ... The surgeon deposited it in her arms. She imprinted her cold white lips passionately on its forehead; passed her hands over her face; gazed wildly round; shuddered; fell back—and died.

Figure 2: An example of comparison of generated text and *Oliver Twist*

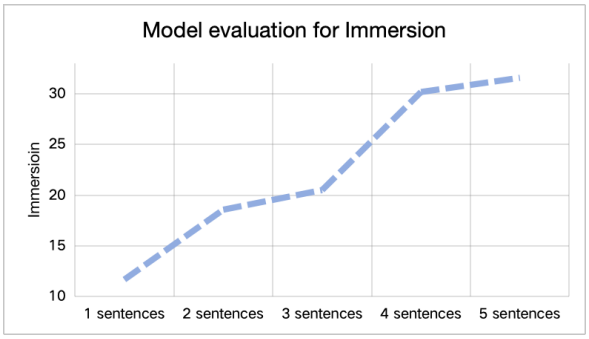

Figure 3: Model evaluation for immersion. When input outlines are more detailed, the higher immersion score the generated texts get.

4.2 Frame work

To make the corpus closer to the premise, we choose to hierarchically generate this corpus:

Story Genres

At the initial stage of the framework, the decision of the genre plays a crucial role as it significantly impacts the development of the story's background. For instance, as depicted in the upper left quadrant of figure 4, the ultimate choice made by the female character is influenced by the chosen story genre. In the case of a "Romance story" genre, she may decide to marry the man, whereas in a "Realistic fiction story" genre, her decision may be influenced by the discovery of certain secrets about the man. We constructed these 10 story types through the Wikipedia fiction categories. The details can be found in Appendix A.

Guidelines

We follow and modify the setup in the book *Outlining Your Novel: Map Your Way to Success* (Weiland, 2011). The idea proposed in that book is to list some indispensable novel points for creators to fill them. We organize and adjust these points as the ground of our guidelines.

The guidelines devised in this framework are characterized by their high level of abstraction, allowing them to be applicable across various story genres. They are intentionally designed to be inde-

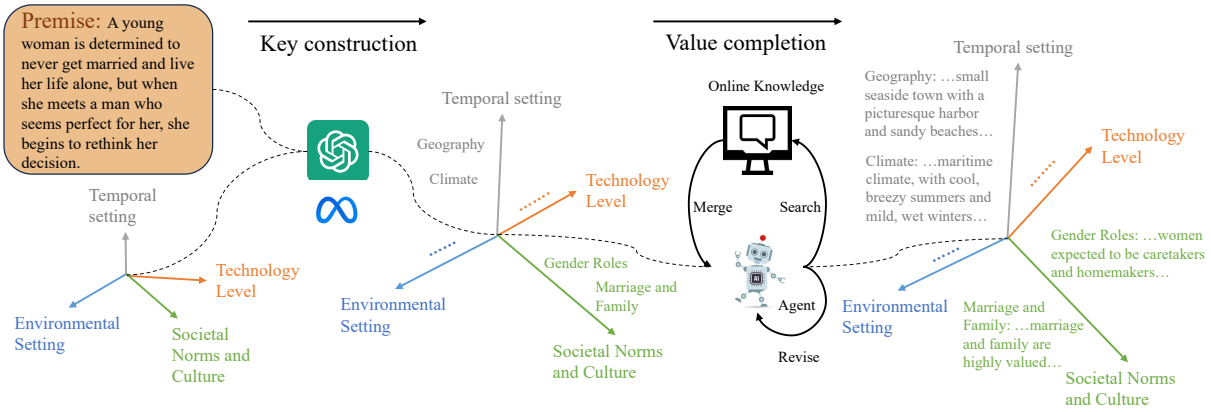

Figure 4: Flow chart of our proposed SSP.

pendent to avoid repetition of key elements. Each guideline serves as a guiding principle for LLM to generate premise-related keywords that align with its specific direction. The guidelines have been meticulously crafted and designed by humans. Eventually, there are a total of 7 guidelines to provide meaningful dimensions for the construction of the story space. The details can be found in Appendix B. In figure 4, we only demonstrate 4 directions for spanning the space of the story world.

Key construction

The keys in each guideline should be related to the premise. We generate keywords as follows:

$$(k_1, k_2, \ldots)_{pg} = \text{LLM}(p, \text{prompt}_g)$$

where prompt_g is a prompt designed in a guideline g. (The detailed prompts can be find in Appendix C) To mimic the brainstorming, we sample n times on the key generated distribution, where n is a hyperparameter, and make a selection to ensure the desired results. The selection methods are as follows: Inspired by the method proposed by Weng et al. (2023), which concludes that LLMs have a strong ability to self-verify, allowing them to accurately assess the conclusions they generate. Our selection method simply involves putting all of the generated keys into the LLM and getting the selected keys.

Value completion

After constructing our desired key on a guideline, we need a detailed description corresponding to each key. For instance, in Figure 1, after we generate a key "Geography" in the guideline "Environmental Setting", we need a detailed description, "a small seaside town with a picturesque harbor and sandy beaches", to explain the correlation between the key and the premise. The requirement for this description should also be informative and based on real-world knowledge. We extend the generations in Part 4.2 with value initialization.

$$(k_1 : v_1, k_2 : v_2, \ldots)_{pg} = \text{LLM}(p, \text{prompt}_g)$$

Real-time retrieval

We adopt a similar approach to the paradigm introduced by Yao et al. (2023) known as ReAct, which combines the ability of LLM at reasoning and acting. In our case, we develop an agent specifically designed for value completion. The three available choices are as follows: (1) Search: If we choose the "Search" action, we will conduct a search using an entity generated by LLM (Entity $=$ LLM(prompt, v_i, k_i, p), where i stands for i-th key-value pair) to further improve the existing knowledge. Differing from ReAct, which relies on searching for entities on Wikipedia and reasoning based on the search results, our agent takes a different approach. It retrieves knowledge from both the commercial system Bing Chat and Wikipedia. (2) Revise: Alternatively, we can choose the "Revise" instead of the original reasoning part, which involves making revisions to the existing knowledge in order to make it more closely correlated to the premise inspired by the discovery of self-improve ability proposed by Fu et al. (2023). (3) Finish: "Finish" action indicates that LLM believing that the current knowledge is sufficient and that no further modifications are required.

The 'Search' action ensures a wide range of up-to-date and comprehensive information. For example, the original outline is *Shannon lands her first major assignment, a feature on the inner city.* After retrieving results from the 'Search' action,

the entity 'New York' is added to our final outline: *Shannon lands her first major assignment, a feature on the inner city, set against the bustling, diverse backdrop of New York City, where vibrant life and stark socioeconomic disparities coexist.*

Correspondingly, we found that sometimes the entity derived by LLM is not relevant to our premise. For instance, the key is *Dimly lit alleyways where shady deals and illicit activities take place*, but the entity from the 'Search' action is 'The Silk Road', which leads LLM to generate a new but wrong key. Nevertheless, the 'Revise' action helps LLM to re-generate a premise-correlated key.

4.3 Plug-and-play

Upon obtaining a complete SSP (details available in Appendix E), we instruct the LLM to generate an outline individually based on this SSP. This outline can then be flexibly adapted to various story-generating frameworks, such as Re^3(Yang et al., 2022b) or DOC(Yang et al., 2023), according to the specific format required by each framework. As demonstrated in Table 2, we generated an outline following the DOC (Yang et al., 2023) format (an example prompt is provided in Appendix C). Once the outline is prepared, it is fed into the framework to generate the full story.

GENERATED OUTLINE:
1. Shannon, driven by the desire to honor her late father's legacy, lands her first major assignment as a journalist in the urban landscape of New York City.
a. Shannon starts her job at a prestigious local newspaper, assigned to cover human interest stories in Queens, ...
b. During her initial days on the job, Shannon struggles with imposter syndrome and doubts about her abilities, ...
2. While navigating the bustling streets and diverse cultures of Queens, Shannon encounters Marcus, a street-wise teenager with a deep understanding of the city's hidden truths.
a. Shannon meets Marcus while researching a story on local community efforts to improve public housing conditions. Marcus, ...
b. Marcus introduces Shannon to Jackson Heights, Corona, and other vibrant neighborhoods where immigrant communities thrive, ...
As Shannon delves deeper into her investigative work, she uncovers systemic inequities and injustices plaguing the inner city, including issues of housing discrimination, inadequate access to education and healthcare, and instances of police brutality, challenging her preconceptions about the city and her role as a journalist.
a. Shannon investigates a series of evictions in a predominantly immigrant neighborhood,
b. While researching disparities in healthcare access, Shannon interviews families in underserved neighborhoods who share heartbreaking stories of neglect and discrimination, v...

Table 2: An example of generated outline.

5 Evaluation

The generation of stories and the components within our framework cannot be easily defined by rule-based constraints. Instead of imposing rigid rules on acceptable outputs, we adopt a more flexible approach and evaluate the generated content using human-annotated metrics.

5.1 Entity density

As shown in Section 3, we have observed that the informativeness of the outline can significantly impact the generation results. To evaluate the informativeness of an outline, we introduce a metric called entity density. Entity density is a measure that quantifies the concentration of entities within a given segment of text. A higher entity density indicates greater content richness and relevance, especially in certain types of documents. This metric helps us understand the richness or sparsity of entities present in our outline. To apply this measure, we defined the types of entities, such as "person," "organization," and "location." We then use llm-NER (Villena et al., 2024) to identify and count the total number of these entities within the given text segment. Then we use the following formula to calculate the entity density:

$$\text{Entity Density} = \frac{\text{Number of Entities}}{\text{Length of Text Segment}}$$

The result is shown in Table 3. The SSP method shows a significant advantage over the baseline method in terms of entity density, which is a key metric for assessing content informativeness. Specifically, with the average number of entities in each outline point of 7.32 compared to the baseline's 2.84, SSP is nearly three times as many in terms of entities per text segment. Additionally, the entity density identified by SSP is higher (79.2 vs. 63.4), further reinforcing the conclusion that SSP is more effective at representing the content's richness. This demonstrates that SSP is more effective at enhancing the informativeness and richness of the generated content, making it a better tool for content generation tasks.

Method	#of Entity	Entity density
Base	2.84	63.4
SSP	**7.32**	**79.2**

Table 3: Pairwise comparisons of SSP against baselines (DOC generated outlines) from 50 outlines. Bold indicates significance with $p < 0.05$. '#of Entity' is the average number of entities in 50 outlines. Entity density is measured in %.

5.2 Agent

Due to online knowledge retrieval, the value generated with the agent has more entity information and it is revised by LLM to have a closer distance from the premise. As shown in Table 4, annotators think that the value with the agent has more informativeness. We find that LLM may provide some entities that do not exist so the Wikipedia wrapper gets some irrelative information. This may affect the LLM to self-improve so that the final value does not have a satisfactory result.

Method	Infomativeness
Initial	17.12
Initial w. Agent	**87.82**

Table 4: Pairwise comparisons of Agent refining against baselines (initial value) from 50 outlines. Bold indicates significance with $p < 0.05$. Refined values are rated with more information.

5.3 Final story Setup

For the main results comparison, our setup is similar to DOC. The input is an English premise and the output is a corresponding story generated automatically. We adopt 20 diverse premises in DOC as a testset. To decrease noise, we compare 2000-2500 word stories corresponding to the same premise. We use three main metrics, also similar to those from DOC.

Specifically, we give two novels (A and B, with random order) generated by different compared methods to human annotators with good English proficiency and instruct them to label whether novel A or novel B is better, or they are indistinguishable, in terms of interestingness, coherence and Immersion. Following the human evaluation settings in DOC, we sample 20 generated novels for each genre and assign 3 annotators for each novel.

GENERATED STORY:

Emma Carter had always cherished her independent lifestyle in the serene seaside town of Willowbrook. With the sound of crashing waves as her symphony and the salty breeze as her constant companion, ...[954 words]...

Emma's unwavering commitment to her independence drew concern from her family, particularly her parents. They held staunchly traditional views on marriage and family, ...[104 words]...

One evening, as they strolled along the shoreline, the topic of marriage emerged. Emma expressed her reservations about the institution, sharing her belief that it could...[419 words]...

In the stillness of her seaside home, Emma would sift through memories like grains of sand slipping through her fingers. She recalled the times she had opened her heart, only to have it shattered by betrayal and disappointment. ...[195 words]...

Table 5: An example of generated stories by the given premise.

- *Interesting* The percentage of passages judged interesting.
- *Coherent* The percentage of passages judged plot-coherent by human annotators.
- *Immersion* The percentage of passages judged as having an immersion by human annotators.

Baselines

Our primary baseline for story generation is based on the DOC system proposed by Yang et al. (2023), which is known to automatically generate stories of similar length. DOC is an extension of the Re^3 system, but it includes an outline controller that allows for longer outlines to be generated. In our analysis, we explore different ways of incorporating our story space into the DOC system.

Method	Interestingness	Coherence	Immersion
DOC	40.3	45.3	34.7
DOC w. SSP	**62.4**	**63.2**	**57.5**

Table 6: Pairwise comparisons between DOC with SSP and the baseline DOC were conducted for 20 stories. Bold indicates significance with $p < 0.05$. The stories generated using SSP are rated substantially more interesting, coherence and immersive compared to the baseline.

Results

The results in Table 6 confirm that our framework does substantially improve the ability of DOC to

make the story more interesting and under the control of our space, the plot is more coherent. The generated story in Table 5 serves as a concrete illustration of how the outlined guidelines are effectively implemented. The story demonstrates a clear alignment with our story space, indicating a degree of control in the narrative. Although our framework concatenating DOC achieves a better immersion for the reader, the overall score of immersion remains low compared to human texts, pointing to two issues. First, the current concatenating method could not utilize all the information in the story space. Second, our story space may not be sufficient due to the entity missing problems (in section 5.2) encountered. Therefore, we incorporate the human in these steps as illustrated in section 6.

6 Human-Interactive Story Generation

Furthermore, we evaluate SSP in an interactive setting with a focus on human controllability. As illustrated in Table 2, after generating an outline, users can refine it through manual 'Search' and 'Revise' actions. During the 'Search' action, users can find inspiration from the output entity and substitute it with a more appropriate one. During the 'Revise' action, users can modify both the key and its corresponding value. This iterative process ensures that the generated outline remains flexible and adaptable for user customization.

We asked annotators to compare the following metrics specific to the interactive experience.

1. *Informativeness* same as Section 5.2, we follow the evaluation instruction.

2. *Immersion* same as Section 3, we follow the evaluation instruction.

3. *Coherence* evaluate the coherence of story plot.

4. *Quality.* overall quality of final story outline.

Method	Inform.	Immersion	Coherence	Quality
SSP	61.5	58.4	64.8	13.5
Human	84.3	59.3	89.4	90.0

Table 7: Pairwise comparison of SSP vs. human interaction on 20 story outlines. Numbers indicate the percentage of responses in favor of each outline, with "no preference" responses omitted. SSP is preferred by a wide margin on overall quality and coherence.

Results. Human interaction is preferred by a significant margin in terms of overall quality and coherence, with percentages of 90.0% and 89.4%, respectively. In terms of Informativeness, human interaction scores slightly higher than the human interaction with a score of 84.3% compared to 61.5%. The entity missing problem is alleviated by the human assistant. For immersion, human interaction also has a higher score than human interaction, with 59.3% compared to 58.4%. The high scores for coherence and quality suggest that human interaction produces story outlines that are well-structured and of a high standard. While human interaction leads in immersion, the margin is not as wide as in coherence and quality, indicating that the immersion of generated text still lags behind human writing. Despite human assistance, current methods still do not achieve the level of human work.

7 Conclusion

Our approach involved integrating the SSP framework with an existing long-text story generator, allowing for seamless synergy between the structured story space and the narrative construction process. This integration ensures that the generated stories are not only coherent but also deeply immersive.

Experimental results validate the effectiveness of our method. By employing the SSP framework, we observed significant enhancements in the generated stories across several dimensions. Notably, our method improved the interestingness of the narratives, making them more captivating and engaging for readers. Additionally, the coherence of the stories was markedly better, as the structured story space provided a consistent context that guided the narrative flow. Lastly, the overall immersion of the stories was significantly heightened, enabling readers to become more deeply absorbed in the fictional worlds created.

In summary, the SSP framework represents a significant advancement in the field of automated story generation. By focusing on the foundational element of story space, we have developed a method that not only enhances narrative coherence and engagement but also enriches the overall reading experience. Future work could explore further refinements of the SSP framework and its application to various genres and styles of storytelling, potentially opening new avenues for creative and compelling automated narratives.

Limitations

Indeed, evaluating the quality of story outputs in long-form text generation is a challenging task that often requires human annotations. While human annotations can be expensive and time-consuming, the difficulty of evaluation limits us from running more detailed ablations. Many of our prompts are carefully engineered in English cause two limitations: the process of prompt engineering involves human intervention, which can introduce unintentional bias. Meanwhile, designing carefully engineered prompts can be time-consuming and resource-intensive.

Ethics Statement

We place great importance on ethical considerations and adhere strictly to the ACL Ethics Policy. While we recognize the potential for harm associated with powerful language generation models, our specific focus on generating fiction stories helps mitigate certain ethical concerns compared to broader language models. By emphasizing the generation of fictional content, we aim to minimize the potential for generating toxic or untruthful text that could be used maliciously. we believe that this research will not pose ethical issues.

References

Prithviraj Ammanabrolu, Wesley Cheung, William Broniec, and Mark O. Riedl. 2020. Automated storytelling via causal, commonsense plot ordering. In *AAAI*.

Antoine Bosselut, Hannah Rashkin, Maarten Sap, Chaitanya Malaviya, Asli Celikyilmaz, and Yejin Choi. 2019. Comet: Commonsense transformers for automatic knowledge graph construction. In *ACL*.

Woon Sang Cho, Pengchuan Zhang, Yizhe Zhang, Xiujun Li, Michel Galley, Chris Brockett, Mengdi Wang, and Jianfeng Gao. 2019. Towards coherent and cohesive long-form text generation. In *WNU*.

Angela Fan, Mike Lewis, and Yann Dauphin. 2018. Hierarchical neural story generation. *arXiv preprint arXiv:1805.04833*.

Yao Fu, Hao Peng, Tushar Khot, and Mirella Lapata. 2023. Improving language model negotiation with self-play and in-context learning from ai feedback. *arXiv preprint*.

Pierre Gander. 1999. *Two myths about immersion in new storytelling media*. Lund University.

Seraphina Goldfarb-Tarrant, Tuhin Chakrabarty, Ralph Weischedel, and Nanyun Peng. 2020. Content planning for neural story generation with aristotelian rescoring. *arXiv preprint arXiv:2009.09870*.

Jian Guan, Fei Huang, Zhihao Zhao, Xiaoyan Zhu, and Minlie Huang. 2020. Knowledge-enhanced pretraining model for commonsense story generation. *TACL*.

Mandy Guo, Joshua Ainslie, David Uthus, Santiago Ontanon, Jianmo Ni, Yun-Hsuan Sung, and Yinfei Yang. 2021. Longt5: Efficient text-to-text transformer for long sequences. *arXiv preprint arXiv:2112.07916*.

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 *NAACL*.

Lesly Miculicich, Dhananjay Ram, Nikolaos Pappas, and James Henderson. 2018. Document-level neural machine translation with hierarchical attention networks. *arXiv preprint arXiv:1809.01576*.

OpenAI. 2023. Gpt-4.

Long Ouyang, Jeff Wu, Xu Jiang, Diogo Almeida, Carroll L Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, et al. 2022. Training language models to follow instructions with human feedback. *arXiv preprint arXiv:2203.02155*.

Xiangyu Peng, Siyan Li, Sarah Wiegreffe, and Mark Riedl. 2022. Inferring the reader: Guiding automated story generation with commonsense reasoning. In *Findings of EMNLP*.

Alec Radford, Jeff Wu, Rewon Child, David Luan, Dario Amodei, and Ilya Sutskever. 2019. Language models are unsupervised multitask learners.

Hannah Rashkin, Asli Celikyilmaz, Yejin Choi, and Jianfeng Gao. 2020. Plotmachines: Outline-conditioned generation with dynamic plot state tracking. *arXiv preprint arXiv:2004.14967*.

R. Rozelle. 2005. *Write Great Fiction - Description & Setting*. Write Great Fiction.

Robyn Speer, Joshua Chin, and Catherine Havasi. 2018. Conceptnet 5.5: An open multilingual graph of general knowledge. *arXiv preprint*.

Bowen Tan, Zichao Yang, Maruan Al-Shedivat, Eric Xing, and Zhiting Hu. 2021. Progressive generation of long text with pretrained language models. In *NAACL*.

Yufei Tian and Nanyun Peng. 2022. Zero-shot sonnet generation with discourse-level planning and aesthetics features. In *2022 Annual Conference of the North American Chapter of the Association for Computational Linguistics (NAACL)*.

Fabián Villena, Luis Miranda, and Claudio Aracena. 2024. llmner: (zero|few)-shot named entity recognition, exploiting the power of large language models.

K.M. Weiland. 2011. *Outlining Your Novel: Map Your Way to Success*. Helping Writers Become Authors. PenForASword Pub.

Yixuan Weng, Minjun Zhu, Fei Xia, Bin Li, Shizhu He, Kang Liu, and Jun Zhao. 2023. Large language models are better reasoners with self-verification. In *Findings of EMNLP*.

Peng Xu, Mostofa Patwary, Mohammad Shoeybi, Raul Puri, Pascale Fung, Anima Anandkumar, and Bryan Catanzaro. 2020. MEGATRON-CNTRL: Controllable story generation with external knowledge using large-scale language models. In *EMNLP*.

Kevin Yang, Dan Klein, Nanyun Peng, and Yuandong Tian. 2023. DOC: Improving long story coherence with detailed outline control. In *ACL*.

Kevin Yang, Nanyun Peng, Yuandong Tian, and Dan Klein. 2022a. Re3: Generating longer stories with recursive reprompting and revision. *arXiv preprint arXiv:2210.06774*.

Kevin Yang, Yuandong Tian, Nanyun Peng, and Dan Klein. 2022b. Re3: Generating longer stories with recursive reprompting and revision. In *ENNLP*.

Zichao Yang, Diyi Yang, Chris Dyer, Xiaodong He, Alex Smola, and Eduard Hovy. 2016. Hierarchical attention networks for document classification. In *Proceedings of the 2016 conference of the North American chapter of the association for computational linguistics: human language technologies*, pages 1480–1489.

Lili Yao, Nanyun Peng, Ralph Weischedel, Kevin Knight, Dongyan Zhao, and Rui Yan. 2019. Plan-and-write: Towards better automatic storytelling. In *Proceedings of the AAAI Conference on Artificial Intelligence*, 01, pages 7378–7385.

Shunyu Yao, Jeffrey Zhao, Dian Yu, Nan Du, Izhak Shafran, Karthik Narasimhan, and Yuan Cao. 2023. React: Synergizing reasoning and acting in language models. In *ICLR*.

A Story Types

- Adventure stories
- Romance stories
- Science fiction stories
- Horror stories
- Mystery stories
- Fantasy stories
- Historical fiction stories
- Realistic fiction stories
- Fairy tales
- Weird fiction stories

B Guidelines

Temporal Setting Describes the time period in which the story takes place, including historical, present, or future settings.

Environmental Setting Details the physical environment and location of the story, such as urban, rural, natural landscapes, or fantastical worlds.

Societal Norms and Culture Covers the social structures, customs, beliefs, and values that influence the characters and plot.

Technology Level Indicates the level of technological advancement within the story, which can range from primitive to futuristic.

Political Structure Describes the form of government, power dynamics, and political atmosphere that affect the story's world.

Historic Setting Provides context by referencing real historical events or periods that shape the story's backdrop.

Conflicts and Tensions Highlights the central conflicts, struggles, and sources of tension that drive the story forward.

C Prompts in Methodology

C.1 Prompt in story genres

Instruction: Given [premise], select [story type] from the list: ['Adventure stories', 'Romance stories', 'Science fiction stories', 'Horror stories', 'Mystery stories', 'Fantasy stories', 'Historical fiction stories', 'Realistic fiction stories', 'Fairy tales', 'Weird fiction stories'].
Here is the [premise]:
premise: After the loss of her father, Shannon is determined to follow in his footsteps ...

C.2 Prompt in Key construction

Story Type: Realistic fiction stories
Premise: After the loss of her father, Shannon is determined to follow in his footsteps...
Give a numbered list (not exceed 7) of Environmental Setting, without explanation.
(For different guideline, substitute red part)

C.3 Prompt in Value completion

Story Type: Realistic fiction stories
Premise: After the loss of her father, Shannon is determined to follow in his footsteps...
Environmental Setting: Busy city streets filled with diverse people and bustling activity.
According to the {premise}, provide each {environmental setting} with some facts.
(For different guideline, substitute red part; for different keyword in guideline, substitute yellow part)

C.4 Prompt in Agent

Instruction: Imagine you are a story writer, now you collect the background knowledge for the story. The background should be informative and precise. Do not need any creation.
{} stands for the given context. [] stands for the generated texts.
Given the {premise} of the story, a {keyword} for building the background of the story, and the {knowledge} is corresponding to the {keyword}. Reason these three contexts [reason], and choose one of [action]s.
There are three [action]s for you to choose:
1.Search, which returns a [new keyword] to search for improving the {knowledge}. (Template: Action: Search New keyword: [new keyword])
2.Revise, which makes a revision on the {knowledge} so that {knowledge} is more corelated to the {premise} and returns [new knowledge] (Template: Action: Revise New knowledge: [new knowledge])
3.Finish, which ends the task when you think the {knowledge} is good enough. (Template: Action: Finish)
Here are {premise}+{keyword}+{knowledge}:
Premise: After the loss of her father, Shannon is determined to follow in his footsteps...
Keyword: Busy city streets filled with diverse people and bustling activity
Knowledge: The bustling city streets of New York, with its diverse population and vibrant activity, serve as ...

C.5 Prompt in plug and play

First point:
Generated SSP (content is omitted)
Let do it step by step, use these settings (Environmental Setting, Societal Norms and Culture, Temporal setting, Technology Level, Political Structure, Historical setting, Conflicts and tensors) to write a story outline for the premise.
Give me the first point (one sentence) of outline:

Following points:
Generated SSP (content is omitted)
Generated Outlines (content is omitted)
Give me the next point; use the information in these settings (Environmental Setting, Societal Norms and Culture, Temporal setting, Technology Level, Political Structure, Historical setting, Conflicts, and tensors).

D Human annotation example

In the questions, annotators are required to choose one option. However, "no preference" and "both" are also available as options.
An example of human annotation is show in Figure 5

E Example of SSP

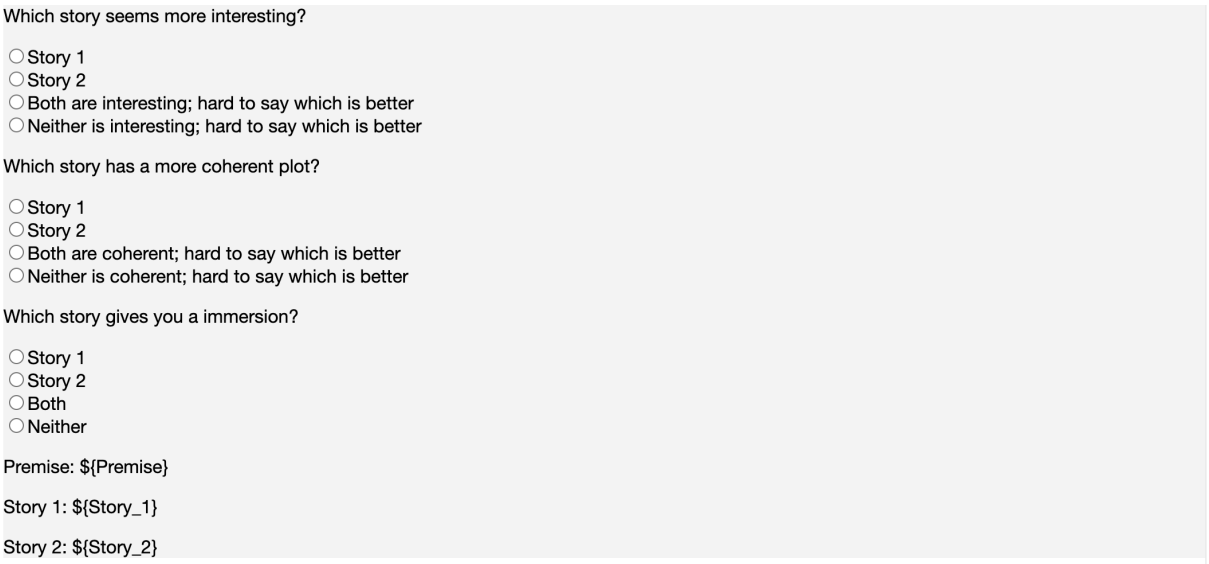

Figure 5: An example of human annotation UI.

SSP:

Premise: After the loss of her father, Shannon is determined to follow in his footsteps and become a successful journalist. However, when she lands her first major assignment, she quickly discovers that the ugly reality of life in the city is far different from the dream she imagined. With the help of her new friend, a street-wise teenager, Shannon comes to understand the harsh realities of life in the inner city and learns that sometimes the truth is much more than just a story.

Story type: Realistic fiction story

Temporal setting:The story is set in contemporary times

Environmental Setting:
Urban Landscape of New York City: The story unfolds amidst the iconic skyline of New York City, where towering skyscrapers scrape the sky, and the streets pulse with the rhythm of millions of inhabitants. From the dazzling lights of Times Square to the historic brownstones of Brooklyn, the cityscape is a patchwork of architectural marvels and cultural landmarks.
Cultural Melting Pot: Queens serves as a vibrant melting pot of cultures and ethnicities, with immigrant communities from around the world calling it home. Bustling ethnic enclaves like Jackson Heights and Corona pulsate with the sights, sounds, and flavors of diverse cultures, offering a glimpse into the rich tapestry of human experience.
Cultural Landmarks: Queens is dotted with cultural landmarks and iconic destinations that add depth to its urban landscape. From the historic Queens Museum and Flushing Meadows-Corona Park to the vibrant culinary scene of Queens Night Market, the borough is a treasure trove of cultural experiences waiting to be explored.
Transit Hub: As a key transportation hub, Queens buzzes with activity day and night. The borough is crisscrossed by an intricate network of subways, buses, and commuter trains, connecting its diverse neighborhoods to the rest of the city. From the bustling terminals of JFK and LaGuardia airports to the bustling transit hubs of Jamaica and Long Island City, Queens is a gateway to the world beyond its borders.

Societal Norms and Culture:
Inequities and Injustices: As Shannon delves deeper into her investigative work, she uncovers the systemic inequities and injustices that plague the inner city. Issues such as housing discrimination, lack of access to quality education and healthcare, and police brutality become stark realities that challenge her preconceptions about the city she thought she knew. She realizes that the truth goes beyond surface-level stories and confronts the harsh realities of structural inequality.
Activism and Advocacy: Inspired by the resilience of the community and driven by a desire for justice, Shannon becomes involved in local activism and advocacy efforts. She joins grassroots organizations fighting for affordable housing, racial justice, and community empowerment, using her platform as a journalist to amplify the voices of those whose stories are often overlooked or ignored. Through her involvement, she learns that the truth is not just a story but a catalyst for meaningful change.

Technology Level:
Social Media Platforms: Platforms like Facebook, Instagram, and Twitter are popular among Queens residents for staying connected with friends, family, and the community. Shannon, the protagonist, uses social media to share updates on her investigative work and engage with her audience.
Instant Messaging Apps: Apps such as WhatsApp, Telegram, and Signal are commonly used for instant messaging and group chats. Shannon communicates with sources, colleagues, and her new friend from the inner city through these platforms, exchanging information and coordinating meetings.
Video Conferencing: Video conferencing tools like Zoom and Skype are essential for remote communication and virtual meetings. Shannon conducts interviews with sources and collaborates with her editor and fellow journalists using these platforms, allowing for face-to-face interactions despite physical distance.
Email and Cloud Services: Email remains a primary mode of communication for professional correspondence. Shannon uses email to pitch story ideas, coordinate interviews, and communicate with sources and editors. Cloud storage services like Google Drive and Dropbox enable seamless sharing and collaboration on documents and media files.
Cryptocurrency Laundering: In the underbelly of Queens, cryptocurrency transactions are often used as a tool for money laundering and illicit activities. Shannon discovers shadowy online forums and dark web marketplaces where criminals exchange cryptocurrencies to obfuscate the origins of their funds. She delves into stories of cybercrime syndicates and underground networks exploiting the anonymity of cryptocurrencies like Bitcoin to launder money obtained through illegal activities such as drug trafficking, extortion, and hacking.

Political Structure:
Government Accountability: Shannon is dedicated to uncovering issues of government accountability and transparency within Queens. She meticulously investigates allegations of corruption, conflicts of interest, and abuse of power, holding elected officials and public institutions accountable for their actions. Through her reporting, she advocates for greater transparency and integrity in government operations.
Party Politics: Queens boasts a politically diverse landscape. While the Democratic Party holds sway over most public offices, the borough is considered a swing county in New York politics, with a significant Republican presence. Sixty-three percent of registered Queens voters affiliate with the Democratic Party, with local party platforms focusing on issues such as affordable housing, education, and economic development. However, contentious political matters, including development policies, noise regulation, and housing affordability, underscore the diversity of viewpoints within the borough.

Historic setting:
Single Parent Household and the Introduction of a Step-Mother: Following her mother's departure early in her life, Shannon's father eventually remarried, introducing a step-mother into Shannon's life. While the transition was not without its challenges, Shannon's step-mother brought a new dynamic to their family structure.
Father-Daughter Bond: Shannon shared a close bond with her father, who not only served as a loving parent but also as a mentor and confidant. He was her rock, guiding her through life's challenges and nurturing her passion for journalism from a young age.

Conflicts and tensors:
Conflict with Truth vs. Ethics: Shannon grapples with the ethical dilemma of uncovering uncomfortable truths in her journalistic pursuits. She must navigate the fine line between reporting the truth and respecting the privacy and dignity of those involved, leading to internal conflict over her professional responsibilities.
Conflict in Relationships: Shannon experiences conflict in her relationships, both with her newfound friend Marcus and with her family members. Differing perspectives, misunderstandings, and emotional barriers create tension that must be addressed and resolved as Shannon navigates her personal and professional journey.
Conflict with Expectations vs. Reality: Shannon experiences conflict as she confronts the stark disparity between her idealized expectations of journalism and the harsh realities she encounters in her first major assignment. The disillusionment she feels challenges her perception of the profession and forces her to reevaluate her goals.
Conflict with Self-Identity: Shannon grapples with conflict surrounding her self-identity and sense of purpose. As she embarks on her journey to honor her father's legacy, she must confront insecurities, doubts, and fears about her abilities and whether she is truly capable of fulfilling her dreams.

Table 8: One detailed example of story spaces generated by SSP