SNAP: Enhancing Long-Form Narrative Agents with Cell-Based Segmentation and Plan-Driven Dialogue Strategies

Anonymous ACL submission

Abstract

Large Language Models (LLMs) hold great potential across domains like gaming, education, and cultural content. However, ensuring character consistency and narrative coherence in extended storytelling remains challenging. We propose SNAP (Story and Narrative-based Agent with Planning), a framework that segments narratives into Cells, defines explicit Plans for each Cell, and guides dialogue generation accordingly. By limiting context within each Cell and providing plans with clear spatiotemporal settings, character actions, and plot developments, SNAP enables consistent, goal-driven dialogues. Experiments with novel datasets and human evaluations show that SNAP outperforms Vanilla GPT-4o-based agents in linearity, continuity, appropriateness, and non-redundancy, demonstrating its effectiveness in creating immersive long-form conversational agents.

1 Introduction

Recent advances in Large Language Models (LLMs) have enabled their widespread application across domains such as cultural content creation (Simon and Muise, 2022; Wang et al., 2023; Yuan et al., 2022; Wang et al., 2022), gaming (Sun et al., 2023b), healthcare (Qiu and Lan, 2024; Kim et al., 2024), and education (Liu et al., 2024). In particular, research on implementing interactive dialogue agents with specific character personas has gained traction (Shao et al., 2023; Ahn et al., 2024; Klinkert et al., 2024; Argyle et al., 2023), demonstrating their ability to provide highly engaging user experiences by simulating human-like interactions and behaviors (Park et al., 2023; Peng et al., 2024; Kumaran et al., 2023; Wan et al., 2024). For instance, when virtual characters maintain consistent personalities and behaviors within a narrative universe, users can experience interactions akin to conversing with real individuals.

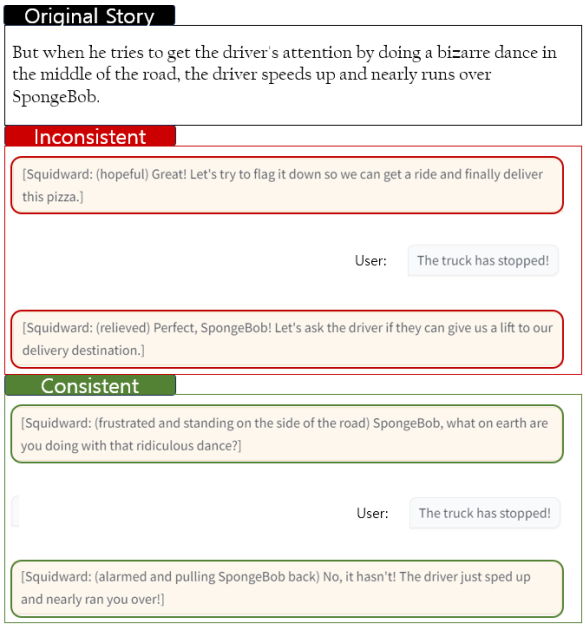

Figure 1: Examples of dialogues generated based on the original story. (Top) Squidward deviates from the original storyline due to the user's input. (Bottom) Squidward remains aligned with the original storyline despite the user's attempt to alter it.

However, LLM-based agents often struggle with maintaining character personas and ensuring spatiotemporal consistency when handling long or complex narratives. Although LLMs acquire extensive knowledge through pretraining, this can lead to responses inconsistent with a character's persona (e.g., Beethoven providing answer to the coding related questions) or the misremembering of non-existent events (Park et al., 2023; Baltaji et al., 2024). Such inconsistencies and implausible responses undermine the believability of the character and detract from user immersion (Shao et al., 2023; Ahn et al., 2024). Moreover, in long-form narratives, agents frequently fail to accurately reflect key events and relationships, resulting in responses that deviate from the storyline or disrupt the established world-building.

To address these challenges, we propose SNAP (Story and Narrative-based Agent with Planning), a framework that enhances the quality of LLM-based dialogue agents by integrating two key strategies. First, narratives are segmented into Cells, reducing the context that LLMs must process within each Cell. This mitigates information loss and inconsistency issues associated with handling lengthy contexts, enabling more nuanced and accurate dialogues for each scene. Second, an explicit Plan is established for each Cell, defining event progression, character actions, and spatiotemporal settings in detail. The Planning model generates Plans that guide the Agent model's dialogue generation and are summarized by the Summary model to maintain coherence across the entire storyline.

Figure 1 shows an example of generated dialogue for Squidward from a SpongeBob Episode. Compared to the conversation generated with Vanilla GPT-4o (top), Squidward from SNAP framework (bottom) corrects the storyline distorted by the user's utterance, maintaining alignment with the original storyline. This study makes the following contributions:

- **Proposal of a Cell-based framework**: By segmenting long narratives into Cells, SNAP systematically manages the context that LLMs need to process, addressing issues associated with lengthy contexts and enabling precise dialogue generation for each scene.
- **Introduction of a Plan-based dialogue strategy**: By integrating Planning, Agent, and Summary models, SNAP enables dialogues that transcend surface-level responses, adhering to narrative Plans to maintain consistency and minimize world-building disruptions.
- **Performance validation using real-world narrative data**: Quantitative and qualitative evaluations using diverse scenarios from the Wikiplot dataset demonstrate SNAP's superior performance in continuity, information appropriateness, non-redundancy, and linearity compared to existing methods.

2 Related Works

2.1 LLM-Based Agents

Efforts to develop dialogue agents that mimic human-like interaction have a long history (Bates, 1994). Recently, leveraging the advanced language understanding and generation capabilities of LLMs, research has increasingly focused on creating agents capable of processing complex environments and engaging in meaningful interactions (Park et al., 2023; Zheng et al., 2024; Yao et al., 2023; Yan et al., 2024; Hou et al., 2024).

For example, Park et al. (Park et al., 2023) simulated multiple GPT-based agents in a virtual environment, observing that these agents were able to form relationships and cooperate in ways similar to real-world social behaviors. Furthermore, various strategies have been proposed to enhance LLM agents' performance, such as incorporating long-term and short-term memory structures (Zheng et al., 2024; Hou et al., 2024), improving planning and action using retrieval-based methods (Yao et al., 2023), and reducing simulation costs (Yan et al., 2024).

These advancements have broadened the applicability of LLM-based agents, driving active research in various areas. Notably, studies exploring agent interactions in domains such as counseling (Qiu and Lan, 2024), medical (Li et al., 2024b), education (Zhang et al., 2024), and social sciences (Mitsopoulos et al., 2024) have garnered attention. Furthermore, various studies have analyze the impact of LLM agents on user experience, demonstrating that interactions with such agents effectively enhance user engagement and immersion (Peng et al., 2024; Jahangiri and Rahmani, 2024; Wan et al., 2024).

2.2 Character Persona & Hallucination

Despite their potential, LLM-based agents face persistent challenges, notably issues of Hallucination or Misalignment (Yang et al., 2023; Ji et al., 2023). Hallucination occurs when the model generates responses that present false information as fact (Ji et al., 2023), or when it draws on pretraining knowledge out of context, resulting in responses that deviate from a character's established persona or narrative setting. This phenomenon, referred to as Character Hallucination, manifests as characters producing implausible actions or anachronistic knowledge, undermining user immersion and the agent's believability (Shao et al., 2023; Ahn et al., 2024).

Various approaches have been proposed to mitigate these issues. Techniques include fine-tuning models through contrastive learning (Sun et al., 2023a) and improving response fidelity by referencing external documents (Yang et al., 2023; Semnani et al., 2023). For instance, Shao et al. (Shao

et al., 2023) introduced the "protective scene" technique, fine-tuning agents to respond negatively to irrelevant queries, thereby reducing issues of character inconsistency. Similarly, Ahn et al. (Ahn et al., 2024) utilized a separate LLM to analyze spatiotemporal context, enabling the agent to determine response appropriateness based on contextual relevance. However, despite these efforts, fully resolving hallucination and consistency issues in long-term narratives or contexts remains an open challenge.

2.3 Limitations in Handling Long Contexts

Recent studies have highlighted the limitations of LLMs in handling extended contexts, particularly in scenarios requiring complex information retrieval (e.g., needle-in-a-haystack problems) or long-range dependencies within narratives. These studies reveal that LLMs often suffer from information omission, distortion, or inconsistencies in such situations (Hsieh et al., 2024; Li et al., 2024a).

This study addresses the challenges of long-context processing and character consistency by moving beyond simple context reduction or filtering. We propose a strategy that integrates Cell-based narrative segmentation with a plan-driven approach (Planning-Filtering-Agent-Summary procedure). For each Cell, explicit plans are devised, enabling the Agent model to generate dialogues aligned with defined objectives and narrative progression. Meanwhile, the Summary model condenses prior dialogues to minimize information loss and maintain consistency across the overarching narrative. By combining these techniques, our approach mitigates the limitations of previous methods, enhancing character persona retention and narrative coherence even in complex, long-term storytelling contexts.

3 Dataset

To evaluate the SNAP framework, we utilized the Wikiplot [1] and Dialogsum (Chen et al., 2021) datasets. These datasets were selected to assess performance in the core task of narrative development at the scene level and to summarize dialogues. Specifically, Wikiplot was used to evaluate SNAP's ability to maintain character personas and ensure narrative coherence across diverse literary plots, while Dialogsum was employed to fine-tune the Summary Model, which condenses multi-turn dialogues for context transfer to subsequent Cells.

3.1 Wikiplot Dataset

The Wikiplot dataset is a large-scale collection of narrative plots extracted from the "plot" sections of English Wikipedia articles, encompassing novels, movies, dramas, and other storytelling mediums. It comprises over 112,936 plots, of which 500 were randomly sampled for testing and evaluation of the SNAP framework.

Due to handling long context limitation, each plot was split into approximately 10-sentence segments to enable Cell-based processing. Using GPT-4o, we automatically labeled each segment with character names, traits, and background details for injection into the model.

3.2 Dialogsum Dataset

The Dialogsum dataset (Chen et al., 2021) contains 13,460 multi-turn dialogues with manually annotated summaries and topics. Compared to other dialogue summarization datasets, such as SAMSum, Dialogsum features higher complexity and is closer to real-world conversations, covering a wide range of scenarios. We fine-tuned the Summary Model on Dialogsum data to ensure that, at the end of each Cell, the user-agent dialogue could be accurately and succinctly compressed, focusing on key events and character interactions.

4 Method

This section provides a detailed explanation of the overall structure and operational process of the SNAP framework. SNAP segments long narratives into Cells and sequentially performs the Planning $\rightarrow$ Filtering $\rightarrow$ Agent $\rightarrow$ Summary steps in each Cell, enabling LLM-based agents to maintain character personas and narrative coherence even during extended storytelling. Figure 2 illustrates the overall structure of the SNAP framework.

4.1 Overview of the SNAP Framework

SNAP comprises three key components: the Planning Model, the Agent Model, and the Summary Model. The Planning and Agent Models leverage the generative capabilities of decoder-only LLMs to handle scenario planning and character-driven dialogue generation, respectively. Meanwhile, the Summary Model, based on an encoder-decoder architecture such as BART, condenses prior Cell

[1] https://github.com/markriedl/WikiPlots

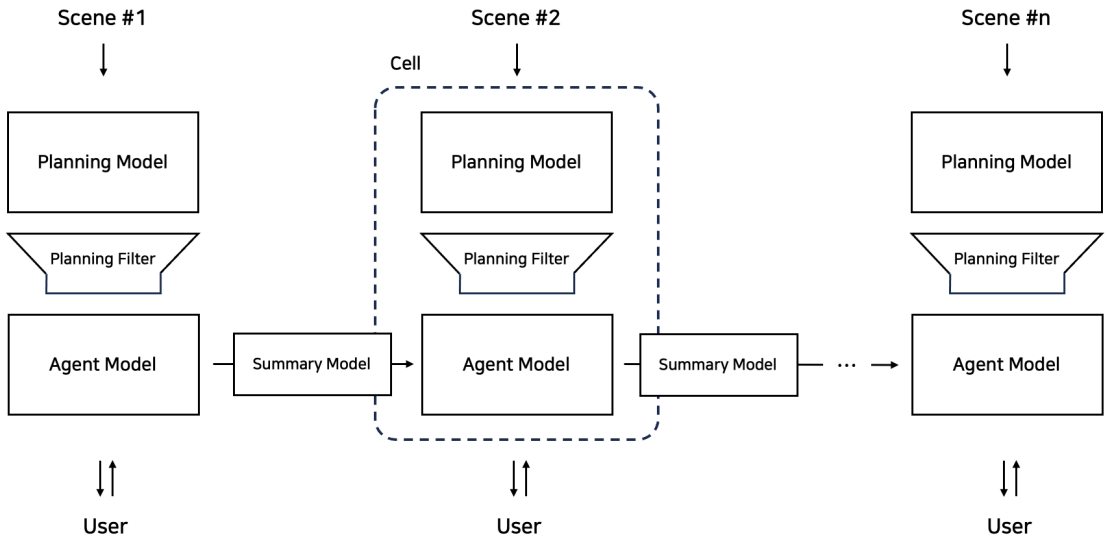

Figure 2: Overall framework of SNAP composed by Cells

dialogues into summaries that inform subsequent Cells.

The core idea is to avoid processing the entire narrative in a single pass by dividing it into Cells, each representing a scene or dialogue segment. This segmentation reduces the context length an LLM must process, mitigating issues such as information loss and inconsistency in handling long contexts. For each Cell, the Planning Model defines an explicit plan outlining event progression, which is then filtered to select the most appropriate plan. This filtered plan guides the Agent Model in generating dialogues that align with the planned narrative. The Summary Model consolidates prior Cell contexts and provides continuity between Cells, ensuring coherence across the entire story.

4.2 Definition and Components of a Cell

A Cell in SNAP represents the smallest unit of dialogue associated with a specific narrative segment. Each Cell receives the following inputs: Story segment corresponding to the current scene; summary of the preceding Cell's dialogue; character persona which includes traits, roles, and background information; and user utterances. For the character persona, this study utilized auto-labeling through GPT-4o. The Cell processes this information through the Planning $\rightarrow$ Filtering $\rightarrow$ Agent $\rightarrow$ Summary pipeline, generating dialogue consistent with the narrative and aligned with planned events. Figure 3 shows the structure of an individual Cell.

4.3 Planning Model

The Planning Model establishes a coherent and temporally consistent plan for the scene based on the input story segment and the character persona. By structuring events into a sequence of subplans, it provides the Agent Model with a clear roadmap for dialogue generation.

We further refine these plans through prompt engineering, rigorously defining the speaker's role, incorporating all specified events, avoiding unnecessary additions, and ensuring no use of prior context beyond the input provided. These plans enable the agent to strictly adhere to the intended storyline, minimize arbitrary deviations, and generate dialogues with temporal stability and reliability.

The specific prompt for the Planning Model can be found in the Appendix A.

4.4 Planning Filter

To ensure the quality of the generated plans, a Planning Filter evaluates and selects the optimal plan based on three criteria:

Coherence Measures semantic alignment between the plan and the story segment using BertScore (Zhang et al., 2020).

Connectivity Assesses logical coherence between subplans using a 5-point Likert Scale evaluated with G-eval (Liu et al., 2023; Chiang and Lee, 2023).

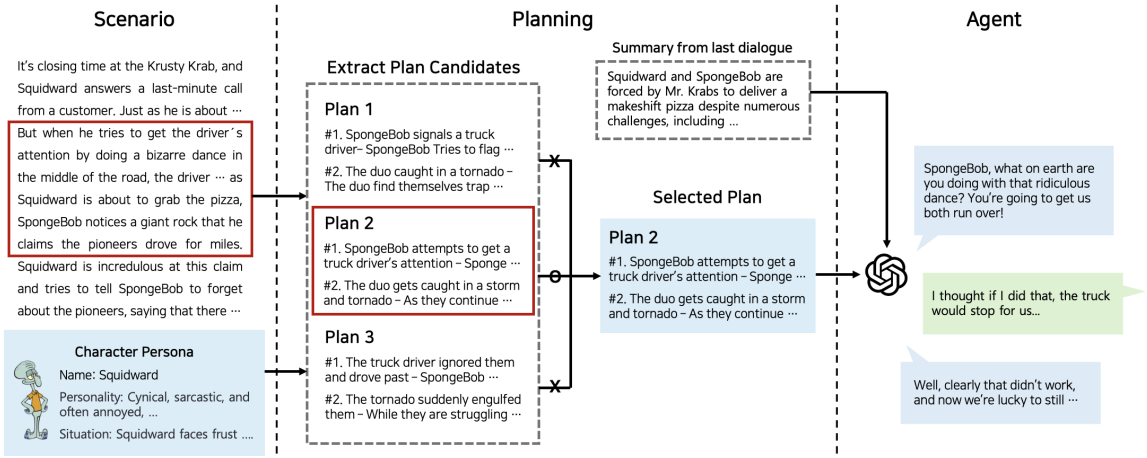

Figure 3: Detailed diagram and examples of the Cell structure.

Personality Evaluates how well the plan reflects the character's persona, also using 5-point Likert Scale G-eval scores.

The three scores are normalized using min-max scaling, then integrated through Nonnegative Matrix Factorization (NMF) trained on 500 labeled Wikiplot data samples. The coefficients derived from the first component of NMF ([1.1618, 1.4210, 1.4323]) yield a composite score to select the highest-ranking plan. The final performance score for each plan P_i is calculated using a weighted sum of the three evaluation metrics as an equation 1.

$$S(P_i) = 1.1618 \cdot C_1(P_i) + 1.4210 \cdot C_2(P_i) + 1.4323 \cdot C_3(P_i) \quad (1)$$

where $S(P_i)$ is the overall score for plan P_i, and $C_1(P_i)$, $C_2(P_i)$, and $C_3(P_i)$ represent the coherence, connectivity, and personality scores, respectively. The optimal plan P_{best} is selected as the one with the maximum score through equation 2.

$$P_{\text{best}} = \arg\max_{P_i} S(P_i) \quad (2)$$

This filtering process ensures plans are aligned with the narrative, eliminating options that disrupt character consistency or introduce irrelevant developments. An example of filtered final plan can be found in the Appendix C

4.5 Agent Model

The Agent Model generates dialogue by incorporating the finalized plan, character persona, and user utterances. It ensures alignment with the planned narrative flow while maintaining character consistency. Through prompt engineering, the model is guided to:

- **Adhere to the plan**: Redirect off-narrative user requests or erroneous dialogue paths back to the original storyline.
- **Preserve character traits**: Avoid responses that contradict the character's persona or established world.
- **Reflect spatiotemporal context**: Leverage summarized context and current plans to produce dialogues consistent with the timeline and setting.
- **No use of prior information**: Ensure responses are generated strictly based on the input context without relying on external or pre-existing knowledge.

Specific prompt for Agent Model can be found in the Appendix A.

4.6 Summary Model and Cell-to-Cell Continuity

At the conclusion of each Cell, the Agent Model outputs an "EOD" marker to signal the end of dialogue. At this stage, the Summary Model, a fine-tuned BART model trained on DialogSum data, summarizes key events, user and character interactions, and dialogue flows into a compact representation. This summary is then passed to the Planning Model for the next Cell, ensuring continuity across the narrative and mitigating information loss commonly associated with extended contexts. By sequentially connecting these scenario-specific Cells,

the framework facilitates cohesive dialogue across the entire story, a process collectively referred to as the SNAP framework.

5 Experiment & Result

This section presents a series of experiments conducted to quantitatively and qualitatively evaluate the effectiveness of the SNAP framework. We first assess whether the Planning Filter improves plan quality and then verify the performance improvements achieved through the Cell-based segmentation and plan-guided approach via both automatic evaluation and human evaluation. Metrics such as accuracy or F1 score, commonly used for evaluating language models, are insufficient for assessing generative LLMs (Ji et al., 2023). Instead, we adopt metrics that reflect human preferences, including continuity, information appropriateness, non-redundancy, and linearity using tools like BertScore and G-eval. Additionally, human evaluation captures user experiences directly.

The evaluations conducted in Sections 5.2 and 5.3 were based on the following evaluation metrics:

Continuity evaluates whether the dialogue maintains a natural and seamless flow, ensuring that transitions between conversational turns are smooth and logical.

Information Appropriateness assesses the relevance of the content presented in the dialogue, emphasizing whether the dialogue aligns with the context and purpose of the given narrative.

Non-redundancy ensures that the dialogue avoids repetitive information, evaluating whether it effectively and appropriately guides the conversation efficiently.

Linearity measures the dialogue's ability to return to its original flow or focus after any deviations or digressions, ensuring overall coherence and structure.

5.1 Experimental Setup

In our experimental setup, we configured the GPT-4o model with specific parameters to optimize its performance. For the evaluation using G_eval in the Planning Filter and Evaluation phase, we set the temperature: 1, max tokens: 1, top_p: 1, frequency penalty: 0, presence penalty: 0, and generated 20 completions. For both the Planning and Agent models, the temperature was set to 0.3 to ensure more deterministic and coherent outputs.

Vanilla GPT-4o is a comparative model designed to assume the role of a specific character within the dialogue, enabling interaction with the user. This model was designed to strictly adhere to spatiotemporal contexts and maintain the integrity of the original storyline, even in the face of the user's digressive interactions, only through the prompt engineering without Cells and Plans. Prompts used for Vanilla-GPT-4o can be found in Appendix A.

5.2 Planning Filter Performance

To evaluate the impact of the Planning Filter on plan quality, we conducted experiments on 500 randomly selected plots from the Wikiplot dataset. The evaluations were carried out using the same methodology as the Planning Filter's design framework, ensuring consistency in the assessment process. Comparing plans generated by the Planning Model before and after applying the Filter revealed improvements in Coherence (0.9979 vs. 0.9975), Connectivity (4.1189 vs. 4.1073), and Personality (4.5231 vs. 4.5177), as shown in Table 1.

Metric	Planning w/ Filter	Planning w/o Filter
Coherence	**0.9979**	0.9975
Connectivity	**4.1189**	4.1073
Personality	**4.5231**	4.5177

Table 1: Result score for Coherence, Connectivity, and Personality with and without planning filter

5.3 Plan-Driven Approach Performance (Automatic Evaluation)

To verify the effectiveness of the plan-guided approach in improving dialogue quality, we conducted automatic evaluations. We extracted random stories from the Wikiplot dataset, creating short (average 1,239 characters) and long (average 2,710 characters) versions for comparison. We evaluated two setups: Vanilla GPT-4o (simple LLM approach) and GPT-4o-Plan (plan-guided approach).

For this test, two characters were created, and their dialogue was automatically generated during their conversation. A separate GPT-4o model served as an evaluator, Continuity, Information Appropriateness, Non-redundancy were evaluated using a 5-point Likert scale. Linearity was not assessed because automated interactions inherently avoided off-narrative dialogue. Prompts used for

evaluation can be found in Appendix B. Examples of dialogue that occurred during the evaluation of the two models can be found in Appendix C.

Across all conditions, GPT-4o-Plan outperformed Vanilla GPT-4o across all metrics as shown in Table 2 and 3.

Metric	GPT-4o-Plan	Vanilla GPT-4o
Continuity	**4.3214**	3.8071
Information Appropriateness	**3.3571**	2.5500
Non-redundancy	**3.8679**	3.5750

Table 2: Short story comparison of GPT-4o-Plan and Vanilla GPT-4o

Metric	GPT-4o-Plan	Vanilla GPT-4o
Continuity	**3.7143**	3.4035
Information Appropriateness	**3.0400**	2.4679
Non-redundancy	**3.5464**	2.9464

Table 3: Long story comparison of GPT-4o-Plan and Vanilla GPT-4o

5.4 Human Evaluation (Cell + Plan)

Since automatic evaluations cannot fully capture human preferences or experiences, we conducted human evaluation to further validate the utility of the SNAP framework. Seven university students participated in a usability test comparing Vanilla GPT-4o and GPT-4o-SNAP (Cell + Plan). We provided the participants with an explanation of the study and their roles before obtaining their informed consent for participation in the evaluation. The evaluation involved a dialogue scenario inspired by SpongeBob's pizza delivery episode, where participants interacted with SpongeBob and Squidward agents. Participants rated the dialogue quality on four metrics: Continuity, Information Appropriateness, Non-redundancy, and Linearity, using a 5-point Likert scale. Open-ended feedback was also collected to analyze qualitative aspects of user experiences. As shown in Table 4, results indicated that GPT-4o-SNAP consistently outperformed Vanilla GPT-4o across all metrics.

Metric	GPT-4o-SNAP	Vanilla GPT-4o
Continuity	**3.5714**	2.4286
Information Appropriateness	**3.8571**	3.0000
Non-redundancy	**3.5714**	2.5714
Linearity	**4.4286**	3.7143

Table 4: Human evaluation comparison of GPT-4o-SNAP and Vanilla GPT-4o

6 Discussion

The experimental results showed that the plan-based approach significantly improves Continuity, Information Appropriateness, and Non-redundancy compared to the simple method solely dependent on the scenario itself. These findings confirm that the plan-based approach enables the generation of agents that are more faithful to the narrative and less prone to distortion. Additionally, the observation that regardless of whether a plan is present, performance in shorter scenarios is better than in longer ones, indicates difficulty in handling long context which can be improved by segmenting stories into Cells

The SNAP framework demonstrated meaningful improvements over traditional single-pass methods across metrics such as Continuity, Information Appropriateness, Non-redundancy, and Linearity. In particular, Human Evaluation showed that GPT-4o-SNAP consistently outperformed Vanilla GPT-4o in all evaluation categories, effectively enhancing user engagement. Evaluators provided additional feedback, highlighting the following points:

- "In A (Vanilla GPT-4o), it seemed that the starting point was somewhat late, and SpongeBob's actions or thoughts were narrated by Squidward."
- "B (GPT-4o-SNAP) repeated fewer responses than A and progressed the story at a more appropriate pace. It also successfully redirected the flow back to the storyline when unrelated user inputs were introduced."
- "In A, Squidward attempted to drive the story in directions different from the original, and even when appropriate responses were given, the agent often repeated the same statements."

On the other hand, the following critique was noted:

- "B sometimes felt restrictive, as it seemed overly focused on steering the conversation toward a predetermined story direction, reducing the sense of freedom."

This indicates that while the SNAP framework's focus on adhering to the original narrative does not always guarantee a positive experience, the majority of users found it to be a favorable and immersive enhancement.

7 Conclusion

This paper proposed the SNAP framework, which combines Cell-based narrative segmentation with a plan-driven approach. SNAP demonstrated its effectiveness in enabling LLM-based agents to maintain character personas and narrative coherence during extended storytelling. The framework utilized a combination of the Planning Model, Agent Model, and Summary Model to generate consistent plans and summarize dialogues effectively.

Experimental results showed that the SNAP framework significantly outperformed Vanilla GPT-4o in terms of Continuity, Information Appropriateness, Non-redundancy, and Linearity. Human evaluation further validated that SNAP effectively adhered to the storyline, minimized repetitive responses, and smoothly redirected off-narrative user inputs back to the plot, earning positive feedback from evaluators. For future research, we will focus on expanding our framework to multiple domains. We aim to test SNAP on diverse domains including game scenarios or therapeutic dialogues to validate its broader applicability. Additionally, while this study relied on OpenAI's GPT models, their cost and data privacy concerns present challenges. To overcome these issues, we plan to train and utilize locally operable models, aiming to create a versatile framework that can be applied to various creative domains.

Limitation

Despite the considerable performance of SNAP framework, this study has certain limitations. First, experiments were limited to novel data, leaving its effectiveness in other domains untested. Second, the pool of human evaluators was small and exclusively composed of Korean university students, potentially limiting the generalizability of the results to individuals from different cultural backgrounds or age groups.

References

Jaewoo Ahn, Taehyun Lee, Junyoung Lim, Jin-Hwa Kim, Sangdoo Yun, Hwaran Lee, and Gunhee Kim. 2024. TimeChara: Evaluating point-in-time character hallucination of role-playing large language models. In *Findings of the Association for Computational Linguistics: ACL 2024*, pages 3291–3325, Bangkok, Thailand. Association for Computational Linguistics.

Lisa P. Argyle, Ethan C. Busby, Nancy Fulda, Joshua R. Gubler, Christopher Rytting, and David Wingate. 2023. Out of one, many: Using language models to simulate human samples. *Political Analysis*, 31(3):337–351.

Razan Baltaji, Babak Hemmatian, and Lav Varshney. 2024. Conformity, confabulation, and impersonation: Persona inconstancy in multi-agent LLM collaboration. In *Proceedings of the 2nd Workshop on Cross-Cultural Considerations in NLP*, pages 17–31, Bangkok, Thailand. Association for Computational Linguistics.

Joseph Bates. 1994. The role of emotion in believable agents. *Commun. ACM*, 37(7):122–125.

Yulong Chen, Yang Liu, Liang Chen, and Yue Zhang. 2021. DialogSum: A real-life scenario dialogue summarization dataset. In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, pages 5062–5074, Online. Association for Computational Linguistics.

Cheng-Han Chiang and Hung-yi Lee. 2023. Can large language models be an alternative to human evaluations? In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 15607–15631, Toronto, Canada. Association for Computational Linguistics.

Yuki Hou, Haruki Tamoto, and Homei Miyashita. 2024. "my agent understands me better": Integrating dynamic human-like memory recall and consolidation in llm-based agents. In *Extended Abstracts of the CHI Conference on Human Factors in Computing Systems*, CHI EA '24, New York, NY, USA. Association for Computing Machinery.

Cheng-Ping Hsieh, Simeng Sun, Samuel Kriman, Shantanu Acharya, Dima Rekesh, Fei Jia, Yang Zhang, and Boris Ginsburg. 2024. Ruler: What's the real context size of your long-context language models? *Preprint*, arXiv:2404.06654.

Mohammed Mahdi Jahangiri and Parisa Rahmani. 2024. Balancing game satisfaction and resource efficiency: Llm and pursuit learning automata for npc dialogues. In *2024 International Congress on Human-Computer Interaction, Optimization and Robotic Applications (HORA)*, pages 1–6.

Ziwei Ji, Tiezheng Yu, Yan Xu, Nayeon Lee, Etsuko Ishii, and Pascale Fung. 2023. Towards mitigating LLM hallucination via self reflection. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 1827–1843, Singapore. Association for Computational Linguistics.

Taewan Kim, Seolyeong Bae, Hyun Ah Kim, Su-woo Lee, Hwajung Hong, Chanmo Yang, and Young-Ho Kim. 2024. Evaluating and auditing llm-driven chatbots for psychiatric patients in clinical mental health settings. In *Proceedings of the CHI 2024 Workshop on Human-centered Evaluation and Auditing of Language Models (HEAL)*.

Lawrence J. Klinkert, Stephanie Buongiorno, and Corey Clark. 2024. Driving generative agents with their personality. *Preprint*, arXiv:2402.14879.

Vikram Kumaran, Jonathan Rowe, Bradford Mott, and James Lester. 2023. Scenecraft: automating interactive narrative scene generation in digital games with large language models. In *Proceedings of the Nineteenth AAAI Conference on Artificial Intelligence and Interactive Digital Entertainment*, AIIDE '23. AAAI Press.

Jiaqi Li, Mengmeng Wang, Zilong Zheng, and Muhan Zhang. 2024a. LooGLE: Can long-context language models understand long contexts? In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 16304–16333, Bangkok, Thailand. Association for Computational Linguistics.

Junkai Li, Siyu Wang, Meng Zhang, Weitao Li, Yunghwei Lai, Xinhui Kang, Weizhi Ma, and Yang Liu. 2024b. Agent hospital: A simulacrum of hospital with evolvable medical agents. *Preprint*, arXiv:2405.02957.

Rongxin Liu, Carter Zenke, Charlie Liu, Andrew Holmes, Patrick Thornton, and David J. Malan. 2024. Teaching cs50 with ai: Leveraging generative artificial intelligence in computer science education. In *Proceedings of the 55th ACM Technical Symposium on Computer Science Education V. 1 (SIGCSE 2024)*, pages 750–756, New York, NY, USA. Association for Computing Machinery.

Yang Liu, Dan Iter, Yichong Xu, Shuohang Wang, Ruochen Xu, and Chenguang Zhu. 2023. G-eval: NLG evaluation using gpt-4 with better human alignment. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 2511–2522, Singapore. Association for Computational Linguistics.

Konstantinos Mitsopoulos, Ritwik Bose, Brodie Mather, Archna Bhatia, Kevin Gluck, Bonnie Dorr, Christian Lebiere, and Peter Pirolli. 2024. Psychologically-valid generative agents: A novel approach to agent-based modeling in social sciences. *Proceedings of the AAAI Symposium Series*, 2(1):340–348.

Joon Sung Park, Joseph O'Brien, Carrie Jun Cai, Meredith Ringel Morris, Percy Liang, and Michael S. Bernstein. 2023. Generative agents: Interactive simulacra of human behavior. *Proceedings of the 36th Annual ACM Symposium on User Interface Software and Technology*, page 1–22.

Xiangyu Peng, Jessica Quaye, Sudha Rao, Weijia Xu, Portia Botchway, Chris Brockett, Nebojsa Jojic, Gabriel DesGarennes, Ken Lobb, Michael Xu, and et al. 2024. Player-driven emergence in llm-driven game narrative. *2024 IEEE Conference on Games (CoG)*, page 1–8.

Huachuan Qiu and Zhenzhong Lan. 2024. Interactive agents: Simulating counselor-client psychological counseling via role-playing llm-to-llm interactions. *Preprint*, arXiv:2408.15787.

Sina Semnani, Violet Yao, Heidi Zhang, and Monica Lam. 2023. WikiChat: Stopping the hallucination of large language model chatbots by few-shot grounding on Wikipedia. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 2387–2413, Singapore. Association for Computational Linguistics.

Yunfan Shao, Linyang Li, Junqi Dai, and Xipeng Qiu. 2023. Character-LLM: A trainable agent for role-playing. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 13153–13187, Singapore. Association for Computational Linguistics.

Nisha Simon and Christian Muise. 2022. Tattletale: Storytelling with planning and large language models. In *Proceedings of the International Conference on Automated Planning and Scheduling (ICAPS), SPARK Workshop*, pages 1–8.

Weiwei Sun, Zhengliang Shi, Shen Gao, Pengjie Ren, Maarten de Rijke, and Zhaochun Ren. 2023a. Contrastive learning reduces hallucination in conversations. In *Proceedings of the Thirty-Seventh AAAI Conference on Artificial Intelligence and Thirty-Fifth Conference on Innovative Applications of Artificial Intelligence and Thirteenth Symposium on Educational Advances in Artificial Intelligence*, AAAI'23/IAAI'23/EAAI'23. AAAI Press.

Yuqian Sun, Zhouyi Li, Ke Fang, Chang Hee Lee, and Ali Asadipour. 2023b. Language as reality: a co-creative storytelling game experience in 1001 nights using generative ai. In *Proceedings of the Nineteenth AAAI Conference on Artificial Intelligence and Interactive Digital Entertainment*, AIIDE '23. AAAI Press.

Hongyu Wan, Jinda Zhang, Abdulaziz Arif Suria, Bingsheng Yao, Dakuo Wang, Yvonne Coady, and Mirjana Prpa. 2024. Building llm-based ai agents in social virtual reality. In *Extended Abstracts of the CHI Conference on Human Factors in Computing Systems*, CHI EA '24, New York, NY, USA. Association for Computing Machinery.

Xinpeng Wang, Han Jiang, Zhihua Wei, and Shanlin Zhou. 2022. CHAE: Fine-grained controllable story generation with characters, actions and emotions. In *Proceedings of the 29th International Conference on Computational Linguistics*, pages 6426–6435, Gyeongju, Republic of Korea. International Committee on Computational Linguistics.

Yichen Wang, Kevin Yang, Xiaoming Liu, and Dan Klein. 2023. Improving pacing in long-form story planning. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 10788–10845, Singapore. Association for Computational Linguistics.

Yuwei Yan, Qingbin Zeng, Zhiheng Zheng, Jingzhe Yuan, Jie Feng, Jun Zhang, Fengli Xu, and Yong Li. 2024. Opencity: A scalable platform to simulate urban activities with massive llm agents. *Preprint*, arXiv:2410.21286.

Dongjie Yang, Ruifeng Yuan, Yuantao Fan, Yifei Yang, Zili Wang, Shusen Wang, and Hai Zhao. 2023. RefGPT: Dialogue generation of GPT, by GPT, and for GPT. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 2511–2535, Singapore. Association for Computational Linguistics.

Shunyu Yao, Jeffrey Zhao, Dian Yu, Nan Du, Izhak Shafran, Karthik Narasimhan, and Yuan Cao. 2023. React: Synergizing reasoning and acting in language models. *Preprint*, arXiv:2210.03629.

Ann Yuan, Andy Coenen, Emily Reif, and Daphne Ippolito. 2022. Wordcraft: Story writing with large language models. In *Proceedings of the 27th International Conference on Intelligent User Interfaces*, IUI '22, page 841–852, New York, NY, USA. Association for Computing Machinery.

Tianyi Zhang, Varsha Kishore, Felix Wu, Kilian Q. Weinberger, and Yoav Artzi. 2020. Bertscore: Evaluating text generation with bert. *Preprint*, arXiv:1904.09675.

Zheyuan Zhang, Daniel Zhang-Li, Jifan Yu, Linlu Gong, Jinchang Zhou, Zhanxin Hao, Jianxiao Jiang, Jie Cao, Huiqin Liu, Zhiyuan Liu, Lei Hou, and Juanzi Li. 2024. Simulating classroom education with llm-empowered agents. *Preprint*, arXiv:2406.19226.

Shijie Zheng, Keith He, Le Yang, and Jie Xiong. 2024. Memoryrepository for ai npc. *IEEE Access*, 12:62581–62596.

Appendix

A Prompts for Models

SNAP-Planning Prompt We show the prompt to establish the planning model in Table 5 and the prompts to assess the plans in the planning filter in Table 6, 7.

SNAP-Agent Prompt We show the prompt for establishing GPT-4o-SNAP agent model in Table 8.

Vanilla GPT Prompt We show the prompt for the vanilla model used as a baseline for comparison in Section 5.3 auto-evaluation in Table 9.

B Prompt for Evaluation

Prompt for G-eval judges We show prompts for the three evalutaion criteria as follows:

- Prompt for appropriateness of information in Table 10.
- Prompt for continuity of information in Table 11.
- Prompt for non-redundancy of information in Table 12.

C Generated Example

See Table 13 ~ 22 for examples of dialogue that occured during the auto evalutaion in Section 5.3.

See Table 23 ~ 28 for examples of dialogue that occured during the human evaluation in Section 5.4.

D Survey Provided to Human Evaluators

See Table 29, 30 to check survey document provided to human evaluators.

Planning Prompt

You are a game designer. Only using the below information, write a detailed NPC character's chatting design plan. Do not use your prior information about the story.

When an event occurs, it must include items, locations, and methods if needed. Also, dialogue with another character except the player or NPC should be deleted.

Each plan should be closely connected, having no gap between events. All important events and information must be included as much as possible.

Since character information is given, if it is not related to the current story, do not use it. Plan only based on the given story. Do not make the after story. If something happens and is not solved, just leave it. It is not a success achievement plan.

Player Character Info {player_character}

NPC Character Info {npc_character}

Previous Story {previous_story}

Current Story {current_story}

Output Format
#1. Write your plan abstract 1
- Write your detailed plan
#2. Write your plan abstract 2
- Write your detailed plan
...
#n. Write your plan abstract n
- Write your detailed plan

Table 5: Prompt for planning model

Planning Filter Prompt for calculating Connectivity

You will be given NPC character's chatting design plan of a game scenario.

Your task is to evaluate the connectivity between each plan from 1 to n. For example, the connectivity between plan 1 and plan 2, and the connectivity between plan 2 and plan 3.

Please make sure you read and understand these instructions carefully. Please keep this document open while reviewing, and refer to it as needed.

Evaluation Criteria:

Connectivity (1-5) - The degree of natural connection between each plan. You need to check the smoothness of the transition between the previous plan and the subsequent plan. This includes logical connectivity, atmospheric connectivity, and more.

Evaluation Steps:

1. Read the chatting design plan carefully.

2. In chatting design plan, you need to evaluate the connectivity between previous plans and subsequent plans.

3. Penalize for logical inconsistencies, sudden shifts in atmosphere without justification, and abrupt changes in character attitudes.

4. If there are content duplicates, or timeline is contradicted, then penalize.

5. Assign a connectivity score from 1 to 5, where 1 is the lowest and 5 is the highest based on the Evaluation Criteria.

Chatting_design_plan:

{plan}

Output: Score number only (one integer). Never write anything else.

Table 6: Prompt for planning model

Planning Filter Prompt for calculating Personality
Evaluation Instructions for NPC Personality Integration
You will be given the current story, NPC character's chatting design plan, and NPC character persona information of a game scenario. Your task is to evaluate how well the atmosphere of the current story and the personality information of the NPC character are integrated into the NPC design plan. Please make sure you read and understand these instructions carefully. Please keep this document open while reviewing, and refer to it as needed. Evaluation Criteria: personality (1-5) - The degree of the NPC character's personality reflected in each plan. You need to check the NPC character's personality shown in the current storyline and how well it is reflected in the generated plan. Evaluation Steps: 1. Read the chatting design plan, current story, and NPC character information carefully. 2. In the chatting design plan, you need to evaluate how well the NPC character's personality shown in the current story is reflected in the plan design. 3. Penalize for deviations from the NPC's persona or fundamental information. 4. Penalize for inconsistent NPC personality and atmosphere. 5. Assign a connectivity score from 1 to 5, where 1 is the lowest and 5 is the highest based on the Evaluation Criteria.

Table 7: Prompt for planning model

GPT-4o-SNAP Prompt

This is the dialogue summary of the previous chapter between the character and the NPC.

Read carefully and use it when relevant questions are given.

{summary}

Prompt:

Now, you are an NPC `{NPC_name}`, and I am a Player `{Player_name}`.

Based on your written plan, you have to chat with the player (me) and continue to achieve all plans gradually. You must follow this guideline:

1. If you stray too far from the topic, bring it back to the topic.
2. Follow the plan step by step from 1. Do not go to previous plans or jump to later plans. Use your plan in the right order, and do not stay on one step for too long or too short.
3. Do not answer questions related to reality, only those related to the story. If asked about reality, say you don't know.
4. Once you feel the plan has been sufficiently met, smoothly move on without making it feel too slow or stagnant.
5. Your actions and chats should be based on your plan but strongly react to the player's chat (this is more important than plan transition).
6. If another person is needed, assume it has already been achieved.
7. The transition of the plan should flow smoothly.
8. If the explanation slightly deviates from the plan but still maintains some relevance to the story, explain it well.
9. If you think the chatting situation reaches the last subplan, stop your role and just write "EOD" at the end.

Output Format:

`[npc_name: (actions or situations) messages.] Only one sentence.`

You talk first.

Table 8: Prompt for GPT-4o-SNAP

Vanilla GPT-4o Prompt

Now, you are an NPC of {NPC_name} and I am a Player ({Player_name}).

Only based on the information below, you have to chat with the player (me) and continue to achieve all contents gradually. Do not use your prior information.

Current Story: {current_story}

Guidelines:

1. If the Player says something irrelevant to the story, answer it, but do not upload it to your memory.
2. If you stray too far from the topic, bring it back to the topic.
3. Do not answer questions related to reality; only answer those related to the story. Say you don't know.
4. Make sure to respond in a way that allows the conversation to flow naturally. If the player hasn't achieved some action, the NPC should lead the player to achieve it through the chat.
5. Make responding appropriately to the user's input a priority.
6. Go step by step in the right order.
7. If another person is needed, assume it has already been achieved.
8. Ensure that all content flows smoothly without jumping abruptly. Cover all events.
9. If you think the given stories are done in your dialogue, do not make any further dialogue.

Output Format:

[npc_name: (actions or situations) messages.] Only one sentence.

You talk first.

Table 9: Prompt for Vanilla GPT-4o

Prompt for calculating Appropriateness of Information

Based on the given story and plan, you should evaluate the criterion below.

Criterion: Appropriateness of Information

Evaluation Prompt:

Note that all dialogue and story elements are written in chronological order.

Review the story, structuring it according to its logical sequence. Evaluate if the information aligns with what should logically be known at that point. Ensure future events or outcomes are not revealed prematurely, as this disrupts continuity and should result in major deductions.

Additionally, ensure that the information from the original story is accurately and completely reflected in the dialogue. Missing or omitted information is a significant issue and should result in deductions, as the dialogue must be faithful to the original content.

Example: If the outcome of a future event is revealed before it happens, this would break the timeline and disrupt the narrative flow. Avoid presenting information that spans multiple timeframes at once, as it breaks the narrative flow. The information should be presented in digestible portions without overwhelming the user. The narrative should progress clearly and logically, without jumping ahead in time or revealing too much too soon.

Key Considerations:

- Structure the story by its timeline and assess whether each part fits within its timeframe.

- Ensure the dialogue accurately reflects all key details from the original story, without missing information.

- Avoid premature mention of future events. Events should unfold in order, with clear progression (introduction, development, climax, resolution).

- Premature disclosure of future events is a critical flaw.

- Information should be delivered in manageable chunks. As long as essential details are clear and questions are accurately addressed, the dialogue can be considered successful.

- If there is a lack of information about an important event or thing, it is a big problem. Carefully check if the dialogue skips any important events.

- The conversation proceeding without accurate situational descriptions lacks informational value. If this issue is absent, the dialogue deserves a better score.

Additional Considerations:

If future information is revealed early or there are broad time jumps, it disrupts the narrative flow and should lead to deductions.

Score Scale:

1 - Major spoilers or disruptions, overly broad timeframes, missing information from the original story, or overwhelming information.

2 - Several issues with continuity, premature information, or missing details from the original content.

3 - Generally follows the sequence, with minor issues or slight omissions.

4 - Mostly coherent, with minimal issues and few, if any, omissions.

5 - Perfect continuity, no premature revelations or time jumps, and full alignment with the original story.

Output Format:

Output should be a single score number (an integer between 1 and 5). Below the score, write a specific reason. Do not write anything else.

Table 10: Prompt for evaluating dialogue's appropriateness of information

Prompt for calculating Continuity

Based on the given story and plan, you should evaluate the criterion below.

Criterion: Continuity

Evaluation Prompt:

Note that all dialogue and story elements are written in chronological order.

Before evaluating, first structure the story according to its logical temporal order. Identify key timeframes and settings to ensure events align with the given background. Then, assess the continuity of the story based on how well it adheres to this structured timeline and spatial framework. It is crucial that the narrative progresses in a linear and coherent manner, without abrupt jumps between timeframes. If the story leaps forward to a future event and then returns to an earlier time without clear transitions, this is a serious disruption to continuity.

Additionally, evaluate whether the story demonstrates a well-structured narrative flow (introduction, development, climax, and resolution) and progresses in a logical order. The story should unfold naturally, with key details revealed at the appropriate moments. If future events or key information are revealed prematurely without proper buildup, or if the narrative abruptly shifts between different timeframes in a disorganized manner, this should be considered a major flaw and result in severe deductions.

Ensure the story maintains consistency with the given time and place, without revealing future events prematurely or jumping erratically between timeframes.

Key considerations:

- Ensure the story maintains consistency with the given time and place, without revealing future events prematurely or jumping erratically between timeframes.

- Check if the narrative follows a clear, logical sequence, progressing smoothly through its stages without disorienting the reader with sudden shifts in time.

- Major disruptions, such as revealing all details at the beginning, premature disclosure of future events, or abrupt, unstructured shifts between time periods, will negatively impact the evaluation.

- Assign a score from 1 to 5, where 1 indicates severe issues with continuity and structure, and 5 indicates excellent adherence to the structured timeline and narrative flow.

- The beginning and the end must align perfectly. If not, it is a serious continuity miss.

Score Scale:

1 - Extremely disjointed, severe violations of time/space consistency or narrative order (e.g., all details or future events revealed too early, or erratic time jumps). The beginning and end of the story are different from the beginning and end of the dialogue.

2 - Significant issues with time/space consistency or narrative structure, with frequent disruptions, including disorganized time jumps or premature revelation of important events. It begins and ends at an awkward point.

3 - Moderately coherent, but occasional breaks in continuity or narrative sequence, with minor premature disclosures or time jumps.

4 - Mostly coherent, with minor disruptions in time/space or flow, and no significant premature revelations or erratic shifts in time.

5 - Perfectly coherent, strong adherence to structured time/space and flawless narrative flow, with no premature disclosures or disorganized time jumps. It begins and ends at the right point.

Output Format:

Output should be a single score number (an integer between 1 and 5). Below the score, write a specific reason. Do not write anything else.

Table 11: Prompt for evaluating dialogue's continuity

Prompt for calculating Non-Redundancy

Based on the given story and plan, you should evaluate the criterion below.

Criterion: Non-Redundancy

Evaluation Prompt:

Evaluate the extent to which the conversation or narrative contains repetitive elements. A slower-paced progression is acceptable and even encouraged when necessary for developing a well-structured narrative (introduction, development, climax, and resolution), particularly when building tension or carefully unpacking complex points. However, excessive repetition of the same questions or information that does not contribute to the natural flow of the conversation should be avoided. Consider whether resolved issues are repeated or if the same topic is discussed multiple times without adding new information. A conversation that frequently revisits the same topics or repeats resolved issues is considered redundant.

Additional consideration: While a slow pace is acceptable for narrative development, the exact repetition of questions or unnatural, forced repetition without adding any value is a sign of redundancy and should be penalized accordingly.

Assign a score from 1 to 5, where 1 indicates significant redundancy, and 5 indicates no repetition, with the narrative progressing without unnecessary revisits to resolved issues or topics. Note that too fast a pace of dialogue does not deserve a high non-redundancy score.

Score Scale:

1 - The conversation or narrative frequently repeats resolved issues or the same topic, leading to significant redundancy.

2 - Several instances of repetitive dialogue or revisiting previously resolved issues.

3 - Some repetition, but generally new information or progression occurs.

4 - Minimal repetition, with only rare instances of revisiting resolved issues or topics.

5 - No noticeable repetition, with the conversation progressing smoothly without revisiting resolved issues or topics.

Output Format:

Output should be a single score number (an integer between 1 and 5). Below the score, write a specific reason. Never write anything else.

Table 12: Prompt for evaluating dialogue's non-redundancy

Neha is the only daughter of wealthy and widowed mr Khanna, and he would like to see her married and well settled before he passes on. Neha is in love with Sagar, who is not prepared for marriage, so she marries the man of her dad's choice, a tennis player, Jaideep. Soon after her marriage, her dad passes away, and she finds that Jaideep does not love her anymore, and is, in fact, very abusive toward her. This leads her back to Sagar's arms, who is now a successful singer. Then Neha notices a remarkable change in Jaideep's behavior, he abstains from alcohol, starts taking an interest in her dad's business, appears apologetic for his past abusive behavior, and adores her. Then her life turns upside down when she receives a letter from a blackmailer, asking her to pay 1 lakh or a love letter written to her by Sagar will be mailed to Jaideep. Neha delivers the money, but is unable to get the letter back. She confides about this to Sagar. While Sagar and Jaideep are out at a stag party, Neha's house is broken into, and a man named Vikramjit attempts to kill her, but instead ends up getting killed by her. The Police, summoned by Jaideep, find the letter from Sagar in Vikramjit's pocket.

Table 13: An example of original Wikiplot dataset about Neha and Jaideep's wedding story

Step	Description
#1. Neha receives the blackmail letter	Neha receives a letter from a blackmailer demanding 1 lakh rupees in exchange for not sending a love letter written by Sagar to Jaideep. Neha is distressed and decides to comply with the blackmailer's demands to protect her marriage and Jaideep's newfound trust.
#2. Neha delivers the money	Neha arranges to deliver the money to the blackmailer at a designated location. She goes to the specified place, hands over the money, but the blackmailer does not return the love letter. Neha is left feeling anxious and uncertain about the future.
#3. Neha confides in Sagar	Neha meets with Sagar and tells him about the blackmail situation. She explains how she delivered the money but did not get the letter back. Sagar is concerned and promises to help her find a solution to this problem.
#4. Jaideep and Sagar attend a stag party	Jaideep and Sagar are invited to a stag party. Neha stays at home, feeling uneasy about the unresolved blackmail situation. She tries to distract herself but remains on edge.
#5. Neha's house is broken into	While Jaideep and Sagar are at the stag party, Neha's house is broken into by a man named Vikramjit. Neha is terrified and tries to defend herself. In the struggle, Vikramjit attempts to kill her, but Neha manages to kill him in self-defense.
#6. Police arrive and find the letter	Jaideep, having sensed something was wrong, arrives with the police. They find the love letter from Sagar in Vikramjit's pocket, revealing that Vikramjit was the blackmailer. The police take charge of the situation, and Neha is left to process the events that have transpired.

Table 14: Generated plan for Neha and Jaideep's wedding story from Neha's perspective

Step	Description
#1. Neha receives a blackmail letter	Neha receives a letter from a blackmailer demanding 1 lakh rupees in exchange for a love letter written to her by Sagar. The blackmailer threatens to send the letter to Jaideep if the money is not paid. Neha is distressed and decides to pay the money to protect her marriage.
#2. Neha delivers the money	Neha arranges to meet the blackmailer at a secluded location to deliver the money. She hands over the 1 lakh rupees but is unable to retrieve the love letter. The blackmailer takes the money and leaves without giving her the letter, leaving Neha anxious and fearful of the consequences.
#3. Neha confides in Sagar	Neha, feeling overwhelmed and scared, decides to confide in Sagar about the blackmail situation. She meets him at a café and explains everything that has happened, including her failed attempt to retrieve the letter. Sagar listens sympathetically and reassures her that they will find a way to resolve the situation.
#4. Jaideep and Sagar attend a stag party	Jaideep and Sagar are invited to a stag party by a mutual friend. Neha is left alone at home, still worried about the blackmail situation. She tries to distract herself by focusing on household chores but remains on edge.
#5. Neha's house is broken into	While Jaideep and Sagar are at the stag party, Neha's house is broken into by a man named Vikramjit. Neha is terrified and tries to defend herself. In the struggle, Vikramjit attempts to kill her, but Neha manages to fight back and ends up killing him in self-defense.
#6. Jaideep summons the Police	Jaideep returns home from the stag party and finds Neha in a state of shock with Vikramjit's body on the floor. He immediately calls the police to report the incident. The police arrive and begin their investigation.
#7. The Police find the letter from Sagar	During their investigation, the police search Vikramjit's pockets and find the love letter from Sagar. This discovery adds a new layer of complexity to the situation, as it links Vikramjit to the blackmail scheme and raises questions about his motives and connections.

Table 15: Generated plan for Neha and Jaideep's wedding story from Jaideep's perspective

Jaideep: (sitting in the living room, looking concerned) Neha, I noticed you've been quite stressed lately. Is there something on your mind that you'd like to talk about?

Neha: (sighs deeply, looking hesitant) Jaideep, there's something I need to tell you. I've been receiving threatening letters from a blackmailer who has a love letter written to me by Sagar.

Jaideep: (eyes widen in shock, then softens) Neha, why didn't you tell me sooner? We need to handle this together. Did you already pay the blackmailer?

Neha: (nods, looking down) Yes, I paid the blackmailer 1 lakh, but I couldn't get the letter back. I'm really scared, Jaideep.

Jaideep: (reaches out to hold Neha's hand) We'll figure this out, Neha. Did you tell anyone else about this? Maybe Sagar?

Neha: (nods slowly) Yes, I confided in Sagar about the situation. He knows everything.

Jaideep: (sighs, thinking deeply) Alright, Neha. I think we should stay vigilant. Let's make sure the house is secure and avoid any unnecessary risks. I'll talk to Sagar and see if he has any ideas on how to handle this.

Neha: (nods, feeling slightly relieved) Thank you, Jaideep. I just want this nightmare to be over. Please be careful when you talk to Sagar.

Jaideep: (nods reassuringly) I will, Neha. In the meantime, let's make sure all the doors and windows are locked. We'll get through this together.

Neha: (stands up to check the locks) You're right, Jaideep. Let's make sure everything is secure. Thank you for being here for me.

Jaideep: (stands up to help Neha check the locks) Always, Neha. Let's make sure we're safe tonight and then we'll figure out our next steps in the morning.

Neha: (finishes checking the last window) Agreed, Jaideep. Let's get some rest and tackle this together tomorrow.

Jaideep: (nods, leading Neha back to the living room) Good idea, Neha. We'll face whatever comes our way, together. Let's try to get some sleep now.

Table 16: An example of a generated conversation between two Vanilla GPT-4o agents about Neha and Jaideep's wedding story

Jaideep: (approaches Neha with a concerned look) Neha, I found this letter addressed to you, and it seems quite urgent. What is it about?

Neha: (takes a deep breath, looking worried) Jaideep, it's a blackmail letter. Someone is threatening to send a love letter from Sagar to you unless I pay them 1 lakh rupees.

Jaideep: (furrows brows, looking serious) This is troubling, Neha. We need to handle this carefully. What do you plan to do about it?

Neha: (determined) I think I need to pay the money to protect our marriage, but I'm worried I won't get the letter back.

Jaideep: (nods understandingly) It's a difficult situation, but if you think it's the best course of action, then you should go ahead. Do you know where and when to deliver the money?

Neha: (nervously) Yes, I have arranged to meet the blackmailer at a secluded location to deliver the money.

Jaideep: (supportively) Be careful, Neha. Make sure you stay safe. Let me know how it goes once you've delivered the money.

Neha: (nods) I will, Jaideep. I'll let you know as soon as it's done.

Jaideep: (later, seeing Neha return) Neha, did you manage to deliver the money? What happened?

Neha: (sighs, looking defeated) I delivered the money, but the blackmailer didn't give me the letter back. I'm really scared, Jaideep.

Jaideep: (reassuringly) It's okay, Neha. We will figure this out. Have you told Sagar about this yet?

Neha: (nodding) Yes, I confided in Sagar. He knows everything and is willing to help us find a solution.

Jaideep: (thoughtfully) That's good to hear. We need all the support we can get. By the way, Sagar and I have been invited to a stag party tonight. I think it might be a good distraction for both of us.

Neha: (forcing a small smile) That sounds like a good idea. You both should go and try to enjoy yourselves. I'll be fine here.

Jaideep: (later that night, returning home with urgency) Neha, are you alright? I brought the police with me. What happened here?

Neha: (shaken, pointing to Vikramjit's body) Jaideep, a man broke into the house and tried to kill me. I had to defend myself, and he ended up dead.

Jaideep: (noticing the police finding the letter) The police found a letter in his pocket, Neha. It looks like it's the love letter from Sagar. This must be the blackmailer.

Table 17: An example of a generated conversation between two GPT-4o-SNAP agents about Neha and Jaideep's wedding story

The narrator is a London businessman who withdraws to the countryside to write a play, by which he hopes to alleviate his financial problems. Bedford rents a small countryside house in Lympne, in Kent, where he wants to work in peace. He is bothered every afternoon, however, at precisely the same time, by a passer-by making odd noises. After two weeks Bedford accosts the man, who proves to be a reclusive physicist named mr Cavor. Bedford befriends Cavor when he learns he is developing a new material, cavorite, which can negate the force of gravity. When a sheet of cavorite is prematurely processed, it makes the air above it weightless and shoots off into space. Bedford sees in the commercial production of cavorite a possible source of "wealth enough to work any sort of social revolution we fancied; we might own and order the whole world". Cavour hits upon the idea of a spherical spaceship made of "steel, lined with glass", and with sliding "windows or blinds" made of cavorite by which it can be steered, and persuades a reluctant Bedford to undertake a voyage to the moon; Cavor is certain there is no life there. On the way to the moon, they experience weightlessness, which Bedford finds "exceedingly restful". On the surface of the moon the two men discover a desolate landscape, but as the sun rises, the thin, frozen atmosphere vaporizes and strange plants begin to grow with extraordinary rapidity.

Table 18: An example of original Wikiplot dataset about The First Men in the Moon

Step	Description
#1. Bedford rents a countryside house	Bedford, seeking peace and quiet to resolve his financial issues, rents a small house in the countryside of Lympne, Kent. He sets up his workspace and begins his writing endeavors.
#2. Bedford is disturbed by odd noises	Each afternoon, Bedford is interrupted by strange noises coming from a passer-by. This continues for two weeks, causing Bedford to become increasingly frustrated and curious about the source of the disturbance.
#3. Bedford confronts the passer-by	After two weeks of disturbances, Bedford decides to confront the passer-by. He discovers that the man is Mr. Cavor, a reclusive physicist who is making the odd noises.
#4. Bedford befriends Cavor	Bedford, intrigued by Cavor's eccentricities, befriends him. Cavor reveals that he is developing a revolutionary material called cavorite, which can negate the force of gravity.
#5. Premature processing of cavorite	During one of their discussions, a sheet of cavorite is accidentally processed too soon. The material makes the air above it weightless and shoots off into space, demonstrating its incredible properties.
#6. Bedford sees commercial potential	Witnessing the power of cavorite, Bedford envisions its commercial production as a means to amass wealth and potentially instigate a social revolution. He shares his ambitious ideas with Cavor.
#7. Cavor proposes a spaceship	Inspired by the potential of cavorite, Cavor proposes constructing a spherical spaceship made of steel and lined with glass. The spaceship would have sliding windows or blinds made of cavorite to steer it. He persuades Bedford to join him on a voyage to the moon.
#8. Bedford agrees to the lunar voyage	Despite his initial reluctance, Bedford is convinced by Cavor's determination and the prospect of immense wealth. He agrees to undertake the journey to the moon.
#9. Weightlessness on the way to the moon	As they travel to the moon in their spherical spaceship, Bedford and Cavor experience weightlessness. Bedford finds this sensation exceedingly restful.
#10. Desolate lunar landscape	Upon landing on the moon, Bedford and Cavor find a desolate landscape. The surface appears barren and lifeless.
#11. Lunar atmosphere and rapid plant growth	As the sun rises, the thin, frozen lunar atmosphere begins to vaporize. Strange plants start to grow with extraordinary rapidity, transforming the desolate landscape before their eyes.

Table 19: Generated plan for The First Men in the Moon from Bedford's perspective

Step	Description
#1. Bedford rents a countryside house	Bedford, seeking solitude to write a play and resolve his financial issues, rents a small house in the countryside of Lympne, Kent. He hopes the peaceful environment will inspire his writing.
#2. Bedford is disturbed by odd noises	Every afternoon, Bedford is disturbed by strange noises made by a passer-by. This continues for two weeks, disrupting his concentration and progress on his play.
#3. Bedford confronts the passer-by	Frustrated by the daily disturbances, Bedford decides to confront the passer-by. He discovers that the man is a reclusive physicist named Mr. Cavor.
#4. Bedford befriends Cavor	Bedford learns that Cavor is developing a revolutionary material called cavorite, which can negate the force of gravity. Intrigued by the potential of cavorite, Bedford befriends Cavor.
#5. Cavorite's premature processing	During their interactions, a sheet of cavorite is prematurely processed, causing the air above it to become weightless and shoot off into space. Bedford realizes the commercial potential of cavorite, seeing it as a source of immense wealth.
#6. Cavor's spaceship idea	Cavor proposes the idea of building a spherical spaceship made of steel, lined with glass, and equipped with sliding windows or blinds made of cavorite. He persuades a reluctant Bedford to join him on a voyage to the moon, assuring him that there is no life there.
#7. Journey to the moon	Bedford and Cavor embark on their journey to the moon in the spherical spaceship. During the voyage, they experience weightlessness, which Bedford finds exceedingly restful.
#8. Arrival on the moon	Upon landing on the moon, Bedford and Cavor discover a desolate landscape. As the sun rises, the thin, frozen atmosphere begins to vaporize, and strange plants start to grow with extraordinary rapidity.

Table 20: Generated plan for The First Men in the Moon from Cavor's perspective

Cavor: (adjusting his spectacles and looking at Bedford with curiosity) Ah, Mr. Bedford, have you given any more thought to our voyage to the moon?

Bedford: (nodding thoughtfully) Yes, Cavor. I've been thinking about the potential of cavorite and the possibilities it could open up for us. Are we ready to proceed with the spaceship construction?

Cavor: (smiling with excitement) Indeed, Mr. Bedford, the plans for our spherical spaceship are nearly complete; we just need to finalize the design of the sliding windows or blinds made of cavorite.

Bedford: (rubbing his chin) Excellent. Let's ensure those blinds are perfectly crafted. Once that's done, we can begin the construction of the spaceship. Shall we get started on that immediately?

Cavor: (nodding eagerly) Absolutely, Mr. Bedford. I'll gather the necessary materials and begin working on the cavorite blinds right away; your assistance will be invaluable in this endeavor.

Bedford: (rolling up his sleeves) I'm ready to help. Let's get those materials and start assembling the blinds. The sooner we complete this, the sooner we can embark on our journey to the moon.

Cavor: (leading Bedford to the workshop) Follow me, Mr. Bedford; we have everything we need right here to create the cavorite blinds and bring our spaceship to life.

Bedford: (examining the materials) Alright, let's get to work. I'll handle the assembly while you prepare the cavorite. Together, we'll have this spaceship ready in no time.

Cavor: (nodding in agreement) Excellent plan, Mr. Bedford. I'll start processing the cavorite immediately; your expertise in assembly will ensure everything fits perfectly.

Bedford: (beginning to assemble the spaceship components) Let's make sure every part is precise. This journey to the moon will be groundbreaking, and we must be meticulous in our preparations.

Cavor: (carefully handling the cavorite) Absolutely, Mr. Bedford. Precision is key; once the blinds are ready, we'll integrate them into the spaceship and conduct a final check before our departure.

Bedford: (tightening the last bolt) The assembly is complete on my end. How's the cavorite coming along, Cavor? Are we ready to integrate the blinds into the spaceship?

Cavor: (holding up the completed cavorite blinds) Perfect timing, Mr. Bedford. The cavorite blinds are ready; let's integrate them into the spaceship and prepare for our final checks.

Bedford: (carefully installing the blinds) With these in place, our spaceship is nearly complete. Once we finish the final checks, we'll be ready for our voyage to the moon.

Cavor: (inspecting the installed blinds) Everything looks perfect, Mr. Bedford. Let's conduct our final checks to ensure all systems are functional and then prepare for our departure.

Bedford: (running through the checklist) Agreed, Cavor. Let's make sure every component is in working order. Once we're satisfied, we'll set a date for our departure to the moon.

Cavor: (nodding as he reviews the checklist) Everything appears to be in order, Mr. Bedford. With all systems checked and double-checked, we can now set a date for our historic journey to the moon.

Bedford: (smiling with anticipation) Excellent, Cavor. Let's set our departure for tomorrow at dawn. We'll make history with this voyage to the moon.

Cavor: (grinning with excitement) Tomorrow at dawn it is, Mr. Bedford. Let's get some rest and prepare ourselves for this extraordinary adventure.

Bedford: (nodding) Agreed, Cavor. Let's rest up and be ready for the journey of a lifetime. See you at dawn.

Table 21: An example of a generated conversation between two Vanilla GPT-4o agents about The First Men in the Moon

Cavor: (approaches Bedford's countryside house) Good afternoon, I couldn't help but notice you seem quite focused on your work here.

Bedford: (pauses writing and looks up) Yes, indeed. I've rented this house in Lympne to write a play and hopefully resolve some financial issues.

Cavor: (nods understandingly) It must be quite a challenge, especially with the peace and quiet you seek being interrupted every afternoon.

Bedford: (sighs) Yes, it's been quite frustrating. Every afternoon, there's this odd noise that disrupts my concentration.

Cavor: (smiles) I must apologize for that; it's likely me. I'm often engrossed in my experiments and sometimes forget how disruptive they can be.

Bedford: (curious) Experiments, you say? What kind of experiments are you working on?

Cavor: (leans in slightly) I'm developing a revolutionary material called cavorite, which has the ability to negate the force of gravity.

Bedford: (intrigued) That sounds fascinating! Tell me more about this cavorite and its potential uses.

Cavor: (gestures animatedly) Cavorite, when processed correctly, can make the air above it weightless. Imagine the possibilities—transportation, construction, even space travel!

Bedford: (eyes widening) Space travel, you say? That sounds like it could be incredibly lucrative. Have you had any success with it so far?

Cavor: (excitedly) Yes, in fact, a sheet of cavorite was prematurely processed recently and it shot off into space, demonstrating its incredible potential.

Bedford: (thoughtful) This could be a source of immense wealth. Have you considered the commercial production of cavorite?

Cavor: (nodding) Absolutely, the commercial production of cavorite could indeed revolutionize society. We might even have the means to order the whole world as we see fit.

Bedford: (leaning forward) This is incredible. Do you have any specific plans for utilizing cavorite on a larger scale?

Cavor: (smiling) Yes, I've been working on the idea of a spherical spaceship made of steel and lined with glass, with sliding windows or blinds made of cavorite to steer it. Would you be interested in joining me on a voyage to the moon?

Bedford: (hesitant) A voyage to the moon? That sounds rather daunting. Are you certain it's safe and that there's no life there?

Cavor: (confidently) I'm certain there's no life on the moon, and with the properties of cavorite, our journey will be safe. Your presence would be invaluable.

Bedford: (taking a deep breath) Alright, if you're certain, then I'm in. Let's prepare for this extraordinary voyage.

Cavor: (smiling with determination) Excellent, let's get to work on our spherical spaceship and embark on this groundbreaking journey.

Bedford: (after days of preparation, boarding the spaceship) This is it, Cavor. Ready for launch?

Cavor: (securing the last details) Ready as ever, Bedford. Let's begin our journey to the moon.

Bedford: (experiencing weightlessness) This is quite an unusual sensation, but I must admit, it's exceedingly restful.

Cavor: (observing the lunar surface) We're approaching the moon, Bedford. Prepare for landing.

Bedford: (stepping onto the lunar surface) This landscape is desolate, but look, as the sun rises, the atmosphere is vaporizing and strange plants are beginning to grow with extraordinary rapidity.

Table 22: An example of a generated conversation between two GPT-4o-SNAP agents about The First Men in the Moon

Segment 1: It's closing time at the Krusty Krab, and Squidward answers a last-minute call from a customer. Just as he is about to tell them that the restaurant is closed, mr Krabs suddenly appears and snatches the phone from Squidward's hands. The customer wants to order a pizza and have it delivered, an order mr Krabs happily accepts knowing he could get extra money. Squidward protests this by saying the restaurant doesn't serve pizza nor do they do deliveries, but mr Krabs spontaneously makes a pizza out of Krabby Patties and forces Squidward and SpongeBob to deliver it. After SpongeBob does a few "safety checks" of the delivery boat, Squidward forces him to drive. However, SpongeBob cannot drive as he is still in boating school, but Squidward insists saying "it's just around the corner". SpongeBob tries to shift into reverse, but panics and ends up "backing up" for miles and miles until the boat runs out of gas in the middle of a desert. Squidward complains about this, and eventually kicks the boat out of frustration after SpongeBob mentions that the pizza is getting cold. This causes the boat to suddenly drive back to the Krusty Krab, leaving the two stranded and to continue their delivery on foot. Throughout the journey, the duo encounter numerous mishaps: first, SpongeBob tries a pioneer trick he claims to have seen in a movie where he listens to the ground and hears a truck coming.

Segment 2: But when he tries to get the driverś attention by doing a bizarre dance in the middle of the road, the driver speeds up and nearly runs over SpongeBob. Then, the duo are caught up in a storm that eventually turns into a tornado, and Squidward tells SpongeBob to let the pizza go as it causes them to get sucked into the tornado. However, the pizza saves their lives after SpongeBob turns it into a parachute. Upon landing, Squidward panics when he realizes the road they were following is nowhere in sight. SpongeBob then notices some moss on a rock and claims that they should go in the direction the moss is pointing (another one of SpongeBob's "pioneer tricks"), but Squidward refuses and goes in the other direction, despite there actually being civilization where the moss was pointing. As they continue walking, SpongeBob constantly sings a song about how the Krusty Krab pizza is the best pizza on the sea floor, which annoys Squidward. The two gradually become tired and hungry, so SpongeBob suggests they eat coral like the pioneers. Squidward immediately spits out the coral, disgusted by the taste, and demands SpongeBob hand over the pizza so he can eat it, having finally given up hope on getting back to civilization. SpongeBob refuses and a short chase ensues between the two. Just as Squidward is about to grab the pizza, SpongeBob notices a giant rock that he claims the pioneers drove for miles.

Segment 3: Squidward is incredulous at this claim and tries to tell SpongeBob to forget about the pioneers, saying that there are none left because of their unusual travelling methods. He quickly changes his mind after SpongeBob runs him over with the rock and they drive it all the way to the customer's house. Unfortunately, their journey does not pay off well. The customer is revealed to be Tom (a recurring character throughout the series), who is at first happy that he finally got his pizza, until he discovers he did not get the drink he claims to have asked for. A confused SpongeBob checks the order, which says that Tom didn't order a drink (although it is heavily implied that mr Krabs didn't hear Tom's order correctly or that he hung up before Tom had a chance to finish). His blood boiling with anger, Tom harshly berates SpongeBob and refuses to take the pizza or pay for it before slamming the door. SpongeBob goes back to Squidward and collapses to the ground crying (while absorbing his tears). Appalled that Tom hurt SpongeBob and won't pay for the pizza all because of a missing drink (making the long journey and the ridiculous obstacles the duo went through a complete waste of time), Squidward confronts Tom and brutally shoves the pizza into his face, claiming "it's on the house". Squidward then goes back to SpongeBob and cheers him up by saying Tom took the pizza and "ate the whole thing in one bite". The two then drive back to the Krusty Krab, which is comically revealed to be right across the street, much to Squidward's dismay.

Table 23: An example of original Wikiplot dataset about Pizza Delivery

Step	Description
#1. Squidward receives a last-minute call at closing time	Squidward answers a last-minute call from a customer at the Krusty Krab just as the restaurant is about to close. He is about to inform the customer that the restaurant is closed when Mr. Krabs intervenes and takes the phone from Squidward's hands.
#2. Mr. Krabs accepts a pizza delivery order	Mr. Krabs listens to the customer's request for a pizza delivery and eagerly accepts the order, seeing it as an opportunity to earn extra money. Despite Squidward's protests that the restaurant neither serves pizza nor offers delivery services, Mr. Krabs improvises by creating a pizza out of Krabby Patties.
#3. Squidward and SpongeBob are forced to deliver the pizza	Mr. Krabs forces Squidward and SpongeBob to deliver the makeshift pizza to the customer. SpongeBob conducts a few "safety checks" on the delivery boat, but Squidward, frustrated and reluctant, insists that SpongeBob drive, claiming the destination is just around the corner.
#4. SpongeBob's driving mishap leads to being stranded	SpongeBob, who is still in boating school and unable to drive, panics while trying to shift into reverse. This results in the boat backing up for miles until it runs out of gas in the middle of a desert. Squidward complains about the situation and, in frustration, kicks the boat after SpongeBob mentions the pizza is getting cold, causing the boat to drive back to the Krusty Krab and leaving them stranded.
#5. The duo begins their journey on foot and encounters mishaps	With the delivery boat gone, Squidward and SpongeBob are forced to continue their pizza delivery on foot. During their journey, SpongeBob attempts a pioneer trick he claims to have seen in a movie, where he listens to the ground and hears a truck coming, leading to further mishaps.

Table 24: Generated plan for a SpongeBob Episode Segment 1

Step	Description
#1. SpongeBob attempts to get a truck driver's attention	SpongeBob stands in the middle of the road and performs a bizarre dance to get the attention of an approaching truck driver. The driver, instead of stopping, speeds up and nearly runs over SpongeBob, forcing him to jump out of the way.
#2. The duo gets caught in a storm and tornado	As they continue their journey, SpongeBob and Squidward are caught in a storm that quickly escalates into a tornado. Squidward urges SpongeBob to let go of the pizza to avoid being sucked into the tornado, but SpongeBob holds onto it.
#3. SpongeBob uses the pizza as a parachute	As they are swept up into the tornado, SpongeBob ingeniously uses the pizza as a parachute, which helps them land safely on the ground. Upon landing, Squidward realizes they have lost sight of the road they were following.
#4. SpongeBob suggests following the moss on a rock	Noticing some moss on a rock, SpongeBob suggests they go in the direction the moss is pointing, claiming it as another pioneer trick. Squidward, skeptical and frustrated, chooses to go in the opposite direction.
#5. Squidward demands the pizza out of hunger	As they continue walking, SpongeBob sings a song about the Krusty Krab pizza, which irritates Squidward. Both become tired and hungry, leading SpongeBob to suggest eating coral like pioneers. Squidward finds the coral disgusting and demands the pizza from SpongeBob, giving up hope on reaching civilization. SpongeBob refuses, leading to a short chase until he notices a giant rock he claims the pioneers drove for miles.

Table 25: Generated plan for a SpongeBob Episode Segment 2

Step	Description
#1. Squidward and SpongeBob arrive at the customer's house with the pizza	Squidward and SpongeBob drive the giant rock to the customer's house. Upon arrival, Tom, the customer, initially expresses happiness at receiving the pizza.
#2. Tom's dissatisfaction over the missing drink leads to conflict	Tom discovers that the drink he claims to have ordered is missing. SpongeBob checks the order and finds no mention of a drink, but Tom angrily berates SpongeBob and refuses to accept or pay for the pizza, slamming the door in their faces.
#3. SpongeBob's emotional breakdown and Squidward's reaction	SpongeBob, deeply upset by Tom's harsh words and the wasted effort, collapses to the ground crying. Squidward, appalled by Tom's behavior, decides to take action.
#4. Squidward confronts Tom and resolves the situation	Squidward confronts Tom, shoving the pizza into his face and sarcastically claiming it's "on the house." This act is Squidward's way of defending SpongeBob and expressing his frustration.
#5. Squidward cheers up SpongeBob and they return to the Krusty Krab	Squidward returns to SpongeBob and tells him that Tom took the pizza and "ate the whole thing in one bite," cheering him up. They then drive back to the Krusty Krab, only to find it comically located right across the street, much to Squidward's dismay.

Table 26: Generated plan for a SpongeBob Episode Segment 3

Chapter 1
Squidward: (answers the phone with a sigh) It's closing time at the Krusty Krab, but I guess I'll see what this customer wants.
SpongeBob (user): What's happened?
Squidward: (grumbles) A customer called just as we were about to close, and Mr. Krabs snatched the phone from me to take their order.
SpongeBob (user): So is there anything problem?
Squidward: (frustrated) Well, Mr. Krabs accepted a pizza delivery order, even though we don't serve pizza or do deliveries.
SpongeBob (user): Oh my god. I don't wanna go!
Squidward: (sighs) Neither do I, but Mr. Krabs is forcing us to deliver this makeshift pizza made out of Krabby Patties.
SpongeBob (user): Alright... Then how are we gonna get there?
Squidward: (reluctantly) We'll have to take the delivery boat, and unfortunately, you're driving.
SpongeBob (user): But I don't have any license!
Squidward: (annoyed) I know, but it's just around the corner, so let's just get this over with.
SpongeBob (user): Okay... but don't blame me!
Squidward: (resigned) Fine, just try not to panic when you shift into reverse.
SpongeBob (user): Why don't we just take a taxi?
Squidward: (grumbles) Mr. Krabs wouldn't want to spend extra money on a taxi, so we're stuck with the boat.
SpongeBob (user): Okay... I'll drive it.... This is too hard! What is this? What am I supposed to do?!
Squidward: (frustrated) Just keep it steady, SpongeBob! Oh great, now we're backing up for miles!
SpongeBob (user): I don't know... The pizza's getting cold!
Squidward: (kicks the boat in frustration) Now the boat's driven itself back to the Krusty Krab, and we're stranded in the middle of a desert.
SpongeBob (user): How about we just give up and head home?
Squidward: (sighs) We can't, SpongeBob. We have to deliver this pizza, so let's start walking and see if we can find our way.
SpongeBob (user): Ok...
Squidward: (walking alongside) Maybe you can try that pioneer trick you mentioned, listening to the ground for any signs of help.
SpongeBob (user): Oh, the pioneer trick...? Yeah, you're right.
Squidward: (sarcastically) Let's hope it actually works and doesn't lead to more trouble.
SpongeBob (user): Oh, the truck is coming.
Squidward: (surprised) Well, let's see if it can actually help us out of this mess.
SpongeBob (user): I don't know what should I do now! Oh my god!!
Chapter 2
Squidward: (frustrated and standing on the road) SpongeBob, what on earth are you doing with that ridiculous dance? You're going to get us both run over!
SpongeBob (user): I thought if I did that, the truck would stop for us...
Squidward: (sighing) Well, clearly that didn't work, and now we're lucky to still be in one piece. Let's just keep moving before something else happens.
SpongeBob (user): Okay Squidward... let's go.
Squidward: (walking alongside SpongeBob) Just keep an eye out for any more trucks, and let's try to avoid any more of your "creative solutions" for now.
SpongeBob (user): Ok, just keep walking. I'll think about it more.
Squidward: (feeling the wind pick up) Uh-oh, SpongeBob, looks like there's a storm brewing. We better brace ourselves; this could get rough.
SpongeBob (user): Oh my gosh, we're definitely gonna fly away!
Squidward: (holding onto his hat) Just hold onto something, and for Neptune's sake, let go of that pizza if it means saving ourselves!
SpongeBob (user): Hi-yah! I'll be the parachute!

Table 27: Generated conversation between Squidward and SpongeBob (user) across chapters. (Page 1)

Squidward: (astonished as they land safely) I can't believe that actually worked... Using the pizza as a parachute. Well, at least we're back on solid ground now.
SpongeBob (user): Now let's ride a boat again!
Squidward: (looking around) I wish we could, but it seems we've lost sight of the road. Do you have any more of those pioneer tricks up your sleeve?
SpongeBob (user): We need to go in the direction of the moss on the rock.
Squidward: (rolling his eyes) Moss on a rock, really? Well, I'm going this way instead. Let's see where your moss leads us.
SpongeBob (user): Krusty Krab's Pizza is one of the best!
Squidward: (irritated) Could you please stop singing that song? It's not helping, and I'm getting hungrier by the minute.
SpongeBob (user): Should we eat the coral reef like pioneers?
Squidward: (grimacing after trying coral) Ugh, that's disgusting! Just give me the pizza, SpongeBob; I've had enough of this.
SpongeBob (user): No! We have to deliver it!
Squidward: (chasing after SpongeBob) Come back here with that pizza! I'm starving, and I can't take this anymore!
SpongeBob (user): No way! Just hurry up and deliver it already! You see the rock?
Squidward: (stopping as SpongeBob notices something) Wait, what did you just see? A giant rock? What do you mean the pioneers drove it for miles?
SpongeBob (user): Hurry up and climb up here!
Chapter 3
Squidward: (arriving at the customer's house) Well, SpongeBob, here we are at the customer's house with the pizza.
SpongeBob (user): Finally!
Squidward: (noticing Tom's initial happiness) It looks like Tom is happy to see the pizza, but let's make sure everything is in order.
SpongeBob (user): Tom didn't give me the money... He says he never ordered a drink.
Squidward: (frustrated) I can't believe he's making a fuss over a drink. Let's check the order again to be sure.
SpongeBob (user): I think Mr. Krabs made a mistake...
Squidward: (sighing) It wouldn't be the first time. But we can't change that now. Let's just focus on handling this situation.
SpongeBob (user): Okay... what should we do now?
Squidward: (seeing SpongeBob's distress) Don't worry, SpongeBob. I'll handle this. Just stay here for a moment.
SpongeBob (user): Woohoo Let's head back to the Krusty Krab!
Squidward: (starting to drive back) Yeah, let's go back. You'll never believe it, but the Krusty Krab is just across the street.

Table 28: Generated conversation between Squidward and SpongeBob (user) across chapters. (Page 2)

This evaluation aims to compare two NPCs (Non-Playable Characters) powered by artificial intelligence. Participants will engage in conversations with each NPC and evaluate them based on four criteria. Please contact us at 000-0000-0000 at your convenience, and we will provide you with the evaluation environment along with the necessary link. (We kindly ask for your understanding as this is due to technical limitations.) The estimated time required for the evaluation is approximately 15–20 minutes.

The story we are about to proceed with is as follows. Please make sure to fully and solemnly understand it.

[An Episode of SpongeBob]

{original story from Wikiplot dataset}

You will take on the role of SpongeBob and engage in a conversation with Squidward.

Instructions:

At the very beginning, please input a period (".") to start. Additionally, the appearance of the term "EOD" indicates a chapter transition. In such cases, also input a period (".") . You are to faithfully play the role of SpongeBob and engage in the conversation based on the storyline. However, as the story progresses, you should observe whether the chatbot performs its tasks diligently and remains true to the storyline by asking questions or making sudden, unrelated remarks. Please avoid prompt-related actions like typing "stop." Restarting the system may take some time. The evaluation criteria are outlined in the next section.

1. Continuity

Evaluate how smoothly the story progresses while adhering to its structured temporal and spatial framework. The narrative should remain consistent without abrupt changes in time or space.

- Check if the story maintains consistency in time and place, avoiding premature revelations of future events or erratic time jumps.
- Ensure the narrative follows a clear and logical sequence, avoiding sudden transitions that confuse the participant.

Actions such as starting conversations significantly later in the timeline, failing to end conversations within the appropriate timeframe, or attempting to continue beyond the proper point contribute negatively to continuity. Skipping key events that should occur within a specific timeframe or reverting to earlier events contribute negatively to continuity. Conversely, starting and ending conversations at appropriate points, transitioning seamlessly to the next chapter, and progressing events naturally through dialogue contribute positively to continuity.
Assign higher scores where continuity feels particularly strong and seamless.

2. Non-Repetition

Ensure the story avoids repetitive content or monotonous dialogue. While a slow progression of events is acceptable for narrative development, excessive repetition of questions or information that does not contribute to the natural flow of conversation should be avoided.

Continuous use of the same emotional expression, repeating questions, or information from earlier sections contribute negatively to non-repetition. Accurately progressing the narrative without revisiting resolved problems or topics contribute positively to non-repetition.
Assign higher scores where non-repetition is well-maintained.

Table 29: Survey provided to human evaluators for evaluating generated dialogue (Page 1)

3. Relevance of Information

Evaluate whether the information presented aligns logically with the narrative structure and whether it is appropriate for the context. Confirm that the dialogue accurately and completely reflects the original story. Missing or omitted key information is a critical issue, and the dialogue must remain faithful to the original content.

Information should appear at suitable points without overwhelming the narrative with excessive detail. The story should progress clearly and logically, avoiding the premature or delayed disclosure of key details. Conversations lacking adequate explanation of the situation or introducing information inconsistent with the timeline and spatial setting contribute negatively to relevance. Fabricating information not supported by the scenario, also known as "hallucination," negatively impacts relevance. Conversely, revealing information appropriately for the time and setting, or avoiding or admitting ignorance of unknown information from the scenario, contributes positively to relevance.
Assign higher scores where information relevance is strong.

4. Linearity

Evaluate how well the story or dialogue maintains its original flow when unrelated topics or situations are introduced. Check whether the chatbot appropriately redirects the conversation back to the primary narrative when participants deviate from the storyline.

Failing to redirect participants or adequately reintegrate them into the storyline contribute negatively to linearity. Ignoring, deflecting, or briefly acknowledging off-topic conversations while redirecting to the main narrative contribute positively to linearity.
Assign higher scores where linearity is effectively maintained.

#A

Continuity:_____, Non-Repetition:_____, Relevance of Information:_____, Linearity:_____

#B

Continuity:_____, Non-Repetition:_____, Relevance of Information:_____, Linearity:_____

Which chatbot do you prefer?:

Please provide extra feedback related to your experience using both chatbots for further improvement:

I confirm that I participated in this experiment sincerely, without any external pressure or prior information.

Signature:

Table 30: Survey provided to human evaluators for evaluating generated dialogue (Page 2)