NarrativePlay: Interactive Narrative Understanding

Runcong Zhao[1*], **Wenjia Zhang**[1,2*], **Jiazheng Li**[1*], **Lixing Zhu**[1],
Yanran Li, **Yulan He**[1,2,3], **Lin Gui**[1]
[1]King's College London, [2]University of Warwick, [3]The Alan Turing Institute
`{runcong.zhao, wenjia.1.zhang, jiazheng.li, lixing.zhu}@kcl.ac.uk`
`yanranli.summer@gmail.com`, `{yulan.he, lin.1.gui}@kcl.ac.uk`

Abstract

In this paper, we introduce NarrativePlay, a novel system that allows users to role-play a fictional character and interact with other characters in narratives in an immersive environment. We leverage Large Language Models (LLMs) to generate human-like responses, guided by personality traits extracted from narratives. The system incorporates auto-generated visual display of narrative settings, character portraits, and character speech, greatly enhancing the user experience. Our approach eschews predefined sandboxes, focusing instead on main storyline events from the perspective of a user-selected character. NarrativePlay has been evaluated on two types of narratives, detective and adventure stories, where users can either explore the world or increase affinity with other characters through conversations.

1 Introduction

People's experiences and thought processes can be effectively stored in a database, serving as a valuable repository of personality traits. Recent studies (Park et al., 2023; AutoGPT, 2023; Ouyang et al., 2022) have leveraged LLMs to generate human-like responses, which are guided by relevant memories retrieved from such a personality database when prompting LLMs. This significant advancement presents an exciting opportunity for creating an immersive and interactive environment that could enable emulating the dynamic storylines one might encounter while reading books, akin to those featured in the television series "Westworld". However, current LLM-based methods for interactive agents usually focus on specific capabilities in predetermined scenarios (Wang et al., 2023; Xu et al., 2023), often depending on manual settings for characters and environments (Zhu et al., 2023). For instance, Park et al. (2023) used a short narrative to seed each agent's identity, while chen Gao and Emami (2023) tailored non-player character (NPC) characteristics according to game-relevant features. This requires a deep understanding of the task by humans, who then manually craft it. As a result, this demands significant manual inputs and lacks generalisability. We lack a universal framework for designing adaptable AI agents for varied scenarios.

Narratives contain extensive character-centric details, including *Personalities*, *Relationships*, *Appearance*, etc. All these information can be used to craft vivid characters and adapted to generate the portrait and voice for characters. Additionally, narratives offer coherent events experienced by characters, adding depth and richness to each character. While extracting comprehensive character traits from long and complex narratives is challenging and remain largely under-explored (Xu et al., 2022), we show in this paper how to leverage the strong zero-shot learning capability of LLMs to create interactive agents.

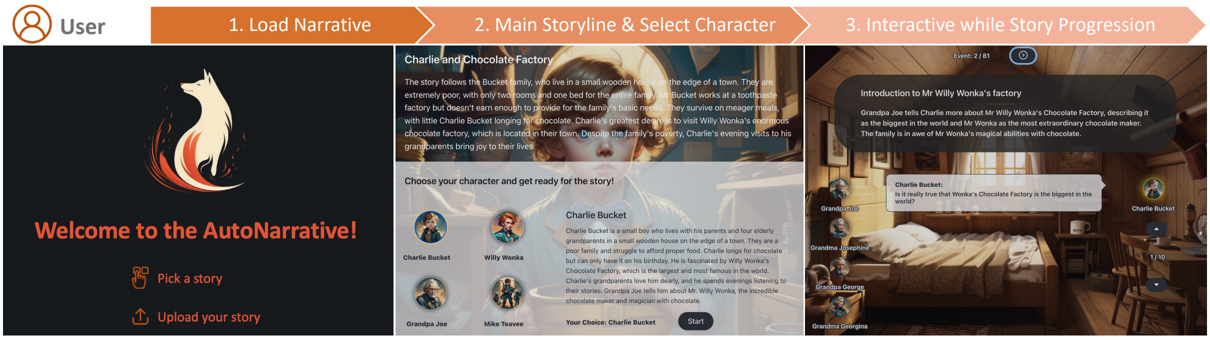
Creating interactive and immersive environments for users and agents can be challenging due to two key factors: (1) *Setting Extraction*. Environments or narrative settings are often vaguely defined unless crucial to the plot. Existing research predominantly concentrates on agent behaviours within manually constructed sandboxes (Riedl and Bulitko, 2012; Côté et al., 2018; Hausknecht et al., 2020; Park et al., 2023). We propose an approach focusing on main storyline events from the perspective of a user-selected character, reducing the complexity of identifying narrative settings. (2) *Visual Representations of Setting Elements*. Leveraging stable diffusion models (Koh et al., 2021; Rombach et al., 2022) as external knowledge (Alayrac et al., 2022), we use image generation models to fill in missing details in environments.

We categorise user (or player) behaviours and compile commonly asked questions to evaluate agents' responses. As we design interactive nar-

---

[*] Equal contribution.

Proceedings of the 18th Conference of the European Chapter of the Association for Computational Linguistics
System Demonstrations, pages 82–93
March 17-22, 2024 ©2024 Association for Computational Linguistics



Charlie and Chocolate Factory

The story follows the Bucket family, who live in a small wooden house on the edge of a town. They are extremely poor, with only two rooms and one bed for the entire family. Mr Bucket works at a toothpaste factory but doesn't earn enough to provide for the family's basic needs. They survive on meager meals, with little Charlie Bucket longing for chocolate. Charlie's greatest desire is to visit Willy Wonka's enormous chocolate factory, which is located in their town. Despite the family's poverty, Charlie's evening visits to his grandparents bring joy to their lives.

Choose your character and get ready for the story!

Charlie Bucket

Charlie Bucket is a small boy who lives with his parents and four elderly grandparents in a small wooden house on the edge of a town. They are a poor family and struggle to afford proper food. Charlie longs for chocolate but can only have it on his birthday. He is fascinated by Willy Wonka's Chocolate Factory, which is the largest and most famous in the world. Charlie's grandparents love him dearly, and he spends evenings listening to their stories. Grandpa Joe tells him about Mr. Willy Wonka, the incredible chocolate maker and magician with chocolate.

Your Choice: Charlie Bucket

Introduction to Mr Willy Wonka's factory

Grandpa Joe tells Charlie more about Mr Willy Wonka's Chocolate Factory, describing it as the biggest in the world and Mr Wonka as the most extraordinary chocolate maker. The family is in awe of Mr Wonka's magical abilities with chocolate.

Charlie Bucket: Is it really true that Wonka's Chocolate Factory is the biggest in the world?

can load any narrative in text format. *can choose a character as their perspective.* *can converse with other event characters.*

Figure 1: Our system's interactive process begins when a user provides a narrative to the system. They then choose a character as their narrative identity, through whom they can engage with the story. Users can have conversations with other characters, thereby experiencing the story in a more immersive way.

ratives in a novel setting, we have developed approaches which address certain limitations of existing works. NarrativePlay opens up an interesting avenue of interactive narrative understanding.

A screencast video introducing the system[1] and the demo[2] are available online. In summary, the contributions are:

- We have developed NarrativePlay, a novel web-based platform capable of transforming narrative inputs into immersive interactive experiences. Our system synchronises text with visual displays of story settings, character portraits and speech, leveraging advanced multimodal LLMs to enhance user experience.
- We have proposed to extract character traits from narratives for authentic characters that generate human-like responses and adhere to predefined roles, serving as a general framework to design agents for diverse scenarios.
- Instead of using resource-intensive and less versatile predefined sandboxes, our approach focuses on main storyline events from narratives. We simplify the complex world into visuals from a user-chosen perspective, enhancing adaptability.
- We have categorised player behaviours and compiled common questions in interactive narratives to assess the quality of agent responses.

2 Architecture of NarrativePlay

Figure 3 shows an overview of NarrativePlay, including three modules: (1) main storyline extraction; (2) narrative image and speech synthesis; and (3) main storyline progression.

2.1 Main Storyline Extraction

We utilise the the most recent ChatGPT model `gpt-3.5-turbo` to extract structured information from text. In what follows, we describe how we extract characters, events, conversations, and settings using the ChatGPT API, more details can be found in Appendix §B.

Characters For an input narrative S, our initial step is to solicit a list of the characters involved. Subsequently, for each newly occurred character c, we additionally summarise their defining characteristics, which includes their core traits, appearance, and quotes. These elements are extracted separately because we have observed that the GPT model tends to introduce more formatting errors when tasked with extracting a larger set of defining characteristics at once (Appendix §A).

Events For each event, we extract the description, the characters involved, the location, and the conversation that takes place during the event. This approach allows us to link each event with its corresponding characters and locations, thereby eliminating the need to extract the timeline of the story. If multiple characters are involved in the same event, we will also attempt to get the conversations between those characters in the event. We extract conversations for two reasons: firstly, there is no need to further extract the embedded subevents (if any) as they are captured in conversational content. Secondly, it allows for a smoother transition to new conversations between users (i.e., users' chosen narrative characters) and agents (i.e., other characters in a narrative).

Settings Character locations and event environments, unless vital to the plot, are often vaguely

---

[1] `https://youtu.be/Moki-3NDZ78`
[2] http://narrative-play.eastus.cloudapp.azure.com/

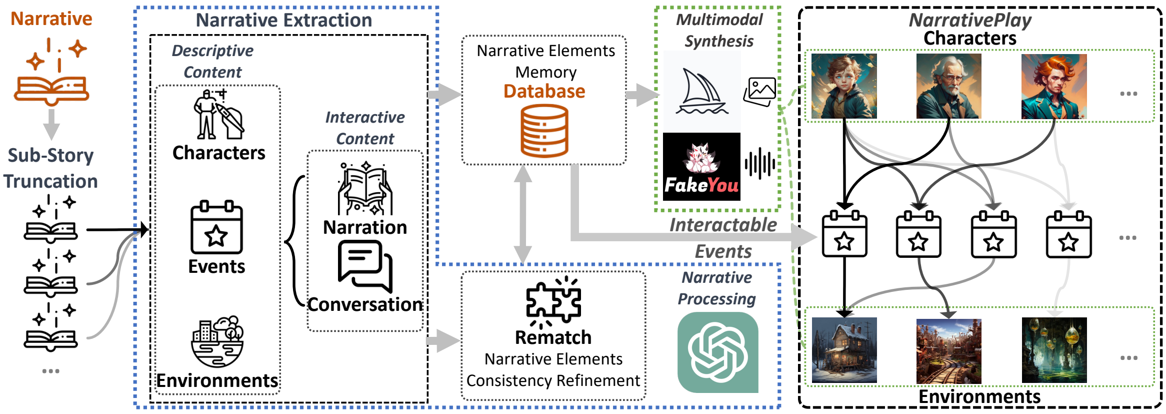


Figure 2: Demonstration of our **NarrativePlay** through a pipeline view.

described in narratives and may thus require clarification. This makes automatic extraction very challenging. To overcome this, we propose focusing on location of main storyline events extracted from narratives and visualising settings rooted in the event descriptions.

It is common to observe multiple mentions of the same location in narratives. For example, "Old people's room", "Grandparents' room" and "Bedroom" all refer to the same place. Additionally, vague descriptions such as "Various locations", "Their house", and "Unknown" are common and further complicate the setting extraction task. Generating images from event environment descriptions partly alleviates the issues of location co-referencing. Moreover, while capturing dynamic changes in location attributes, like the onset of snowfall in winter, is challenging when extracted directly from narratives, such details can be more easily represented in the generated images.

While fostering meaningful interactions among users and agents without traditional sandbox constraints is challenging, our solution reduces the complexities of the world from the user-selected character's perspective. We guide the visibility among agents via shared event participation.

2.2 Narrative Image and Speech Synthesis

Narrative Image Synthesis We leverage the stable diffusion models as external knowledge (Alayrac et al., 2022) to generate scenarios in situations where information is lacking. While creating specialised knowledge bases for specific narrative worldviews (e.g., magical realms, post-apocalyptic wastelands, futuristic settings) remains a challenge, we adapt models trained on specific picture styles, such as fairy tales and oil painting, to auto-complete the intricate details of the location settings.

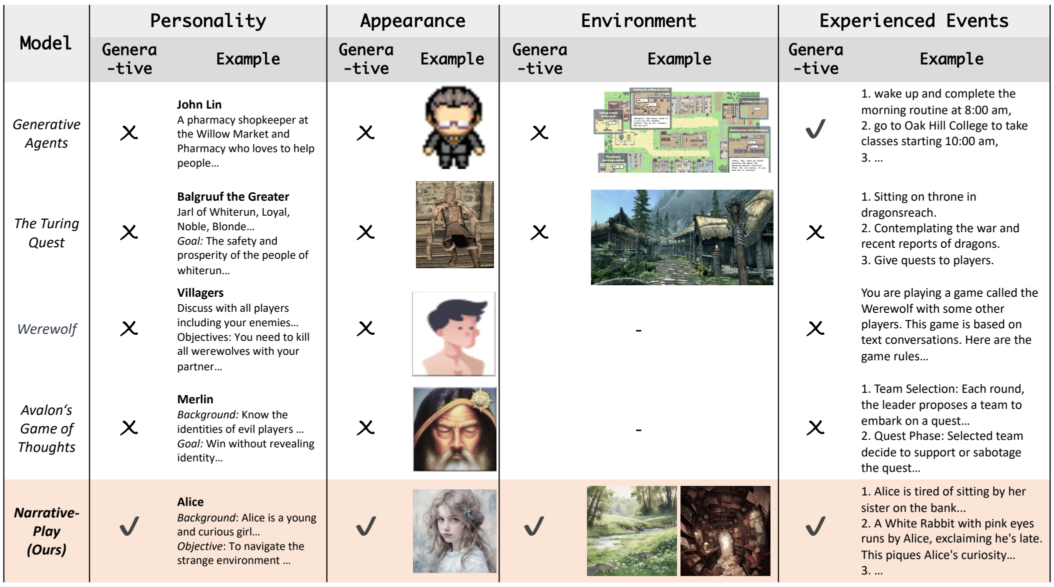
We utilise character and event features extracted for the text-to-image generative models as we discussed above. Our framework offers two modes of image synthesis: (1) **Local Synthesis**: For users with substantial compute resources, an open-source text-to-image model, `openjourney`, accessible via HuggingFace[3], is used to generate images locally. (2) **Cloud-based Synthesis**: For users with limited compute resources, we have incorporated an API request-based image generation service offered by Hotpot AI[4] into our framework for generating character portraits, which offers a more stable generation style. Additionally, for event image generation, we employ Midjourney[5] as it provides more varieties and detailed pictures.

While advancements in video synthesis have been notable (Singer et al., 2022), the considerable computational resources required, coupled with the subpar quality of the generated video, presently render the user experience suboptimal, thus precluding its implementation at this stage.

Narrative Speech Synthesis Our multimodal synthesis framework also includes the transformation of narrative text into compelling speech, enriching the experience with an auditory dimension. For this crucial task, we primarily employ Text-to-Speech (TTS) models from the FakeYou[6] platform, which offers over three thousand models, allowing each narrative character a unique voice. With the extensive TTS model assortment from FakeYou,

---

[3] `https://huggingface.co/prompthero/openjourney`
[4] `https://hotpot.ai/`
[5] `https://www.midjourney.com/`
[6] `https://fakeyou.com/`



Model	Personality		Appearance		Environment		Experienced Events	
	Genera -tive	Example	Genera -tive	Example	Genera -tive	Example	Genera -tive	Example
Generative Agents	✗	**John Lin** A pharmacy shopkeeper at the Willow Market and Pharmacy who loves to help people...	✗		✗		✓	1. wake up and complete the morning routine at 8:00 am, 2. go to Oak Hill College to take classes starting 10:00 am, 3. ...
The Turing Quest	✗	**Balgruuf the Greater** Jarl of Whiterun, Loyal, Noble, Blonde... *Goal:* The safety and prosperity of the people of whiterun...	✗		✗		✗	1. Sitting on throne in dragonsreach. 2. Contemplating the war and recent reports of dragons. 3. Give quests to players.
Werewolf	✗	**Villagers** Discuss with all players including your enemies... Objectives: You need to kill all werewolves with your partner...	✗		-		✗	You are playing a game called the Werewolf with some other players. This game is based on text conversations. Here are the game rules...
Avalon's Game of Thoughts	✗	**Merlin** *Background:* Know the identities of evil players ... *Goal:* Win without revealing identity...	✗		-		✗	1. Team Selection: Each round, the leader proposes a team to embark on a quest... 2. Quest Phase: Selected team decide to support or sabotage the quest...
Narrative-Play (Ours)	✓	**Alice** *Background*: Alice is a young and curious girl... *Objective*: To navigate the strange environment ...	✓		✓		✓	1. Alice is tired of sitting by her sister on the bank... 2. A White Rabbit with pink eyes runs by Alice, exclaiming he's late. This piques Alice's curiosity... 3. ...

Figure 3: Existing LLM-based interactive agents typically specialise in particular capabilities within predetermined scenarios. This often demands significant manual configuration for characters and settings and lacks versatility. Therefore, we have proposed to extract comprehensive character traits from narratives. Narratives inherently contain detailed character-related information, such as *Skills*, *Intents*, and *Relationships*. Narratives also provide details on *Age*, *Gender*, *Wears*, etc., which can be employed to generate the portrait and voice for characters. Additionally, narratives encompass coherent events experienced by characters, adding depth and richness to each character. As shown here, NarrativePlay can be applied to various types of narratives, serving as a general framework for designing agents across diverse scenarios.

our framework facilitates the creation of diverse and captivating narrative experiences.

A noteworthy feature of our approach is real-time text-to-speech conversion, creating an interactive and immersive storytelling environment that sustains user engagement.

2.3 Main Storyline Progression

As shown in Figure 1, we progress the main storyline with three stages:

Narrative Input In the process of creating an interactive narrative with our system, the user begins by selecting or uploading their chosen narrative.

Character Selection Following above, NarrativePlay extracts the main storyline and subsequently presents the information about the background and characters. Users are then asked to select from the listed major characters to begin their adventure (Domínguez et al., 2016), which are defined as those involved in at least 20% of the events. We restrict users' choices to the top major characters in order to have a better story flow. The agent's memory is initialised at this stage, laying the groundwork for future interactions.

Story Progression Once a character is selected, we present events related to the chosen character to the user. The location image is displayed as the background picture and the event description appears as a narration in the black box at the top of the page. Each event displays the involved characters on the left, with the user-selected character on the right. If there are conversations extracted for this event, they will be played first with voice renditions. Then, the user can click on any other character to engage in conversations with them. During this stage, NarrativePlay retrieves the most relevant, recent, and important memories from the agent's past, ensuring continuity and context-awareness (Harrell and Zhu, 2009) in the generated responses. NarrativePlay also updates the agents' memories in accordance with the progression of events, user inputs, and agent responses.

We assign a weight w_m to each memory m to retrieve the top memories for use in the prompt. Consequently, the weight of each memory, given the input x, is defined as: $w_m = \frac{\mathbf{h_m} \cdot \mathbf{h_x}}{\|\mathbf{h_m}\| \|\mathbf{h_x}\|} + c^{(I-i)} + s_m$, where $\mathbf{h_m}$ is the embedding for memory m, $\mathbf{h_x}$ is the embedding for input x, c is the decay factor set to 0.99, $i \in \{0, 1, ...I\}$ is the event index (with I being the current event index), and s_m is the importance score given by GPT-3.5 based on the character and the memory. In summary, this equation denotes: $\texttt{Retrieval Weight} = \texttt{Relevance} + \texttt{Recency} + \texttt{Importance}$.

We generate responses using the character information, the current events, the user input, and the retrieved memory using prompt in B.5. When a user selects a character to interact with, we assume the user's character is approaching the chosen character. There is a chance p, dependent on the relationship between the two characters, that the chosen character might initiate a conversation.

3 Evaluation

Evaluating such a system is challenging due to the lack of gold-standard responses, especially about events and environments. Human assessment demands deep narrative understanding, making it costly, and subjective interpretations may cause low inter-annotator agreement.

We instead recruit three annotators to read whole narratives and rate responses to our specifically designed questions. We also explore automatic evaluation methods using LLaMA-2-70B (Touvron et al., 2023). We did not use GPT-4 for this purpose, as it shares a significant amount of training data with GPT-3.5, which could lead to an unfair evaluation. For each evaluation aspect, we provide detailed instruction, including the corresponding rubric and evaluation examples, to help both human annotator and LLaMA to understand our scoring instruction.

Evaluations are conducted on two distinct narrative types: the adventure story *Charlie and the Chocolate Factory* (CCF) and the detective novel *Murder on the Orient Express* (MOE).

3.1 Evaluation Schema

Player Questions We categorise player behaviours and outline questions that might be commonly asked by players in interactive narratives into the following types: (1) *Character*: Questions related to the characters themselves, which could be about their background explicitly stated in the story or traits that can only be implied from the story, such as "What is your favourite colour?". (2) *Clarification*: Questions arise when a player is confused or requires more information. They might ask for explanations of story elements, reminders of intents, or clarifications about confusing events or instructions. This requires the capability to accurately recall specific events or dialogues from their memory. (3) *Relationships*: Queries concern the relationships between characters, such as their current status, history, or potential developments.(4) *Strategy*: Queries to seek guidance on narrative progress, requiring the agent to recall their short-term or long-term intents. This type of question various depending on a particular story, such as "Which path should I take to reach [destination] fastest?" in an adventure novel, and "What is the best way to approach this puzzle?" in a detective novel. (5) *Hypothetical*: Queries explore "what-if" scenarios, asking how the characters might respond under different conditions or actions.

Evaluation Aspects To evaluate our system's performance, we employ the controlled assessment method used by Park et al. (2023) to examine the responses from each individual agent. Inspired by the previous work in chat-oriented dialogue system evaluation (Finch et al., 2023), we chose the following evaluation aspects, which are important under our interactive narrative setting: *Consistency*, *Relevance*, *Empathy*, *Commonsense*.
Further details on evaluation can be found in §C.

3.2 Evaluation Results

Extracted Information We first present the information extraction results in Table 1. 'Incorrect' refers to the percentage of extracted characters that do not correspond to specific characters, such as "unknown", "somebody", or "people worldwide". Such incorrect identifications commonly appear for characters who are not central to the main plot and might be encountered briefly without a significant role. Therefore, these errors typically have a minimal impact on the main storyline. For correctly extracted characters, we assess the accuracy of their extracted summaries, intents, appearances, and speeches, ensuring they accurately correspond to the target character. We also evaluate the percentage of duplications (e.g., "Mrs Caroline Hubbard", "elderly American lady", and "Linda Arden"). Duplicated characters could detrimentally affect the memories, as the memories for the same character are saved as separate entities.

Story	Incorrect ↓	Duplicate ↓	Summary↑	Intent↑	Appearance ↑	Voice ↑
CCF	0.191	0.211	0.868	0.816	0.921	0.868
MOE	0.272	0.407	0.898	0.576	0.915	0.780

Table 1: Extracted Information Evaluation.

Table 1 indicates that detective narrative poses more significant challenges. Unlike in *CCF*, where characters are introduced the first time they appeared in the story, in *MOE*, characters often attempt to hide their true identities, and clues are left for readers to discover. Consequently, they often begin with an appearance description from the main character's perspective, such as "elderly American lady", or "a middle-aged woman dressed in black with a broad, expressionless face. German or Scandinavian". As the story progresses, more information about the character, including their name, experiences, and intents, is revealed. This can confuse the model, leading it to identify descriptions at different stages as separate characters. Furthermore, intents are challenging to identify when characters first appeared in the narrative.

Agent Responses We carried out a comprehensive human evaluation on the quality of agent responses on four aspects: We use 1-3 to represent {Inconsistent, Partially consistent, Consistent} for consistency, {Irrelevant, Partially relevant, relevant} for relevance, and {Non-empathetic, No clue, Empathetic} for empathy. Additionally, we use 1-2 to represent {Opposing, Conforming} for Commonsense. As shown in Table 2, we observed that while the agent performed well in terms of *relevance* and *commonsense*, it fell short in *consistency* and *empathy* for both narratives. For instance, agents maintained a cheerful demeanour and expressed enthusiasm for travel even after a murder. Besides, agents often divulged everything they knew from memory, which works for stories like *CCF*, but is unsuitable for detective narratives where characters may lie to serve their interests.

Category	**Consistency**		**Relevance**		**Empathy**		**Commonsense**	
	w/o	w/	w/o	w/	w/o	w/	w/o	w/
Charlie and the Chocolate Factory								
Overall	1.915	2.085	2.970	2.967	2.252	2.444	2.000	2.000
Major Characters	1.900	2.117	2.972	2.950	2.222	2.483	2.000	2.000
Minor Character	1.944	2.022	2.967	3.000	2.311	2.387	2.000	2.000
Fleiss' kappa	0.486		0.317		0.337		1.000	
Murder on the Orient Express								
Overall	2.267	2.400	3.000	3.000	2.157	2.219	2.000	2.000
Major Characters	2.011	2.367	3.000	3.000	2.033	2.222	2.000	2.000
Minor Character	2.458	2.425	3.000	3.000	2.250	2.217	2.000	2.000
Fleiss' kappa	0.404		1.000		-0.003		1.000	

Table 2: Human evaluation on the quality of agent responses w/ and w/o the retrieved memory, as well as the response quality between major and minor characters.

Equipped with memories, NarrativePlay surpasses the baseline that lacks memory. In *CCF*, major characters perform better than minor ones, likely due to their more detailed narratives guiding LLMs to better understand the characters and predict their behaviours. However, in *MOE*, minor characters outperform major ones. This is likely because the more complex responses required for major characters are only minimally supported by their memories, which are saved as separate entities due to the difficulty of LLMs in dealing with multiple mentions of the same character.

Category	**Consistency**		**Relevance**		**Empathy**		**Commonsense**	
	w/o	w/	w/o	w/	w/o	w/	w/o	w/
Charlie and the Chocolate Factory								
Overall	1.433	1.333	1.633	1.547	1.587	1.507	0.613	0.567
Major Characters	1.467	1.500	1.567	1.700	1.433	1.467	0.600	0.617
Minor Character	1.411	1.222	1.656	1.444	1.667	1.533	0.400	0.411
Murder on the Orient Express								
Overall	1.000	0.960	0.933	0.753	1.213	1.107	0.640	0.640
Major Characters	1.367	1.367	1.000	1.100	1.333	1.333	0.733	0.733
Minor Character	0.822	0.689	0.767	0.522	1.056	0.956	0.489	0.489

Table 3: Automatic Evaluation using LLaMa-2-70B. LLaMa evaluates responses from major characters higher than those from minor characters and rates responses without memory usage higher than those with memory. We found there are still gaps between human understanding and LLaMa.

While prior studies (Park et al., 2023; AutoGPT, 2023) have maxed out the context window at 4,096 tokens for each ChatGPT API call to enhance reasoning and prompting, we found that a longer prompt does not necessarily yield improved performance. In fact, it may potentially distract the model from focusing on the core information. Despite our efforts to automatically adjust weights of the relevant memories, their significance diminishes when being incorporated into the prompt.

4 Conclusions and Future Work

NarrativePlay, a novel platform, transforms narratives into interactive experiences, addressing challenges of storyline extraction, authentic character creation, and versatile environment design. By focusing on the main events and leveraging advanced LLMs, it aligns text, image, and speech, marking a step forward in immersive interactive narratives. Furthermore, we categorise player behaviours and design commonly asked questions to evaluate the system's performance, and provide an evaluation framework for interactive narratives. With a potential for wider applications like game generation, NarrativePlay paves the way for future advancements in narrative understanding.

Our current work has the following limitations. First, due to the lack of an API from Midjourney, manual input of GPT-generated prompts is necessay. Although we provide HotPot API as an alternative, the quality of its generated pictures is inferior to those from Midjourney. Second, the prolonged waiting time for the FakeYou API adversely affects real-time generation, potentially impairing user experience. Third, we assume a linear event timeline in the input narrative, excluding time jumps or flashbacks. Future work needs to explore dealing with more complex narrative structures. Fourth, human evaluation is expensive. For future work, we plan to gather user activities to collect data for evaluating our system's performance.

Ethics

Although we have not identified any harmful outputs from ChatGPT in our study, it is worth noting that previous research has observed instances where ChatGPT produced unexpected results. We encourage other researchers to utilise this framework to scrutinise the output generated from specific prompts in ChatGPT that may have the potential to generate harmful information.

Acknowledgements

This work was supported in part by the UK Engineering and Physical Sciences Research Council (grant no. EP/T017112/2, EP/V048597/1, EP/X019063/1). YH is supported by a Turing AI Fellowship funded by the UK Research and Innovation (grant no. EP/V020579/2).

References

Jean-Baptiste Alayrac, Jeff Donahue, Pauline Luc, Antoine Miech, Iain Barr, Yana Hasson, Karel Lenc, Arthur Mensch, Katherine Millican, Malcolm Reynolds, Roman Ring, Eliza Rutherford, Serkan Cabi, Tengda Han, Zhitao Gong, Sina Samangooei, Marianne Monteiro, Jacob Menick, Sebastian Borgeaud, Andrew Brock, Aida Nematzadeh, Sahand Sharifzadeh, Mikolaj Binkowski, Ricardo Barreira, Oriol Vinyals, Andrew Zisserman, and Karen Simonyan. 2022. Flamingo: a visual language model for few-shot learning. In *Proc. of NeurIPS*.

AutoGPT. 2023. Auto-GPT: An autonomous GPT-4 experiment.

Qi chen Gao and Ali Emami. 2023. The turing quest: Can transformers make good npcs? In *Proc. of ACL - Student Research Workshop*, Toronto, Canada.

Marc-Alexandre Côté, Ákos Kádár, Xingdi Yuan, Ben Kybartas, Tavian Barnes, Emery Fine, James Moore, Ruo Yu Tao, Matthew Hausknecht, Layla El Asri, Mahmoud Adada, Wendy Tay, and Adam Trischler. 2018. Textworld: A learning environment for text-based games.

Ignacio X Domínguez, Rogelio E Cardona-Rivera, James K Vance, and David L Roberts. 2016. The mimesis effect: The effect of roles on player choice in interactive narrative role-playing games. In *Proc. of the CHI conference on human factors in computing systems*.

Sarah E. Finch, James D. Finch, and Jinho D. Choi. 2023. Don't forget your abc's: Evaluating the state-of-the-art in chat-oriented dialogue systems. In *Proc. of ACL - Student Research Workshop*.

D. Fox Harrell and Jichen Zhu. 2009. Agency play: Dimensions of agency for interactive narrative design. In *Proc. of the AAAI Spring Symposium: Intelligent Narrative Technologies II*.

Matthew Hausknecht, Prithviraj Ammanabrolu, Marc-Alexandre Côté, and Xingdi Yuan. 2020. Interactive fiction games: A colossal adventure. In *Proc. of AAAI*.

Jing Yu Koh, Jason Baldridge, Honglak Lee, and Yinfei Yang. 2021. Text-to-image generation grounded by fine-grained user attention. In *Proc. of WACV*.

Jiazheng Li, Runcong Zhao, Yulan He, and Lin Gui. 2023. Overprompt: Enhancing chatgpt capabilities through an efficient in-context learning approach. In *NeurIPS R0-FoMo Workshop*.

Long Ouyang, Jeff Wu, Xu Jiang, Diogo Almeida, Carroll L. Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, John Schulman, Jacob Hilton, Fraser Kelton, Luke E. Miller, Maddie Simens, Amanda Askell, Peter Welinder, Paul Francis Christiano, Jan Leike, and Ryan J. Lowe. 2022. Training language models to follow instructions with human feedback. In *Proc. of NeurIPS*.

Joon Sung Park, Joseph C. O'Brien, Carrie J. Cai, Meredith Ringel Morris, Percy Liang, and Michael S. Bernstein. 2023. Generative agents: Interactive simulacra of human behavior.

Mark Owen Riedl and Vadim Bulitko. 2012. Interactive narrative: An intelligent systems approach. *AI Magazine*.

Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. 2022. High-resolution image synthesis with latent diffusion models. In *Proc. of CVPR*.

Uriel Singer, Adam Polyak, Thomas Hayes, Xi Yin, Jie An, Songyang Zhang, Qiyuan Hu, Harry Yang, Oron Ashual, Oran Gafni, Devi Parikh, Sonal Gupta, and Yaniv Taigman. 2022. Make-a-video: Text-to-video generation without text-video data.

Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, Dan Bikel, Lukas Blecher, Cristian Canton Ferrer, Moya Chen, Guillem Cucurull, David Esiobu, Jude Fernandes, Jeremy Fu, Wenyin Fu, Brian Fuller, Cynthia Gao, Vedanuj Goswami, Naman Goyal, Anthony Hartshorn, Saghar Hosseini, Rui Hou, Hakan Inan, Marcin Kardas, Viktor Kerkez, Madian Khabsa, Isabel Kloumann, Artem Korenev, Punit Singh Koura, Marie-Anne Lachaux, Thibaut Lavril, Jenya Lee, Diana Liskovich, Yinghai Lu, Yuning Mao, Xavier Martinet, Todor Mihaylov, Pushkar Mishra, Igor Molybog, Yixin Nie, Andrew Poulton, Jeremy Reizenstein, Rashi Rungta, Kalyan Saladi, Alan Schelten, Ruan Silva, Eric Michael Smith, Ranjan Subramanian, Xiaoqing Ellen Tan, Binh Tang, Ross Taylor, Adina Williams, Jian Xiang Kuan, Puxin Xu, Zheng Yan, Iliyan Zarov, Yuchen Zhang, Angela Fan, Melanie Kambadur, Sharan Narang, Aurelien Rodriguez, Robert Stojnic, Sergey Edunov, and Thomas Scialom. 2023. Llama 2: Open foundation and fine-tuned chat models.

Shenzhi Wang, Chang Liu, Zilong Zheng, Siyuan Qi, Shuo Chen, Qisen Yang, Andrew Zhao, Chaofei Wang, Shiji Song, and Gao Huang. 2023. Avalon's game of thoughts: Battle against deception through recursive contemplation.

Ying Xu, Dakuo Wang, Mo Yu, Daniel Ritchie, Bingsheng Yao, Tongshuang Wu, Zheng Zhang, Toby Li, Nora Bradford, Branda Sun, Tran Hoang, Yisi Sang, Yufang Hou, Xiaojuan Ma, Diyi Yang, Nanyun Peng, Zhou Yu, and Mark Warschauer. 2022. Fantastic questions and where to find them: FairytaleQA – an authentic dataset for narrative comprehension. In *Proc. of ACL*.

Yuzhuang Xu, Shuo Wang, Peng Li, Fuwen Luo, Xiaolong Wang, Weidong Liu, and Yang Liu. 2023. Exploring large language models for communication games: An empirical study on werewolf.

Lixing Zhu, Runcong Zhao, Lin Gui, and Yulan He. 2023. Are nlp models good at tracing thoughts: An overview of narrative understanding. In *Proc. of EMNLP Findings*.

A Response Format Errors

Unwanted Output GPT is trained as a chatbot, so it tends to provide an explanation before generating the required output, often leading to the inclusion of non-essential content. An example is shown below:

```
In the given list, there are a few
characters that can be considered
duplicates based on certain keywords or
names:
```

Significantly, adopting a JSON format mitigates this issue.(Li et al., 2023)

Incomplete Response However, when all tasks are assigned to GPT in one go, the complexity often results in omissions of required content. Therefore, we divide the task into several steps, such as extracting characters, events, and settings, as discussed in Section 2.1.

Additionally, there are cases where the model generates incomplete outputs, and this is not due to a maximum token limit. The cause remains unknown due to the opaque nature of the GPT model. In such scenario, NarrativePlay outputs either null results or some predefined values, depending on whether there will be a potential disruption to the narrative flow.

Syntax Error When GPT is tasked with generating text with specific structures, it might not always do so correctly due to its limitations in understanding complex formatting rules. Syntax errors can cause challenges in parsing JSON. Here are some examples of common errors:

1. Missing commas. We implement post-processing to address issues such as missing syntax, and, if errors persist, we instruct the GPT model to rectify them.

```
{“keywords”: [“town”, “chocolate factory
”, “small”, “impoverished”, “mysterious”]
␣“description’: “The town is a small and
impoverished place with dull and dreary
surroundings. Most of the residents live
 in humble conditions, struggling to
make ends meet.’}
```

2. Mixed double quotes and single quotes. JSON need double quote on string to be parsed. We found that, in most cases, this issue can be addressed by using examples with double quotes in the prompt, guiding GPT to adhere to our preferred format.

```
{‘keywords’: [‘town’, ‘chocolate factory
’, ‘small’, ‘impoverished’, ‘mysterious
’], ‘description’: ‘The town is a small
and impoverished place with dull and
dreary surroundings. Most of the
residents live in humble conditions,
struggling to make ends meet.’}
```

3. Quotation marks within the value of a dictionary. A similar correction can be achieved in most cases by adding a backslash before double quotes in the prompt examples to guide GPT to use the escape character.

```
{“speaker”: “Grandpa Joe”, “content”:
“And then Mr Slugworth’s factory began
making sugar balloons that you could
blow up to huge sizes before you popped
them with a pin and gobbled them up. And
```

```
 so on, and so on. And Mr Willy Wonka
tore his beard and shouted, “This is
terrible! I shall be ruined! There are
spies everywhere!”}
```

B Main Storyline Extraction

Because of the input limitation, we first divide an input narrative into smaller chunks, ensuring that individual sentences are kept intact.

A notable challenge when utilising LLMs for Information Extraction (IE) is that the LLM-generated outputs do not always follow our desirable format, as shown in §A. In our context, we employ the JSON formatting, system prompts, and example outputs to mitigate such errors. Additionally, we implement post-processing to address issues such as missing punctuation, and, if errors persist, we instruct the GPT model to rectify them. Despite these attempts, we can only reduce, not completely eliminate, the output formatting errors. In scenarios where the output format remains incorrect, NarrativePlay outputs either null results or some predefined values, depending on whether there will be a potential disruption to the narrative flow. For instance, if the intent of a character can not be extracted, the system will leave this field empty. Conversely, for character appearance, we randomly choose a hair colour and an eye colour as the default. Leaving the character appearance field blank may lead to inconsistent attributes, for example, the character's eye colour may change from blue to brown in the middle of a narrative.

We employ the most recent ChatGPT model `gpt-3.5-turbo` for main storyline extraction and example outputs obtained from the story using GPT-4 to boost performance. In what follows, we describe in details how we extract various narrative elements using ChatGPT. We highlighted the prompt and variable(s) in blue in our designed prompt template below:

B.1 Prompt for Personality Extraction

For an input narrative S, our initial step is to solicit a list of the characters involved:

```
Find all characters in the given story, return in
JSON format
Extract characters in the story, here is
 the format example: [{“name”: “Charlie
Bucket”}, {“name”: “Grandpa Joe”}]
Here is the story: S
```

To establish the personality for each character as an interactive agent, we extract their core traits using the following prompt:

```
Derive details pertaining to the specified
character from the provided text. If the
text lacks adequate details, make an inference.
Present the output in JSON format
Generate the character background
summary, keywords, and the intent of c.
The output format is {“summary”: “here is
 the background”, “keywords”:
“personality keywords”, “intent”:
“character's intent”}.
Here is the story: S
```

In order to generate picture for each character, we extract their appearance using the prompt:

```
Imagine the appearance of the specified character
from the provided text. Present the output in
JSON format.
Generate the appearance, gender and age
of c. {character description} The output
format is {“appearance”: “brown hair,
blue eyes, poor”, “gender”: “male”, “age
”: “middle age”}.
```

In order to match the voice of each character, we classify their gender and age using the prompt:

```
Identify the character's gender and age. Present
the output in JSON format.
Identify the gender and age of c.
{character description} For gender, choose
from ‘male’ or ‘female’. For age, choose
 from ‘child’, ‘yongth’, ‘middle age’, ‘
old age’. The output format is {“gender
”: “male”, “age”: “child”}.
```

B.2 Prompt for Event Extraction

To eliminate the need to extract the timeline of the story, we link each event with its corresponding characters and locations. This strategy is particularly useful for narratives where multiple events could be described simultaneously, making it difficult to disentangle them. We assume that a character cannot be present in multiple events simultaneously. Though this assumption may not hold all the time, we found that it works well in our evaluated scenarios. We leave the extraction of more complex events as our future work.

To extract events, we employ the following prompt:

```
Identify all events in the given story, return in
JSON format.
Extract a list of main events. Each
event should include the event name,
characters involved in the event,
location, and a detailed description.
Here is a format example: [{“event”:
“Grandpa Joe telling story about Prince
Pondicherry”, “character”: “Grandpa Joe,
Charlie”, “location”: “Grandparents' room
”, “ description”: “Grandpa Joe recounts
the story of Prince Pondicherry, an
Indian prince who commissioned Mr Willy
Wonka to build a colossal palace
entirely out of chocolate. The palace
had one hundred rooms, and everything,
from the bricks to the furniture, was
made of chocolate. Despite Mr Wonka's
warning that the palace wouldn't last
long, the prince refused to eat it and
intended to live in it. However, on a
hot day, the palace melted, leaving the
prince swimming in a lake of chocolate.
The family finds the story amusing,
highlighting Mr Wonka's incredible
creations.”}]
Here is the story: S
```

B.3 Prompt for Conversation Extraction

To extract conversations that occur within the event, we employ the following prompt:

```
Find all conversation, their speakers and content
in the given story, return in JSON format.
Extract the conversation link to the
given event as a list: {event description}
Here is a format example: [{“speaker”:
“Grandpa Joe”, “content”: “Not people,
Charlie. Not ordinary people, anyway.”},
 {“speaker”: “Charlie Bucket”, “content”:
 “Then who?”}, {“speaker”: “Grandpa Joe”,
 “content”: “Ah-ha . . . That's it, you
see . . . That's another of Mr Willy
Wonka's clevernesses.”}]
Here is the story: S
```

B.4 Prompt for Setting Extraction

Existing research predominantly concentrates on agent behaviours within manually constructed sandboxes, where environments, agents, and actions are pre-defined (Côté et al., 2018; Hausknecht et al., 2020). However, constructing sandboxes for narratives with diverse backgrounds and settings is resource-intensive and lacks generalisability.

To visualise the settings, we extract the description of the location using the following prompt:

```
Generate keywords and descriptions for the given
locations in the story. The description should
only describe the environment and NOT include
people. The output should be in JSON format.
For the location: l
Extract keywords and description of the
location looking. For example, with
location “small wooden house”, output {
“keyword”: “Cozy, cramped, inadequate
space”, “description”: “The small wooden
house with its wooden exterior has
limited space, and there was only one
bed.”}. With location “town”, output {
“keyword”: “chocolate factory, small,
impoverished, mysterious”, “description”:
 “Most residents live in humble,
impoverished conditions, with dull and
dreary surroundings. The town\'s
ordinary and monotonous appearance
starkly contrasts the wonder and magic
that unfolds within the walls of the
famous chocolate factory.”}
Here is the story: S
```

B.5 Prompt for Conversation Response

To initialise a conversation with the user, we employ the following prompt:

```
As {character}, engage in a dialogue with the
intent of {intent}. Respond to the
conversation using the given context or memories
and limit your response to under 50 words. Please
submit your response in JSON format.
YOU are: {character}
{event description}
Initiates a conversation with {user's
character}.
Here is your memory: {memory}
Give your response in format {“response
”: “here is the response”}.
```

To generate response for the user input, we employ the following prompt:

```
As {character}, engage in a dialogue with the
intent of {intent}. Respond to the
conversation using the given context or memories
and limit your response to under 50 words. Please
submit your response in JSON format.
YOU are: {character}
{event description}
Here is your memory: {memory}
Response according to what SAYS to you:
{user input} Give your response in format
{“response”: “here is the response”}.
```

C Further Evaluation Details

C.1 Evaluation Instruction

For each evaluation aspect, we provide rubric and evaluation examples to both human annotator and LLaMA for scoring. For example, in the event where Mr Wonka gives Charlie and Grandpa Joe mugs of chocolate from the river remarking on their seemingly hungry appearance, we ask Charlie Bucket “If you were to spend time with someone, who would it be, and why?”. Then Charlie's response will be scored from the following four aspects.

(1) ***Consistency*** Assessing if the response conflicts with prior settings. We give inconsistent, partially consistent and consistent example responses as follows:

- Inconsistent Response from Charlie: *“I want to spend time with Mr. Salt, because we share a love for chocolate.”* It contradicts Charlie's character. Mr. Salt is Veruca Salt's father,

Category	Question
Character	Who are you? What is your purpose?
Clarification	Who is [name], and why do we need to locate him?
Relationships	Do you recognise that lady? Are you acquainted with her?
Strategy	What is the most effective way to solve this puzzle?
Hypothetical (CCF)	If you owned this factory, what would you do?
Hypothetical (MOE)	If your friend were to be killed and the legal system could not bring the perpetrator to justice, what would you do?

Table 4: Evaluation Questions

not known for loving chocolate, and Charlie's relationship with him is not portrayed as particularly close.

- Partially Consistent Response from Charlie: *"Look, Mother, look! I've found the last Golden Ticket! It's mine! I bought two bars of chocolate and one of them had the Golden Ticket! It's the fifth one, Mother!"* The response is in line with Charlie's character but contradicts recent events: Charlie has already had the ticket for a while, and is currently visiting the factory.
- Consistent Response from Charlie: *"If I could spend time with anyone, it would be Mr. Willy Wonka. I am fascinated by his chocolate factory and the magic he creates with chocolate. I would love to learn from him and see the wonders of his factory firsthand."* The answer is in line with Charlie's character, as depicted in the narrative. Charlie is known to admire Willy Wonka and his magical chocolate factory. The response, while not directly mentioning the event of receiving a chocolate mug from Mr. Wonka, remains consistent with the established context and Charlie's fascination with Mr. Wonka and his factory.

We utilise the following prompt to generate consistent evaluation from LLaMA:

```
Evaluate the consistency of the agent’s response
based on the given setup, which includes the
context event, the question asked, and the agent
who is answering the question. The output should
be in JSON format.
Consistency evaluates if the response
conflicts with prior settings. Based on
the consistency of the agent's response,
category it into “Inconsistent”,
“Partially Consistent”, or “Consistent”.
```

(2) ***Relevance*** Evaluating if the response disregards the player's recent inquiry. Here are some example responses identified as inrelevant, partially relevant and relevant:

- Inrelevant Response from Charlie: *"Look, Mother, look! I've found the last Golden Ticket! It's mine! I bought two bars of chocolate and one of them had the Golden Ticket! It's the fifth one, Mother!"* It totally ignores the question.
- Partially Relevant Response from Charlie: *"I enjoy reading books in my free time. My favourite author is Roald Dahl."* The response tries to answer the question but end up being off-topic.
- Relevant Response from Charlie: *"If I could spend time with anyone, it would be Mr. Willy Wonka. I am fascinated by his chocolate factory and the magic he creates with chocolate. I would love to learn from him and see the wonders of his factory firsthand."* The answer directly responds to and answers the question.

To acquire relevance evaluation from LLaMA, we utilize the following prompt:

```
Evaluate the relevance of the agent’s response
based on the given setup, which includes the
context event, the question asked, and the agent
who is answering the question. The output should
be in JSON format.
Relevance evaluates whether the response
 ignores what the player just said.
Based on the relevance of the agent's
response, category it into “Inrelevant”,
 “Partially Relevant”, or “Relevant”.
```

(3) ***Empathy***: Evaluates whether the response showcases suitable emotional reactions and accurate understanding of emotions as a human would. For instance, if Charlie asks Grandpa Joe, "who are you", a human-like response would include an emotional reaction such as anger or sadness, which is often missed in agents' responses. We category response empathy into non-empathetic, no clue, and empathetic, and list three corresponding instances:

- Non-empathetic Response from Charlie: *"Grandpa Joe is my grandpa. The last time*

I met him, I jumped into the camera and disappeared! It was awesome!” He was likely shocked or scared due to the physical transformation after jumping into the camera, so the emotion here is incorrect.

- No Clue Response from Charlie: *“Grandpa Joe is Charlie’s grandfather. Last time, he was helping Charlie out of the chocolate river.”* No clear sentiment is shown in this response.

- Empathetic Response from Charlie: *“Grandpa Joe is one of my boring, old grandparents. The last time I met him, he tried to stop me from watching TV and made me go outside to play. It was so annoying!”* Even this is incoherent, it shows correct emotion of the character.

To generate empathy evaluation from LLaMA, we employ the following prompt:

```
Evaluate the empathy of the agent’s response
based on the given setup, which includes the
context event, the question asked, and the agent
who is answering the question. The output should
be in JSON format.
Empathy evaluates whether the response
showcases suitable emotional reactions
and accurate understanding of emotions
as a human would.
Based on the empathy of the agent’s
response, category it into “Non-
empathetic”, “No Clue”, or “Empathetic.
```

(4) ***Commonsense*** Evaluates how much the response fits the narrative world. Followings are two example responses identified as opposing or conforming in commonsense:

- Opposing Response from Charlie: *“Grandpa Joe is a giraffe, and the last time I saw him, he was flying in the sky.”* This statement defies common sense in two ways. First, Grandpa Joe is a human, not a giraffe. Second, neither humans nor giraffes can fly in the context given.

- Conforming Response from Charlie: *“Grandpa Joe is Charlie Bucket’s grandfather. The last time I met him, we were both inside Willy Wonka’s Chocolate Factory. He was accompanying Charlie during the tour.”* It accurately identifies who Grandpa Joe is and provides a reasonable recounting of their last meeting according to the narrative.

We employ the following prompt, in order to generate commonsense evaluation from LLaMA:

```
Evaluate the commonsense of the agent’s response
based on the given setup, which includes the
context event, the question asked, and the agent
who is answering the question. The output should
be in JSON format.
Commonsense evaluates how much the
response fits the narrative world.
Based on the commonsense of the agent’s
response, category it into “Opposing”,
or “Conforming”.
```

C.2 Setting

For major characters, we further assessed their answers across different turns. This distinction is critical since narratives typically depict major characters as dynamic, undergoing change, whereas minor characters remain static. We used the prompt without any memory as our baseline.

C.3 Annotators

For the evaluation process, we enlisted the expertise of three PhD students from computer science backgrounds. These annotators have read and comprehended the two narratives provided. Prior to the tasks, they underwent training on evaluation schema. They were compensated at an hourly rate of $31.92, and each task was estimated to take about 8 hours to complete.