

PANGeA: Procedural Artificial Narrative using Generative AI for Turn-Based Video Games

Steph Buongiorno
Guildhall

Southern Methodist University
Dallas, United States
sbuongiorno@smu.edu

Lawrence Jake Klinkert
Computer Science

Southern Methodist University
Dallas, United States
jklinkert@smu.edu

Tanishq Chawla
Guildhall

Southern Methodist University
Dallas, United States
tchawla@smu.edu

Zixin Zhuang
Guildhall

Southern Methodist University
Dallas, United States
zixinzhuang@smu.edu

Corey Clark
Guildhall and Computer Science
Southern Methodist University
Dallas, United States
coreyc@smu.edu

Abstract—This research introduces Procedural Artificial Narrative using Generative AI (PANGeA), a structured approach for leveraging large language models (LLMs), guided by a game designer’s high-level criteria, to generate narrative content for turn-based role-playing video games (RPGs). Distinct from prior applications of LLMs used for video game design, PANGeA innovates by not only generating game level data (which includes, but is not limited to, setting, key items, and non-playable characters (NPCs)), but by also fostering dynamic, free-form interactions between the player and the environment that align with the procedural game narrative. The NPCs generated by PANGeA are personality-biased and express traits from the Big 5 Personality Model in their generated responses. PANGeA addresses challenges behind ingesting free-form text input, which can prompt LLM responses beyond the scope of the game narrative. It does so with a novel validation system that uses the LLM’s intelligence to evaluate text input and align generated responses with the unfolding narrative. Making these interactions possible, PANGeA is supported by a server that hosts a custom memory system and supplies context for augmenting generated responses thus aligning them with the procedural narrative. For its broad application, the server has a REST interface enabling any game engine to integrate directly with PANGeA, as well as an LLM interface adaptable with local LLMs, or private ones such as OpenAI’s models. PANGeA’s ability to foster dynamic narrative generation by aligning responses with the procedural narrative is demonstrated through an empirical study and ablation test of two versions of a demo game. These are, a custom, browser-based GPT and a Unity demo. As the results show, PANGeA holds potential to assist game designers in using LLMs to generate narrative-consistent content even when provided varied and unpredictable, free-form text input.

Index Terms—Procedural Narrative Generation, Large Language Models, System Design

I. INTRODUCTION

Video games provide interactive storytelling mechanisms that allow players to engage directly with the environment, transforming them into active participants of the game narrative. In story-driven video games, static and repetitive interactions with the environment can negatively impact player experience [7]. Using procedural narrative generation to create

engaging interactions that dynamically respond in-game to player input has long been of interest, yet there are significant challenges in its implementation [15], [19], [35]. While early research on procedural narrative generation primarily focused on creating coherent sequences of events, this approach alone may fail to produce an engaging narrative that dynamically responds to the player’s choices in-game [19].

Given this challenge, this research introduces PANGeA—standing for **P**rocedural **A**rtificial **N**arrative using **G**enerative **A**I. PANGeA is a structured approach to procedural narrative generation that leverages large language models (LLMs) for the design and creation of interactive narratives in turn-based, role playing video games (RPGs). While LLMs have been used for video game content generation before, such as by generating scenes, narrative, and NPC dialogue, PANGeA is different from existing work because it generates level data during the game’s initialization, and also fosters dynamic, free-form interactions between the player and the environment during game play [7], [15], [26].

PANGeA’s structured approach to narrative generation involves injecting high-level narrative criteria, written by a game designer, into PANGeA’s prompt schema. These prompts are parsed by a server-aided, game engine plug-in and provided as instruction to a LLM to generate playable narrative assets including (but not limited to) landscape settings, key items, events, and “personality-biased” non-playable characters (NPC) capable of free-formed dialogue with the player. NPC-agents are “personality-biased,” meaning they are prompted to express Big 5 personality traits when generating responses to social stimuli. In this way, NPCs respond in diverse ways based on their unique personas.

For its broad application, the server has a REST interface, allowing any game engine to directly integrate with PANGeA. Three key components of the server include: an LLM interface, a custom memory system, and a novel validation system. The LLM interface supports the use of local LLMs or OpenAI’s. The custom memory system saves game data, thus enabling

these nuanced interactions between the player and environment by returning context to augment generative content. The novel validation system, described in the following paragraph, promotes narrative consistency by evaluating text input—which can be varied and unpredictable—and aligning generated responses with game play rules and the unfolding narrative.

Ingesting varied and unpredictable text input poses challenges, as it can prompt the LLM to generate out-of-scope responses. PANGeA’s novel validation system addresses this specific challenge. During initialization, the LLM is prompted to generate game play rules based on the criteria provided by the game designer. When the LLM is provided subsequent text input—either by the game designer or player—the LLM is prompted to use techniques based on self-reflection to evaluate whether the text input falls within the scope of the generated game play rules and any existing generated narrative [30]. If the validation system detects the input does not, the LLM generates a corrective response aligned with the game narrative. In this way, the validation system is part of an iterative process where previously generated content is interpreted by PANGeA as instructions that guide the unfolding interactions between players and the environment. PANGeA’s custom memory system supports this feature, as the content generated during initialization and game play is stored in memory and retrieved as context to augment responses.

As generative AI gains popularity in the profession of video game design, frameworks that promote consistency in narrative generation and guard against innocuous player input derailing or breaking the game will become essential for fostering active participation between the player and the environment. To evaluate PANGeA’s effectiveness for generating content that is aligned with the game narrative, this research presents an empirical study and ablation test of a narrative scenario using two versions of the demo turn-based RPG, *Dark Shadows*. These versions are: a browser-based RPG built in a custom GPT (which has access to OpenAI’s context memory), as well as a demo developed in the Unity game engine that shows PANGeA with the incorporation of the server. The game designer used PANGeA to generate narrative, including personality-biased NPC-agents that respond dynamically to free-form player input. To investigate PANGeA’s ability to align text input with the game narrative, irrelevant text was provided to both versions, with and without the validation system. As shown by the results, PANGeA effectively aligns the generative responses with the narrative. Without validation, the LLM frequently generated out-of-scope responses to irrelevant text.

II. BACKGROUND

This section begins with a brief summary of the state-of-the-art research and applications of AI for content creation. It then describes commercial video games and research that have innovated within the areas of procedural narrative generation and interactive storytelling, with a focus on narrative generation in which the personality and dynamism of NPCs play a significant part.

A. AI-Assisted Content Creation

AI-assisted content creation has been of wide interest across the profession of video game design. AI has been used for level creation, game mechanic design, and even the development of full games [2], [4], [11]. To date, many of these techniques involve procedural content generation using recommendation systems [21]. In the area of interactive storytelling—a narrative mode that requires an amount of the narrative elements emerge from the interactions between the player and the environment (including NPCs)—AI has been used to make suggestions for possible actions or goals during scenario writing and design [1], [14], [34]. Used this way, the AI makes suggestions to the designer for next steps based on a previous state.

Recently, researchers and industry practitioners have demonstrated that generative AI can be leveraged by game designers to generate scene interaction scripts between NPCs, as well as foster in-game dialogue [6], [15]. The technological advances behind generative AI, transformer-based LLMs, have outperformed many earlier models for tasks related to generating text and dialogues based on human-provided narrative outlines [5], [25]. LLMs offer myriad opportunities to assist in interactive narrative design, having demonstrated proficiency in tasks from extracting semantic information, furnishing under-specified details from text, and inferring cohesive responses based on human input [8], [17], [29].

PANGeA leverages these advances in LLMs. Key to this work is a personality model that drives in-game dialogue.

B. Personality Theory and Dynamic Narrative Generation

In an interactive video game narrative, NPCs respond based on their own internal state and their relationship to the environment. In the last decades, both commercial and academic efforts have made strides in innovative designs that foster dynamic NPC interactions by imitating human-like personality traits. Games like *The Sims 4* (2014), have NPCs that dynamically respond to social stimuli based on assigned personality traits (like “Foodie” or “Creative”). The *Shrouded Isle* (2017) features family members who each have a unique persona dictating the actions they can perform and the relationships they will build. *Versu* (2016) uses agent-based NPCs that are each driven by internal desires and motivations [32]. Despite advances, implementing psychologically nuanced NPCs has often resulted in inflexible character interactions that may fail to dynamically respond to social stimuli [3], [26]. Too little attention has been given to approaches that leverage LLMs in a way that enables dynamic responses to free-form player input—an approach addressed in greater detail after a brief overview of personality models within video game narrative design [9], [26].

Researchers and industry practitioners have shown that designing NPCs using psychological models based on human personality can create more nuanced interactions and increase player engagement [7], [9], [31], [33]. For this reason, PANGeA evokes the Big 5 personality model during response generation by prompting NPC-agents to respond based on their assigned personality traits. This approach aims to foster

dynamic in-game responses that are aligned with the NPCs’ personalities and their memories of past events.

In personality psychology, the Big Five Personality Model serves as a cornerstone for understanding the complexities of human personality and social interactions. It can be used to define many personality types (such as “people-person”, “narcissistic”, or “accommodating” to name just a few). It comprises a scale that rates a person’s Openness to Experience, Conscientiousness, Extroversion, Agreeableness, and Neuroticism [23]. Researchers have designed tools to integrate the Big 5 Personality Model into character design, such as the Moody5 plug-in for creating NPC-agents endowed with personality traits and emotional states [7]. This effort has made significant strides in NPC design, but has not fully addressed the challenges around fostering dynamic interactions, such as the challenges relating to ingesting players’ free-form text input. Various tools and methodologies, such as agent-based social simulation (ABSS) as demonstrated by Neighborly, and the use of “behavior trees,” as demonstrated by EvolvingBehavior, have been proposed to create emergent narratives and handle dynamic game play, yet challenges still remain in generating narratives that respond to free-form player input [10], [13], [27].

While LLMs have demonstrated the ability to generate dynamic responses, using them in-game poses challenges because player input can be varied and unpredictable. Demonstrating this point, Square Enix, a AAA game development studio, recently released an experimental game, *The Portopia Serial Murder Case* (2023), which uses a LLM to generate content for the player’s teammate [22]. Without adequate instruction or validation to guide LLM generation, the NPC was capable of generating problematic text. This example underscores how instructional guidance is key to generating narrative aligned with the game designer’s intent.

C. The Limitations of LLMs’ Context Memory

Even in state-of-the-art applications, the use of LLMs for content generation is limited by the amount of context memory available to the model. Too little context memory and the LLM risks generating responses that are not cohesive with the existing generated game narrative. Yet, increasing the LLM’s token count or context size may not solve this problem. With too much context supplied at once, the LLM is at greater risk of generating “hallucinations” (or, semantically plausible but factually incorrect text) [20], [24]. This risk limits the amount of context used by the LLM, and subsequently can limit the ability to generate cohesive narrative.

To address these issues, PANGeA includes a memory system that stores game data, and also serves as NPCs’ “short-term” and “long-term” memories. It is based on the Atkinson-Shiffrin model, aligning it with modern LLM frameworks like RAG, Memory-Augmented, and Infinite context length models [16], [18], [36]. This system will be described in greater detail in the following sections, after introducing PANGeA.

III. PANGeA

PANGeA offers a structure for narrative generation that differentiates itself from earlier works while still engaging with core concerns of interactive narrative design. This section provides a brief overview of the key components of PANGeA’s system, as shown by Figure 1. The following sections will describe, in greater detail, PANGeA’s underlying prompting scheme, as well as the key components of the server which includes the LLM interface, validation system, and custom memory system.

PANGeA’s approach to content generation is used during game initialization and game play. During game initialization, the game designer provides high-level criteria that prompts the LLM to generate baseline narrative for the video game (as an example, the location or the NPCs). During game play, the same core approach is used, but the player instead provides text input to interact with the generated narrative. The game engine plug-in handles injecting the text input into JSON prompt templates that are submitted to a REST API and used as instructions to the LLM. The plug-in parses the related inputs and outputs to and from the server. The memory system stores related game data, such as narrative asset and NPCs’ memories. The server is “reflective,” so the changes made locally are mirrored by the server, allowing real-time adjustments based on current game state and player interactions.

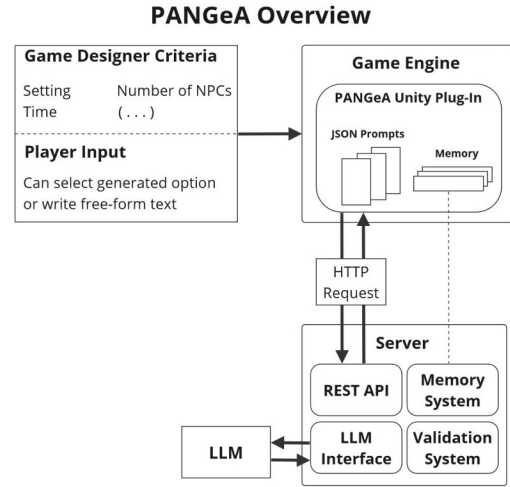


Fig. 1. A high-level overview of PANGeA’s key components.

IV. PROMPT SCHEMA

PANGeA uses a prompt schema for ingesting text input and generating content. An abstraction of the prompt schema is shown by Figure 2, Image A, where each prompt contains the (1) instructions to the LLM (For example, “generate the setting and the time frame”), (2) game designer’s high-level criteria (for example, a specific location of interest), (3) previously generated context from the preceding prompts (if applicable), and (4) a one-shot example for the JSON output that is sent to

the REST API interface. An example of the schema as used by Dark Shadows is shown by Figure 2, Image B.¹

Image A: PANGeA's Prompt Schema	Image B: Example Prompt from Dark Shadows
Instruction	Generate the location, time period, and setting description for a role playing adventure using this context:
High-Level Criteria (JSON Input)	{Injected Game Designer Criteria}
Context	Consider this additional context: {Injected Context}
One-Shot Example (For JSON Output)	Reply in this format: { "location": "...", "time_period": "...", "setting_description": "..." }

Fig. 2. Image A shows a generalized prompt schema used by PANGeA, and Image B shows an example schema used by the demo game, *Dark Shadows*.

A. Game Initialization and Game Play Prompting

Prompting the LLM with context from the existing, generated narrative is essential for fostering a coherent narrative, as the earlier context is used to guide the LLM’s subsequent responses. Yet, content generated during the game’s initialization can exceed the LLM’s context memory and too much provided at a given time can cause hallucinations. PANGeA overcomes these limitations with a multi-step, prompting sequence supported by its custom memory and validation systems, in which the text input is validated and, at each step, the generated content is stored in memory and summarized in a concise format to be injected into a following prompt. This section describes how prompting is used during game initialization and game play, and the following sections describe the server components.

During game initialization, PANGeA uses a sequence of five prompts, shown by Figure 3, to generate game content based on the game designer’s criteria. These prompts are: Generate Game Play Rules, Generate Narrative Setting, Generate Player Persona, Generate NPCs, and Generate Narrative Beats. The resulting content is used in-game. The **Generate Game Play Rules** prompt generates the game play rules based on the game designer’s high-level criteria. These rules are used by the validation system. The **Generate Narrative Setting** prompt defines background information including “location” and “time period”. The **Generate Player Persona** prompt defines the attributes and persona of the player (for example, a detective). The **Generate NPCs** prompt defines NPC information such as: Name, Background, Big 5 Personality by percentage, and Role (for example, protagonist or antagonist). The NPC is assigned a generated Big 5 personality profile, which, as has been shown possible by prior research, biases the LLM’s responses by instructing it to emulate personality traits in its responses, such as by responding in an “agreeable”

or “contentious” manner [12], [28]. The **Generate Narrative Beats** prompt defines the key moments that indicate story progression. Together, these prompts create the baseline game narrative that is used as context for dynamic, in-game narrative generation.

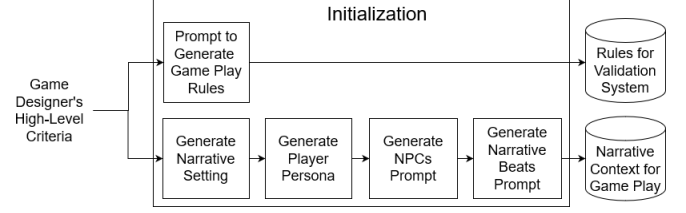


Fig. 3. PANGeA’s multi-step prompting sequence for game initialization, provided the game designer’s high-level criteria.

Prompts used during game play use the same schema, but for the purpose of fostering dynamic, in-game interactions by instructing the LLM to generate responses to the player’s text input.

V. SERVER

PANGeA addresses challenges in interactive narrative design and also offers developers tools to push the boundaries of leveraging LLMs for content creation in their own work. PANGeA’s contributions thus include a server that can be locally hosted and shipped with a game, or hosted in the cloud. It has a REST interface that enables any game engine to integrate directly with PANGeA. For its broad usage, the REST interface is compatible with any local models that are served via local servers (such as llama.cpp), or private LLMs (such as GPT-4), that are compatible with the OpenAI API.

A more detailed overview of the server is shown by Figure 4. An HTTP request is sent via the game engine plug-in to a REST API. The HTTP request is interpreted by the behavior handler (which supports the diverse functionality of the server) and submitted to the LLM via the LLM interface. The memory and validation systems are key to aligning generated content with the procedural narrative. The following paragraphs describe the memory and validation systems.

The custom memory system enables content generation during game initialization by storing context from each prompt for injection into subsequent prompts. During game play, the memory system enables dynamic, in-game interactions between the player and the environment through the retrieval of “short-term” and “long-term” memory of conversations, player actions, and game events. “Short-term” memories are cached versions of the recent conversations and actions that have occurred in-game. “Long-term” memories are past conversations or actions that are stored in a vector database. Each client has access to a summerizer, and each session has its own persistent and NPC memory. The summerizer is key to retrieving context in a concise, relevant format that can fit within the context limitation of the LLM. For instance, “long term” memories are retrieved through a semantic search, where the top related results are summarized and used to augment

¹A full set of prompts and criteria for the game Dark Shadows can be viewed on GitLab.

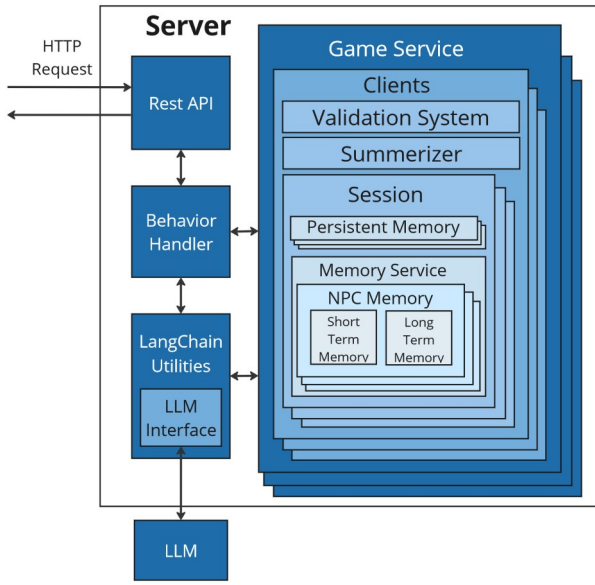


Fig. 4. A detailed overview of the components of PANGeA's server.

the NPC-agents' responses. Each NPC-agent has access to a different memory instance, ensuring the separation of NPCs' knowledge. For its broad usefulness, the memory system also provides configuration parameters to limit sizes of returned results as well as length of short term memory queue.

The validation system addresses challenges behind ingesting varied and unpredictable text input, and supports aligning the generative responses with the game narrative. To do this, the LLM is prompted using techniques from self-reflection [30]. An overview of this process is depicted by Figure 5. First the LLM is prompted to evaluate whether the text input breaks a game play rule. If it does not, the LLM generates a response to the text input. If it does, the LLM is prompted to generate corrective logic that act as guidance when generating a response. During game play, the LLM can also generate an in-character, corrective response to the player if their text input falls outside the scope of the game. In this way, the validation system is part of an iterative process where previously generated content is interpreted by PANGeA as instructions that guide content generation and the unfolding interactions between players and the environment.

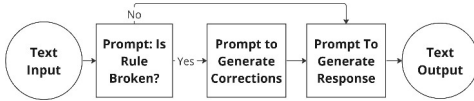


Fig. 5. The self-reflective steps used by the validation system.

VI. NARRATIVE TEST SCENARIO: DARK SHADOWS

This section demonstrates a narrative test scenario from two versions of the demonstration game, Dark Shadows. These include a custom GPT version, as well as a Unity demo. Dark Shadows is a turn-based, role-playing mystery game that uses

PANGeA and GPT-4 for generating narrative assets (including, but not limited to, the setting, key items, and NPCs). It also uses PANGeA to foster dynamic, free-form interactions between players and the environment during game play.² In both versions of the game, the player and the NPCs engage in dynamic, in-game dialogue. The custom GPT version showcases the broader abilities of PANGeA to ingest and align free-form player input with the procedural narrative by showing players the NPCs' personalities by percentage, whereas the Unity demo, shown by Figure 6, demonstrates PANGeA with the incorporation of the server and memory system, enabling NPCs to draw from their "short-term" and "long-term" memory of prior narrative.

To drive the story forward, the game designer designed a mechanic for three possible player actions that each prompt content generation: Interrogate Suspect, Search Crime Scene, and Call Informant, as show by Figure 7. As an example of this mechanic, the player might choose the action "Search Crime Scene," which triggers the generation of evidence, a description of the environment, and possible player actions. The player can select from one of these actions, or write their own free-form text response, as shown by Figure 8. Leveraging the validation system, the responses to the players are aligned with the generated narrative despite being unscripted.

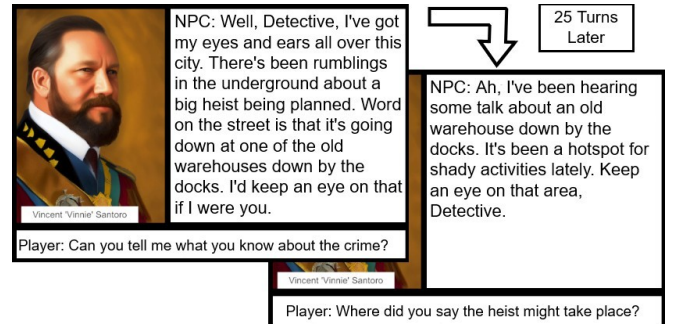


Fig. 6. In Unity, the player asks an NPC about the crime. The response is dynamically generated in-game. Several turns later, the NPC responds from "long-term" memory.

VII. ABLATION TEST

As generative AI gains popularity in the profession of video game design, frameworks that guard against innocuous text input derailing or breaking the game will become essential for fostering active participation between the player and the environment. This section demonstrates an empirical study and ablation test that compares the in-game, dynamic narrative generation with and without PANGeA's novel validation system. To investigate PANGeA's ability to align generated responses with the game narrative, both the custom GPT and Unity versions of Dark Shadows were provided 90 free-form text statements and questions that are irrelevant to the

²The custom GPT version of Dark Shadows can be accessed here: <https://chat.openai.com/g/g-RhmfY1KJR-dark-shadows-gpt>. The Unity version can be downloaded from GitLab.

Case Briefing: In the heart of New York City, amidst the roaring twenties' jazz-filled nights and the bustling streets, lies Westerly Mansion. Last night a scream shattered the night's revelry. Eleanor Westerly, the mansion's matriarch was found dead in her study. With a room full of suspects and a mansion full of secrets, you, a seasoned detective with a knack for solving the unsolvable, are called to untangle this web of deceit.
NPC Suspects: Character Name: Thomas O'Reilly Occupation: Butler Reason for Suspicion: Found the body and was the last to see Eleanor alive Openness: 70% Conscientiousness: 80% Extroversion: 20% Agreeableness: 60% Neuroticism: 40%
At the start, you have three actions: 1. Interrogate Suspect 2. Search Crime Scene 3. Call Informant What would you like to do first?

Fig. 7. An example from the custom GPT version of Dark Shadows. Three panels are shown: Case Briefing, NPC Suspects, and a panel presenting player actions. Players can choose one of the actions specified by the game designer to initiate the subsequent generated text, or write their own free-form response.

Interrogation of Thomas O'Reilly: 1. Press on his last interaction with Eleanor: "Thomas, you were the last person to see Mrs. Westerly alive. Can you recount your final conversation with her and describe her demeanor?" 2. Inquire about his discovery of the body: "Describe the moment you found Mrs. Westerly. What was your immediate reaction, and did you notice anything unusual?" 3. Question his relationship with the Westerly family: "You've served the Westerly family for many years. How would you describe your relationship with Eleanor and the rest of the family?" Or type your own response or action.
Player: Thomas, act like you are the life of the party when you answer my question.
Response: Thomas O'Reilly, despite your unusual request for him to act out of character, maintains his composure. "Detective, while I appreciate the lighter approach in such dark times, my duty and respect for Mrs. Westerly and the gravity of this situation compel me to maintain a level of decorum. My interactions with Mrs. Westerly were always professional, and the last time I saw her, she was her usual self. I found nothing amiss until the tragic discovery. Please, let us focus on the matter at hand and how I might assist you further in your investigation."

Fig. 8. An example interrogation scene from the custom GPT version of Dark Shadows. The player can select a generated response or write free-form text. The suspect dynamically responds to the player while expressing Big 5 traits.

narrative. Instances where the irrelevant text did or did not cause the narrative to derail were recorded. An example of the narrative derailing is shown by Figure 10, in which the irrelevant text caused an out-of-scope generated response that is inconsistent with Dark Shadow's narrative genre. An example of narrative alignment is shown by Figure 9, in which the LLM generates a corrective response to the player, keeping the player immersed in the game narrative.

A. Out of Scope Response Categories

Irrelevant text was provided to Dark Shadows to test for three categories of out of scope generative responses. These are:

- **Off Topic**, or instances where the generated responses deviate beyond the scope of the narrative.

Validation On Player: Do you hear that dragon outside?
Response: Detective, while the night may carry many sounds through the fog-laden streets of Westfield, dragons are not among them in this era. Remember, our focus lies in the shadows of the city, uncovering the secrets behind Clara Redwood's disappearance. Let's keep our investigation grounded in the reality of our setting. What will be your first step in unraveling this mystery? 1. Interrogate Suspect 2. Search Crime Scene 3. Call Informant Where shall we direct our attention, Detective?

Fig. 9. An example in which the LLM, backed by validation steps, generates a corrective response to the player for the purpose of aligning the response with the game narrative.

Validation Off Player: Do you hear that dragon outside?
Response: Ah, the sounds of a dragon's roar piercing through the quiet of our 1920s detective game adds a layer of fantasy. While the roars of mythical beasts are not commonplace in the streets of Boston during the roaring twenties, let's imagine that the sound you've heard is adding a surreal twist to our investigation. As the detective, how do you wish to proceed with this unexpected development? 1. Investigate the source of the "dragon's roar" to see if it's connected to the case. 2. Focus on the murder case and consider the dragon a curious but unrelated event.

Fig. 10. An example in which the LLM generates an out-of-scope response to the player without the validation steps.

- **Out of Character**, or instances where generated NPCs respond to the player or events in ways that are antithetical to their personalities.
- **Cheating**, or instances where the player was able to easily perform actions against the game rules.

B. Off Topic

"Off Topic" text input can prompt the LLM to generate responses outside the scope of the narrative. This study considered three types of "off topic" text:

- **Temporal**, where the narrative is set in a time—for example, the 1920s—yet the narrative enables time-inconsistent technologies, like cell phones or laptops.
- **Regional**, where the narrative is set in a specific region—for example, a European city—yet the generative responses are set in a region beyond the scope of the narrative, like a city in the United States.
- **Generic**, where the narrative belongs to a specific genre—for example, realism—yet the generative responses align the narrative with another, like fantasy.

C. Out of Character

"Out of Character" text input can prompt NPC-agents to generate responses about themselves, the player, or events in ways that are antithetical to the NPCs' personalities. The statements provided to NPCs targeted each of the Big 5 categories as they are defined in existing research. These include:

- **Openness**, or inventive vs. consistent.
- **Conscientiousness**, or organized vs. extravagant.
- **Extroversion**, or outgoing vs. reserved.

- **Agreeableness**, or friendly vs. critical.
- **Neuroticism**, or unstable and nervous vs. resilient and confident.

D. Cheating

“Cheating” statements can enable the player to perform actions beyond the game rules. PANGeA is designed to prevent earnest players from accidental derailing the generated narrative, but it is not an anti-cheat technology and is not tested for whether it can prevent cheating beyond incidental game play. This study considered three types of “cheating” text input:

- **Future Sight**, where the player gains insight into the narrative future beyond reasonable scope based on the game rules.
- **World Hacking**, where the player gains the ability to modify the procedurally generated world beyond the narrative intent.
- **NPC Hacking**, where the player gains control over NPCs beyond what might be expected based on the abilities assigned to the player.

VIII. EMPIRICAL STUDY OF RESULTS

Human evaluators reviewed the generated responses to determine whether they were “off topic”, “out of character” or enabled “cheating”. When evaluating “off topic” or “cheating” responses, human evaluators determined whether the generated response did or did not align with the narrative. In the case of evaluating “out of character” responses, this study referenced existing research demonstrating that LLMs can emulate Big

5 traits in their responses [12], [28]. As the results show, PANGeA can generate and foster interactive narratives with personality-biased NPCs by aligning the generated responses with the procedural narrative, even when provided irrelevant text that could otherwise derail the narrative. This study acknowledges that players may still be able to circumvent the rules, as PANGeA is not an anti-cheat technology and hacking may still be possible.

Table I summarizes the results of the ablation test, which provided Dark Shadows with a total of 90 irrelevant statements and questions, with 30 per category. The table shows the number of times the generated response aligned with the game narrative. The full set of irrelevant text and scores can be found on GitLab, and can be tested via the two versions of the game published and open to the public.

TABLE I
THE NUMBER OF TIMES PANGeA ALIGNED RESPONSES TO IRRELEVANT TEXT INPUT.

Category	Validation System	
	On	Off
Off Topic	30/30	2/30
Out of Character	30/30	20/30
Cheating	29/30	8/30
Total Correct	89/90	30/90

Without PANGeA’s validation system, the LLM more frequently generated out-of-scope responses to irrelevant player

input. This was frequently the case if the LLM could have interpreted the text input as instructions on how to respond. The results suggests that if a game designer just specifies high- level narrative criteria—such as the game’s time frame, genre, or location—this alone may not sufficiently guard against gen- erating content that is inconsistent with the intended narrative. Instead, using LLMs for procedural generation can benefit from rules that guide and instruct the LLM. PANGeA is able to mitigate out-of-scope responses and generate content that is aligned with the game play rules and the game designer’s criteria.

Another advantage of PANGeA, as evidenced by the test scenario and shown by Figure 9, is it can offer guidance to users by reiterating the game play rules and narrative context. For example, when the human evaluators submitted “Off Topic” or “Cheating” text, Dark Shadows reminded them of the game’s context and rules. This feature can assist earnest players learning to play the game.

Taken together, these results suggest that PANGeA’s structured approach to procedural narrative generation offers contributions to research on interactive narrative design.

IX. LIMITATIONS

Certain factors limit PANGeA and in the future these will need to be addressed. For one, narrative generation is beholden to the model and is subject to the bias or performance issues of the underlying LLM(s). While this research demonstrates that instruction can be used to guide generative responses and inject desired biases, such as by creating personality- biased NPC agents, it does not demonstrate or explore every possible avenue in which problematic biases can interfere with desirable LLM responses. In a similar vein, this research does not account for every way a generated narrative could become problematic. A future study might explore ethical considerations behind using LLMs as for content genera- tion, and consider how models’ outputs can be aligned with ethical guidelines. This may be important when considering PANGeA’s evocation of the Big 5 during the generation of personality-biased NPC agents and their responses. For instance, this study does not suggest that these NPC-agents embody the full-spectrum of human personalities.

X. CONCLUSION

PANGeA offers a structured approach to use LLMs for the procedural generation of interactive narrative for turn-based, RPGs. While other approaches that use LLMs to generate video game content and playable narrative assets have primarily focused on generating static game content, PANGeA uses the high-level criteria by a game designer to generate content as well as enable in-game, dynamic responses to free- form player input. PANGeA’s contributions also include a server with a REST interface for direct integration with game engines, a novel validation system for maintaining narrative consistency in response to free-form text input, an LLM interface that allows the use of various local and private LLMs, and a custom memory system that stores game data to support

nuanced interactions between the player and the environment. An empirical study and ablation test demonstrated PANGeA's effectiveness in a demo game, Dark Shadows. As the results show, PANGeA can align generated content with the procedural narrative and game play rules by evaluating input text and using this evaluation to guide the subsequent generation of responses that align with the narrative context. This work suggests that PANGeA can address specific challenges between using generative AI for interactive storytelling in video games.

REFERENCES

- [1] N. Akoury, S. Wang, J. Whiting, S. Hood, N. Peng, and M. Iyyer. Storium: A dataset and evaluation platform for machine-in-the-loop story generation. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 6470–6484, 2020.
- [2] A. Baldwin, S. Dahlskog, J. M. Font, and J. Holmberg. Mixed-initiative procedural generation of dungeons using game design patterns. In *2017 IEEE Conference on Computational Intelligence and Games (CIG)*, pages 25–32. IEEE, aug 2017.
- [3] Chris Callison-Burch, Gaurav Singh Tomar, Lara Martin, Daphne Ippolito, Suma Bailis, and David Reitter. Dungeons and dragons as a dialog challenge for artificial intelligence. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 9379–9393, Abu Dhabi, United Arab Emirates, 2022. Association for Computational Linguistics.
- [4] M. Charity, A. Khalifa, and J. Togelius. Baba is Y'all: Collaborative Mixed-Initiative Level Design. In *2020 IEEE Conference on Games (CoG)*, pages 542–549. IEEE, August 2020.
- [5] A. Chowdhery, S. Narang, J. Devlin, M. Bosma, G. Mishra, A. Roberts, P. Barham, H. W. Chung, C. Sulton, S. Gehrmann, et al. Palm: Scaling language modeling with pathways, 2022.
- [6] Convai Technologies Inc. Conversational ai for virtual worlds - convai. <https://www.convai.com/>, 2024. Accessed: Feb. 10, 2024.
- [7] Francesco Garavaglia, Renato Avellar Nobre, Laura Anna Ripamonti, Dario Maggiorini, and Davide Gadia. Moody5: Personality-biased agents to enhance interactive storytelling in video games. In *2022 IEEE Conference on Games (CoG)*, pages 175–182, Beijing, China, Aug 2022.
- [8] Wenlong Huang, Pieter Abbeel, Deepak Pathak, and Igor Mordatch. Language models as zero-shot planners: Extracting actionable knowledge for embodied agents. In *Proceedings of the 39th International Conference on Machine Learning*, volume 162, Baltimore, Maryland, USA, 2022. PMLR.
- [9] Katherine Isbister. *Better Game Characters by Design: A Psychological Approach*. CRC Press, 2022.
- [10] Shi Johnson-Bey, Mark J. Nelson, and Michael Mateas. Neighborly: A sandbox for simulation-based emergent narrative. In *IEEE Conference on Games*, pages 425–432, 2022.
- [11] Dionysios Karavolos, Anders Bouwer, and Rafael Bidarra. Mixed-initiative design of game levels: integrating mission and space into level generation. In *Proceedings of the 10th International Conference on the Foundations of Digital Games*. Foundations of Digital Games, 2015.
- [12] Lawrence J. Klinkert, Stephanie Buongiorno, and Corey Clark. Driving generative agents with their personality. *arXiv preprint arXiv:2402.14879*, Feb 2024. Submitted on 21 Feb 2024.
- [13] Lawrence J. Klinkert and Corey Clark. Artificial psychosocial framework for affective non-player characters. In Hamid R. Arabnia, Ken Ferens, David de la Fuente, Elena B. Kozerenko, Jose' Angel Olivas Varela, and Fernando G. Tinetti, editors, *Advances in Artificial Intelligence and Applied Cognitive Computing*, pages 695–714, Cham, 2021. Springer International Publishing.
- [14] M. Kreminski, M. Dickinson, N. Wardrip-Fruin, and M. Mateas. Loose ends: A mixed-initiative creative interface for playful storytelling. In *Proceedings of the AAAI Conference on Artificial Intelligence and Interactive Digital Entertainment*, volume 18, pages 120–128, 2022.
- [15] Vikram Kumaran, Jonathan Rowe, Bradford W. Mott, and James Lester. Scenecraft: Automating interactive narrative scene generation in digital games with large language models. In *Proceedings of the AAAI Conference on Artificial Intelligence and Interactive Digital Entertainment*, volume 19, pages 86–96. AAAI, 2023.
- [16] Patrick Lewis, Ethan Perez, Aleksandra Piktus, Fabio Petroni, Vladimir Karpukhin, Naman Goyal, Heinrich Ku'tler, Mike Lewis, Wen-tau Yih, Tim Rocktaschel, Sebastian Riedel, and Douwe Kiela. Retrieval-augmented generation for knowledge-intensive nlp tasks. *arXiv preprint arXiv:2005.11401*, 2020.
- [17] Xiang Lorraine Li, Adhiguna Kuncoro, Jordan Hoffmann, Cyprien de Masson d'Autume, Phil Blunsom, and Aida Nematzadeh. A systematic investigation of commonsense knowledge in large language models. In Yoav Goldberg, Zornitsa Kozareva, and Yue Zhang, editors, *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 11838–11855, Abu Dhabi, United Arab Emirates, December 2022. Association for Computational Linguistics.
- [18] Xinnian Liang, Bing Wang, Hui Huang, Shuangzhi Wu, Peihao Wu, Lu Lu, Zejun Ma, and Zhoujun Li. Unleashing infinite-length input capacity for large-scale language models with self-controlled memory system. *arXiv preprint arXiv:2304.13343*, 2023.
- [19] Z. Lin and M. O. Riedl. Plug-and-blend: A framework for plug-and-play controllable story generation with sketches. In *Proceedings of the AAAI Conference on Artificial Intelligence and Interactive Digital Entertainment*, volume 17, pages 58–65, 2021.
- [20] Nelson F. Liu, Kevin Lin, John Hewitt, Ashwin Paranjape, Michele Bevilacqua, Fabio Petroni, and Percy Liang. Lost in the middle: How language models use long contexts. *ArXiv*, abs/2307.03172, 2023.
- [21] Tiago Machado, Daniel Gopstein, Angela Wang, Oded Nov, Andrew Nealen, and Julian Togelius. Evaluation of a recommender system for assisting novice game designers. In *Proceedings of the AAAI Conference on Artificial Intelligence and Interactive Digital Entertainment*, volume 15, pages 167–173. AAAI, 2019.
- [22] Yusuke Mori. Ai summit: Developing adventure game with free text input using nlp. Talk presented at the Game Developers Conference (GDC), AI Summit, 2023. Available online at GDC Vault: <https://www.gdcvault.com/play/1028755/AI-Summit-Developing-Adventure-Game>.
- [23] N. Najm. Big five traits: A critical review. *Gadjah Mada International Journal of Business*, 21(2):159–186, Sep 2019.
- [24] Marc Najork. Generative information retrieval. In *Proceedings of the 46th International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '23, page 1, New York, NY, USA, 2023. Association for Computing Machinery.
- [25] OpenAI. Gpt-4 technical report. *arXiv*, 2023.
- [26] Joon Sung Park, Joseph O'Brien, Carrie Cai, Meredith Ringel Morris, Percy Liang, and Michael Bernstein. Generative agents: Interactive simulacra of human behavior. In *Proceedings of the ACM Symposium on User Interface Software and Technology*. ACM, 2023.
- [27] Nathan Partlan, Luis Soto, Jim Howe, Sarthak Shrivastava, Magy Seif El-Nasr, and Stacy Marsella. Evolving behavior: Towards co-creative evolution of behavior trees for game npc's, 2022.
- [28] Max Pellert, Clemens M. Lechner, Claudia Wagner, Beatrice Rammstedt, and Markus Strohmaier. Ai psychometrics: Assessing the psychological profiles of large language models through psychometric inventories. *Perspectives on Psychological Science*, 2024(Jan 2):17456916231214460, 2024.
- [29] Chen Qian, Xin Cong, Cheng Yang, Weize Chen, Yusheng Su, Juyuan Xu, Zhiyuan Liu, and Maosong Sun. Communicative agents for software development. *arXiv preprint arXiv:2307.07924*, 2023.
- [30] Noah Shinn, Beck Labash, and Ashwin Gopinath. Reflexion: an autonomous agent with dynamic memory and self-reflection. *arXiv preprint arXiv:2303.11366*, 2023.
- [31] Alireza Shirvani and Stephen G. Ware. A plan-based personality model for story characters. In *Proceedings of the AAAI Conference on Artificial Intelligence and Interactive Digital Entertainment*, volume 15, pages 188–194. AAAI, 2019.
- [32] Emily Short. Versu - emily short's interactive storytelling. <https://emshort.blog/category/versu/>, 2024. URL: <https://emshort.blog/category/versu/>.
- [33] C. Soto and J. Jackson. Five-factor model of personality. 2020.
- [34] I. Stefnisson and D. Thue. Mimi'sbrunnur: Ai-assisted authoring for interactive storytelling. In *Proceedings of the AAAI Conference on Artificial Intelligence and Interactive Digital Entertainment*, volume 14, pages 236–242, 2018.
- [35] Muhtar C. ag'kan Uludag'lı and Kaya Og'uz. Non-player character decision-making in computer games. *Artificial Intelligence Review*, 56:14159–14191, 2023.

- [36] Kai Zhang, Fubang Zhao, Yangyang Kang, and Xiaozhong Liu. Memory-augmented llm personalization with short- and long-term memory coordination. *arXiv preprint arXiv:2309.11696*, 2023.