Reconceptualizing LLM-Induced Hallucinations as Game Features

Kate Vi
Simian Institute for Advanced Studies in Humanities, East China Normal University
kattystellavi@gmail.com

Mingzhe Chen
School of Humanities Tongji University, Tongji University
chenmingzhelinda@163.com

Yuqian Sun
Computer Science Research Centre, Royal College of Art
yuqiansun@network.rca.ac.uk

Yuhao Ming
Department of Chinese Language and Literature, East China Normal University
2833962906@qq.com

Feng Wang
School of Communication, East China Normal University
wang99feng@126.com

Proceedings of DiGRA 2025

© 2025 Authors & Digital Games Research Association DiGRA. Personal and educational classroom use of this paper is allowed, commercial use requires specific permission from the author.

ABSTRACT

This study presents a novel and systematic framework to integrating Large Language Models (LLMs) into video game design by reconceptualizing the inherent characteristic phenomenon of "hallucinations"—instances where LLMs generate plausible yet inaccurate or fictitious content—as intrinsic game features. Instead of treating hallucinations as errors, we adapt them to enrich narrative complexity and enhance player experience. We introduce two key design strategies: 1) controlling narrative boundaries to limit the disruptive impact of hallucinations and 2) establishing an irrational worldview, which seamlessly incorporates these stochasticities into the game mechanism. We demonstrate these strategies through case studies of three diverse LLM-driven games across different genres. Our work contributes to the game studies community by offering innovative design paradigms that position LLMs as core interactive mechanisms, while considering their unique generative capabilities and implications for game design and research.

Keywords

Large Language Models, Game Design, Hallucinations, Narrative Innovation, Player Engagement, Interactive Mechanics, AI in Games, Dynamic Storytelling

INTRODUCTION

Video games are advancing toward an era of increased intelligence and personalization, with the emergence of Large Language Models (LLMs) offering novel possibilities for this trend (Chen et al. 2021; Kasneci et al. 2023; Zhao et al. 2024). Through LLMs, players can engage in open-ended dialogues with characters, receive generative content in real-time. When combined with multimodal capabilities, LLMs can generate new images and sounds or even serve as environmental perception and decision-making systems for characters and assets within the game. These capabilities arise from the inherent compatibility between video games and LLMs: games provide more comprehensive data inputs than the real world, and the virtual content generated by the models can be "realized" through the audiovisual presentation of the game. Therefore, LLMs present a promising research direction for game design.

However, the application of LLMs in the gaming domain is not without challenges. A key issue is the conflict between the **dynamic and unpredictable content generated by LLMs** and the predefined game world. In the field of artificial intelligence, this is referred to as **the "hallucination" phenomenon**, where LLMs may produce inaccurate, inconsistent, or entirely fictitious information during content generation (Ji et al. 2023). Hallucinations are often viewed as defects in both academic and industrial contexts, necessitating avoidance or correction during training. However, within the "Magic Circle" of video games—a space characterized by fiction and immersion—the boundaries

between reality and fiction are inherently more flexible (Salen and Zimmerman 2003). Consequently, this paper proposes a novel perspective from the standpoint of game design: transforming LLM hallucinations from "errors" into a feature of the game, thereby making them an integral component that drives narrative innovation and enhances player interaction experiences.

The primary focus of this study is to:

1. Define and characterize LLM hallucinations within the gaming context.

2. Propose two major design strategies that enable designers to mitigate the disruptive effects of hallucinations while harnessing their creative potential:

    1) Control narrative boundaries to mitigate the disruptive effects of hallucinations

    2) Establish an irrational worldview to integrate hallucinations into game order

3. Demonstrate these strategies through three game case studies

4. Discuss the value, limitations, and future directions of hallucinations in LLM- driven game design in the discussion and conclusion sections.

2. RELATED RESEARCH AND PROBLEM STATEMENT

2.1 Technical Landscapes of LLMs

Large Language Models, exemplified by GPT, do not "think" by simulating human brain activity but rather generate content based on **autoregressive models**, which predict the probability of the next word (or character) in a given text sequence (Floridi and Chiriatti 2020). However, this data-driven predictive mechanism often generates content that appears plausible but is actually inaccurate or fictitious, known as hallucinations (Ji et al. 2023).

Hallucinations are typically categorized into two types:

- **Factual Hallucination**: Content generated differs from verifiable real-world facts, often manifesting as factual inconsistencies or fabrications.

- **Faithfulness Hallucination**: Content diverges from user interactions or the context provided by the input, as well as from the internal consistency of the generated content. (Huang et al. 2024)

Most existing AI research tends to treat the hallucination problem of LLMs as a technical bottleneck, attempting to circumvent or correct hallucinations through more precise model tuning like RAG (Retrieval-Augmented Generation) or fine-tuning to verify output content or by using more precise training data to reduce the frequency of hallucinations (Lin, Hilton, and Evans 2022; Kadavath et al. 2022; OpenAI et al. 2024; Touvron et al. 2023; Wei et al. 2022; Liu et al. 2024). RAG is a method that enhances language models by integrating external knowledge retrieval into the generation process, allowing the model to access relevant information during response generation and thereby reduce hallucinations (Lewis et al. 2021). However, these methods primarily aim to avoid hallucinations rather than explore their potential value.

Unlike most AI studies aimed at avoiding hallucinations, this paper reassesses the value of hallucinations from a game design perspective, exploring how to leverage hallucinations to expand narrative possibilities and proposing specific design strategies to make them a narrative tool that enhances player experiences.

2.2 Current research on LLMs and game design

Existing research on the integration of LLMs and video games primarily concentrates on two directions: first, from the perspective of **design tools**, utilizing LLMs to support the design of existing games, such as generating storylines, character dialogues, or level content to enhance game development efficiency (e.g., cases like Werewolf, Sandbox game) (Hu et al. 2024; Xu et al. 2024; Yuan et al. 2023); second, from the perspective of **game participation**, employing LLMs as game players, commentators, or Game Masters to engage in game interactions, thereby enhancing the player experience through diverse forms of interaction(Gallotta et al. 2024; Meta Fundamental AI Research Diplomacy Team (FAIR)† et al. 2022; Wang et al. 2023; Gupta 2023; Ma et al. 2024). While some studies have touched upon the use of LLMs as game mechanisms, this aspect has not been the primary focus and remains relatively underexplored. Notable examples include *Death by AI* (Playroom 2023), where players navigate survival scenarios shaped by LLM-generated responses, and *Yandere AI Girlfriend Simulator ~ With You Til The End* (AlterStaff 2024), which engages players in escape dialogues with an LLM-driven character. Similarly, *Infinite Craft* (Neal Agarwal 2024) leverages LLMs to dynamically generate sandbox elements, and *1001 Nights* (Ada Eden 2023) allows players to co-create stories alongside an LLM-driven king.

This paper distinguishes itself by focusing on LLMs as the **central interactive mechanism**, offering native open-ended experiences rather than serving as design tools or peripheral interaction aids. We propose reconceptualizing the content generated by LLMs, including hallucinations, as design materials. By **reframing hallucinations from "errors" to "features,"** we aim to create dynamic game worlds that foster high player engagement and narrative tension. This approach moves

beyond incremental improvements to existing game genres, exploring the novel potential of LLM-driven games in both mechanism design and narrative creation.

3. TWO DESIGN STRATEGIES

When introducing LLMs into games as core interactive mechanisms, hallucinations can significantly influence narratives and player experiences. **Factual hallucinations** may undermine established factual settings within the game, while **faithfulness hallucinations** can erode players' trust in character behaviors and dialogues, causing the entire game world to become ambiguous and chaotic. For example, if Kim Kitsuragi from *Disco Elysium: The Final Cut* (ZA/UM 2021) suddenly claims to be the murderer and intends to overthrow the world due to LLM-driven content, it directly distorts the game's worldview and core narrative logic.

However, within the gaming context, the evaluation criteria for hallucinations differ from those in the real world. They rely more on the player's understanding of the internal logic and consistency of the game rather than comparisons with external reality. As a medium that combines reality and fiction, games operate within a "Magic Circle" that allows for the establishment of unique logics and rules (Salen and Zimmerman 2003). Within this circle, players accept the internal norms of the game world, even if certain content appears to defy real-world common sense, as long as it is consistent with the game's internal logic. For example, in a fictional world where "progatonist became stronger after eating mushrooms," such elements may not seem discordant **if the game provides explanations consistent with its narrative logic.**

Building on this characteristic, hallucinations in games can transcend their role as "errors" and become design features. While they may disrupt narrative consistency, they also introduce unpredictability and creative tension. In open-ended or interpretive games, hallucinations can enhance narrative diversity, encouraging players to explore uncertainty and engage more deeply with the game world. By reevaluating these elements, players actively contribute to constructing the game’s logic, enriching its narrative layers and interactivity.

To this end, this paper proposes two fundamental design strategies to transform hallucinations into game features.These strategies aim to redefine the role of hallucinations in game narratives, transforming them from "errors" into features that enhance narrative tension and interaction depth, thereby opening up new possibilities in game design.

3.1 Strategy One: Control Narrative Boundaries to Mitigate the Disruptive Effects of Hallucinations

Motivation:

As previously elucidated, both factual and faithfulness hallucinations generated by Large Language Models have the potential to disrupt the narrative coherence of the

game world. When players are exposed to extensive explicit background information or engage in prolonged dialogues, the logical inconsistencies introduced by hallucinations become increasingly perceptible, thereby undermining their trust in the game environment. This strategy aims to establish a "controlled hallucination environment" by limiting the flow of information. By doing so, it ensures that any deviations in generated content remain within an acceptable narrative range, thereby mitigating the disruptive effects of hallucinations on the game's narrative coherence.

Implementation Methods:

- **Reducing Disclosure of Verifiable Information**

  By minimizing the exposure of excessive verifiable factual details, deviations in generated content become less detectable as hallucinations.

- **Limiting Dialogue Rounds and Length**

  Restricting the number and length of dialogues diminishes the frequency of interactions between players and LLM-driven NPCs. This reduction in contextual accumulation lowers the consistency requirements for generated content, thereby decreasing the probability of faithfulness hallucinations.

- **Focusing on Specific Themes and Controlling Scope**

  Structuring dialogues to concentrate on specific themes minimizes the potential for LLMs to deviate off-topic. For example, ensuring that NPCs respond exclusively within predefined contexts prevents engagement in overly broad discussions.

3.2 Strategy Two: Introduce an Irrational Worldview to Integrate Hallucinations into Game Content

Motivation:

Hallucinations can not only be constrained but also harnessed for their potential to enhance narrative creativity through thoughtful design. As Foucault posits in *Madness and Civilization*, the boundaries between rationality and irrationality are socially constructed and inherently flexible (Foucault, 2007). Game design can exploit this flexibility by developing irrational or surreal worldviews, thereby seamlessly integrating hallucinations into the game's narrative framework. Additionally, Huizinga's *Magic Circle* theory (Huizinga and Huizinga 2009; Tekinbaş and Zimmerman 2003) and Caillois' explorations of virtuality (Caillois and Barash 2001) highlight that the rules and logic governing a game world can transcend real-world common sense, thereby justifying the inclusion of absurd and non-linear content.

Implementation Methods:

- **Introduce Unusual Backgrounds**

Develop game environments with extreme or surreal settings that inherently incorporate blurred logic and anomalous phenomena. This approach rationalizes hallucinations by embedding them within the game world's established narrative framework, making such deviations an integral and acceptable part of the player's experience.

- **Create Eccentric Characters**

Design NPCs with mental abnormalities or absurd attributes, ensuring that the logically disordered content generated by hallucinations is portrayed as characteristic traits of these characters rather than flaws of the LLM.

- **Cultivate a Carnival Atmosphere**

In comedic or absurd games, transform hallucinations into sources of humor. For instance, when players encounter AI-generated dialogues that are incoherent or nonsensical, these erratic responses can become a source of amusement, thereby enhancing the game's entertainment value.

4. CASE STUDIES: THREE LLM-DRIVEN GAMES

This section presents three specific case studies—"***Yandere AI Girlfriend Simulator ~ With You Til The End***" (AlterStaff 2024), "***1001 Nights***" (Ada Eden 2023), and "***Suck Up!***" (Proxima Enterprises 2024)—to demonstrate the application of Strategy One (controlling narrative boundaries to reduce disruption) and Strategy Two (introducing an irrational worldview) in actual games. These case studies aim to illustrate how LLM-generated "hallucinations" can be transformed into design features that enhance narrative and interactive experiences, and to explore the functions and manifestations of hallucinations in different gaming contexts.

4.1 Case Overview

- "Yandere AI Girlfriend Simulator ~ With You Til The End"

A yandere girlfriend simulation game developed based on ChatGPT, where players awaken in a confined room and engage in dialogues with a "yandere girlfriend" who is obsessed with the player, easily angered, and may suddenly initiate attacks. Players need to gradually uncover background information through interactions while avoiding triggering the girlfriend's negative emotions.

- "1001 Nights"

A fantasy narrative game originating from an academic project, where players assume the role of the legendary Scheherazade, telling stories to a tyrannical king. The AI-driven king generates real-time responses based on the story content, and players guide the king to complete tasks using specific keywords. The game combines

innovative language and strategy gameplay, receiving nominations for the Lumen Prize and various international game awards.

- "Suck Up!"

A parody sandbox game where players act as vampires, gaining the trust of NPC neighbors through disguise and dialogue to gain access to their rooms. The game features an absurd and carnival-like style, with multiple playthrough videos on YouTube garnering over a million views.

These three games, despite differing in genre and theme, share a commonality: **interaction with LLMs constitutes the core game mechanic**. They successfully address the hallucination issue through specific designs, transforming hallucinations from potential defects into organic elements that enhance narrative and interactive experiences.

4.2 "Yandere AI Girlfriend Simulator": High-Pressure Interactions in a Confined Space

Strategy One (Controlling Narrative Boundaries to Mitigate Disruptive Effects):

"Yandere AI Girlfriend Simulator" places the player in an extremely information-restricted "confined room" scenario. Upon awakening, the player's sole means of acquiring information, aside from conversing with the yandere girlfriend, is through observing the room—such as examining wall paintings, witnessing billowing red smoke outside the window, and reading two emails on the computer. These sparse clues gradually emerge throughout the game without detailed explanations, providing players with a disorienting and information-scarce introduction.

- **Thematic Focus**: The content of interactions is concentrated on two main themes—romance and escape—thereby reducing the likelihood of the LLM deviating from the topic.

- **Minimal Information Disclosure**: Players cannot understand the game world through extensive background information. Instead, they gather information through limited environmental clues and sparse dialogues, relying heavily on inference. This approach minimizes the potential for factual hallucinations conflicting with the preset narrative.

- **Limiting Dialogue Rounds and Duration**: In the game, the yandere girlfriend employs an immediate evaluation and punishment mechanism. If players ask inappropriate questions or touch upon topics that displease her, she will attack the player, resulting in death after four hits. This mechanism forces players to speak cautiously within a very short timeframe, avoiding lengthy dialogues that could lead to contextual inconsistencies.

Strategy Two (Introducing an Irrational Worldview):

The very background setting of "Yandere AI Girlfriend Simulator" deviates from conventional logic, constructing an irrational and marginalized scenario:

- **Irrational Background Setting**: The abnormal sights of an apocalyptic world and the high-pressure situation of being kidnapped together create a game world that no longer adheres to realistic rationality standards. The unconventional worldview diminishes the player's expectations of the conventional order. Hallucinations are systematized as part of a "madness" logical structure, integrated into the game narrative.

- **Yandere Character Traits**: The yandere girlfriend's inconsistencies, pathological logic, and aberrant emotional expressions are rationalized as her character traits rather than flaws of the LLM. Hallucinations become a distinctive feature of the character.

The combination of these two strategies ensures that even if the LLM produces some fabricated outputs, they can be interpreted either as new world knowledge (lacking verification methods) or as manifestations of the girlfriend's mental instability. When hallucinations occur, they are not viewed as glitches but as inherent game features.

4.3 "1001 Nights": Stories in a Fantasy World

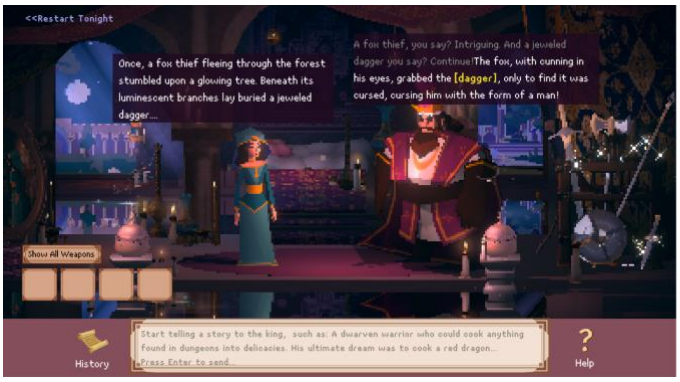
Strategy One (Controlling Narrative Boundaries to Mitigate Disruptive Effects):

"1001 Nights" focuses its scenes within the king’s private chamber and corridors, supplemented by sporadic dialogues with a few court maids and guards, which constitute the primary sources of world information in the game.

- **Limited Scene Presentation**: The game's environments are primarily confined to the king's private chamber and corridors. Minimal dialogues with court maids and guards provide necessary background information, avoiding excessive setting details and reducing the exposure of factual hallucinations.

- **Patience Mechanism Limiting Dialogue Rounds**: The game features a patience meter for the king, restricting players to focus on "storytelling" within a limited number of turns. Repeatedly causing the king's displeasure results in game termination.

- **Strict Thematic Control**: The player's main task is to tell stories to the king, aiming to elicit specific keywords (weapon names) to transition the game from the dialogue phase to the combat phase. Dialogues are strictly confined to the "storytelling" theme, minimizing the possibility of the LLM deviating from the topic. The king's responses are structured, containing both commentary on the story and personalized continuations, keeping the dialogue focused within a narrow thematic scope. The fictional nature of the stories themselves eliminates the potential for factual hallucinations.



Strategy Two (Introducing an Irrational Worldview):

"1001 Nights" adopts a folk-tale-inspired fantasy background, constructing a setting where a tyrannical king executes a new bride every day and endowing the player character with the supernatural ability to transform language into reality:

- **Fantasy Folklore Background**: Building upon the classic "1001 Nights" stories, the protagonist Scheherazade is granted the supernatural ability to turn words into reality. This **extraordinary** trait increases the narrative distance from real-world logic.

- **Extremely Cruel King**: The king's daily practice of marrying and killing a new bride, coupled with his mental instability, is highly plausible within the game's context. Hallucinations serve to reinforce the king's bizarre characteristics (AdaEden1001 2024).

In "1001 Nights," hallucinations are integrated into the legendary story atmosphere. The king's abrupt and bizarre statements reflect his inscrutable psyche and the internal logic of the mysterious world, encouraging players to tell stories in more creative ways. The hallucinations here not only preserve the immersive experience but also enhance the mystique of the legend, infusing the narrative with a sense of experimentation and imaginative potential.

4.4 "Suck Up!": A Farcical Comedy Game

Strategy One (Controlling Narrative Boundaries to Mitigate Disruptive Effects):

In "Suck Up!", players assume the role of an alien vampire with no knowledge of the town, relying solely on conversations at NPC doorsteps, the appearance of houses, and NPC names to gather fragmented information:

- **Minimal Information Disclosure**: The game lacks detailed background settings, with all dialogues occurring at NPC residences' doorsteps. The absence of complex world-building reduces the occurrence of factual hallucinations.

- **Fixed Conversation Locations and Themes**: All interactions are confined to NPC homes, with dialogues centered around gaining trust and permission to enter. NPCs respond immediately to the player's speech; if inappropriate, they shut the door. Players must then knock again with a new disguise to restart the conversation, avoiding the accumulation of extensive contextual dialogue.

- **Short Dialogues to Reduce Context Accumulation**: All dialogues are limited to NPC doorsteps. The trust system ensures that if players perform poorly, NPCs immediately close the door, preventing the buildup of lengthy contexts.

Strategy Two (Introducing an Irrational Worldview):

"Suck Up!" employs a festive, carnival-like, and parodic comedic atmosphere to buffer the presence of hallucinations:

- **Festive and Carnival Atmosphere**: Players can freely disguise themselves to deceive NPCs, such as picking up a cardboard box to masquerade as a delivery person or running naked in the snow, creating an absurd and joyful game setting.

- **Comedic Setting**: NPCs exhibit inconsistent attitudes and may disregard real-world logic at any moment, becoming part of the game's design. The carnival-like festivities and bizarre styles eliminate the need for players to adhere to realistic logic. When NPCs utter strange content, it is perceived as humorous. Hallucinations are thus humorized, transforming "errors" into comedic elements.

- **Deception as Game Objective**: Hallucinations are designed as the player's objective. Players must deceive NPCs to prevent them from recognizing their vampire identity. Consequently, inconsistent NPC responses not only align with game logic but also add fun and satisfaction to achieving objectives.

In "Suck Up!", hallucinations become a source of game humor. NPCs' off-topic or bizarre reactions enhance the story's comedic aspect, supporting parody and uncertainty, thereby enriching the layers of player experience. Hallucinations do not undermine immersion; instead, they enhance the game's entertainment value through comedic elements, allowing players to enjoy humor while experiencing richer interactive enjoyment.

4.5 Case Summary

The three games successfully transform hallucinations into narrative driving forces and gameplay features through the two primary design strategies:

- **"Yandere AI Girlfriend Simulator"** utilizes a confined setting and character design, making hallucinations part of the narrative tension.

- **"1001 Nights"** integrates hallucinations into a legendary storytelling framework, enhancing the game's experimental and creative aspects.

- **"Suck Up!"** employs comedic presentation, turning hallucinations into sources of entertainment and interactive fun.

These case studies demonstrate how controlling and leveraging hallucinations through design strategies can adapt to different game genres, providing practical foundations and insights for future LLM-driven game development.

5. DISCUSSION

The powerful generative capabilities of Large Language Models (LLMs) present new opportunities for game design, allowing narratives and interactive forms to transcend traditional limitations. However, the core challenge for designers is determining

how randomly generated content can maintain acceptability within a preset framework and evolve into creatively valuable interactive experiences. In this context, hallucinations should no longer be simply viewed as technical flaws or logical loopholes but should be integrated into the conceptualization of game design, transforming them into creative driving forces for narrative and gameplay.

The Transformation Path of Hallucinations: From "Errors" to "Features"

In traditional game design, designers tend to organize game narratives in a completely deterministic story tree, viewing any deviations as errors. However, the "Magic Circle" of games provides a natural stage for reshaping logic and rules. Just as players can accept settings that defy real-world common sense within the game world, they can also embrace the uncertainty and tension brought by hallucinations under appropriate design strategies. The two strategies proposed in this paper offer tangible pathways for this transformation.

Narrative and Design Value of Hallucinations

Through these two strategies, hallucinations demonstrate the following values in game narrative and design:

- **Creating More Distinctive Characters**: Characters exhibiting contradictory or unconventional behaviors due to hallucinations are perceived as having inherent traits, thereby deepening their personality and complexity.

- **Enriching World Atmosphere and Narrative Tension**: Uncertainty allows the world to transcend a wholly preset logical framework, transforming it into a dynamic environment filled with bizarre, absurd, or suspenseful atmospheres. The emergence of hallucinations enables players to gain deeper immersion by continuously reconstructing their understanding of the world.

- **Fostering Unexpected Surprises and Creative Interactions**: The randomness and unpredictability brought by hallucinations create unforeseen interactive spaces for players. In dealing with such uncertainty, players themselves become co-creators of the narrative process, making the game experience more open and creative.

This phenomenon is vividly exemplified in a popular *Suck Up!* gameplay livestream, which has garnered over four million views on YouTube (CaseOh 2024). In this session, the player attempts to deceive an NPC by claiming to be their nephew. When asked for a last name, the player improvises a matching surname. Unexpectedly, the AI NPC rejects the claim, saying, “Don’t try to fool me with a made-up name,” and abruptly ends the interaction. Whether this response stems from a hallucinated failure to recognize the fabricated connection or from a plausible in-character suspicion is left unclear.

Instead of breaking the game's logic, the ambiguity generates humor and deepens player engagement. The streamer, frustrated yet amused, shouts "Your last name *IS* Boomer! What do you mean?!"—provoking laughter and excitement in the live chat. Rather than seeing the hallucination as a glitch, the audience perceives it as a delightful twist that fuels improvisational role-play. The incident not only motivates the player to continue pursuing the NPC but also transforms a potential error into a memorable and personalized gameplay moment.

This illustrates how hallucinations, when situated within flexible and humorous narrative structures, can serve as catalysts for emergent player creativity, reinforcing the game's improvisational dynamics and community-based enjoyment.

Design Inspirations for the Future

For game designers, embracing hallucinations does not mean reverting to chaos but redefining design goals and methods. Shifting from "meeting player expectations through strict logic" to "stimulating player creativity through uncertainty" aligns with the current gaming industry's pursuit of open-world, multi-linear narratives, and immersive interactions.

- **Finding Balance Between Uncertainty and Predefined Structures**: Although hallucinations are incorporated as a feature in game design, maintaining narrative coherence remains paramount. Techniques such as Retrieval-Augmented Generation (RAG), which enhances language models by integrating external knowledge retrieval into the generation process, continue to be invaluable in reducing the frequency of hallucinations, thereby ensuring that they do not precipitate the complete collapse of the narrative framework.

- **A New Paradigm for Dynamic Narratives**: Hallucinations offer a non-linearand continuously regenerating possibility for game narratives. Designers can view them as engines for dynamic storytelling, making the game's story not a preset text but a process co-created by players and AI.

- **Revolutionizing the Definition of Immersion**: Immersion does not rely on predefined consistency, but is instead achieved through the dynamic uncertainty introduced by hallucinations. In a game world that embraces hallucinations, players can interact with and attempt to make sense of them, allowing for a deeper and more direct engagement with the narrative space of the game.

Ethical Considerations

While the deployment of Large Language Models (LLMs) in various applications raises significant ethical concerns—such as data privacy, misinformation, and algorithmic bias—this study focuses specifically on their use within game environments. However, even within the magic circle of play, hallucinations are not without risk. In particular, the unpredictable or misleading outputs of generative models may undermine player trust, distort narrative understanding, or cause emotional discomfort—especially in games that simulate real-world scenarios, involve sensitive topics, or claim educational or therapeutic value.

In such cases, hallucinations must be treated with heightened ethical scrutiny. Designers should clearly communicate the potential for AI-generated inaccuracies, and make players aware that certain outputs may not reflect intentional design, but rather the stochastic nature of the model. This transparency is particularly vital in **serious games**, where players may rely on the presented information to form beliefs or decisions. Acknowledging and contextualizing hallucinations helps preserve player agency and supports more responsible player-model interaction.

The industry is already responding to this need for greater transparency. For example, game platforms like Steam followed up with these considerations, and put strict survey for any game that contains AI realated content to disclosure their use of AI on steam page. For instance, 1001 Nights put detailed technical disclosure, including the model and workflow they used, publicly on Steam. We suggest game study

community to actively discuss the ethical and considerate way to adapt AI technology, and mitigating the potential problems and challenges.

We encourage the game studies community to continue exploring ethical frameworks for LLM integration in games, especially regarding transparency, user trust, and responsible content moderation. As AI becomes more embedded in game systems, a nuanced understanding of its affordances and limitations will be essential to maintaining both creative freedom and ethical integrity.

Limitations and Future Work

This study acknowledges limitations related to the scalability and generalizability of the proposed design strategies across diverse game genres. The effectiveness of hallucination-driven interaction is most evident in **single-player, narrative-driven games** that involve close player–NPC engagement and tolerate ambiguity, such as interactive fiction or experimental comedy games. However, in **multiplayer games**, hallucinations may introduce inconsistent world states, disrupt social coordination, or create perceptions of unfairness among players. Similarly, in **serious or educational games**, hallucinations risk spreading misinformation, undermining trust, and compromising the game's intended learning outcomes. Additionally, there remain challenges in ensuring that players interpret hallucinations as **intentional features** rather than **technical glitches**. Player reception may vary across cultural, genre, or gameplay contexts, calling for further study into how designers can frame, signal, or contextualize hallucinations to maintain immersion and credibility. Future research should investigate how LLMs perform across different genres and gameplay modes, examining the specific conditions and constraints under which hallucinations can be meaningfully integrated into the game design. More empirical work is also needed to assess player responses to hallucination-based mechanics in varied gameplay contexts. In summary, LLM-induced hallucinations prompt designers to reevaluate the relationship between narrative and interaction: moving from traditional preset logic to a dynamic, co-creative, and multi-dimensional narrative ecosystem. In such a complex and uncertain environment, hallucinations transform from "errors" into "features," becoming powerful catalysts for expanding game design boundaries and activating player imagination.

6. CONCLUSION AND FUTURE WORKS

This study presents a novel approach to integrating Large Language Models (LLMs) into video game design by reconceptualizing the traditionally problematic phenomenon of hallucinations as intrinsic game features. Unlike research in AI field which predominantly treats hallucinations as technical flaws to be mitigated, this work leverages them to enhance narrative depth and player engagement, positioning LLMs as core interactive mechanisms rather than mere design tools or peripheral elements.

We began by clearly defining and categorizing LLM-induced hallucinations within the context of gaming. We distinguished between **factual hallucinations**—where generated content deviates from real-world facts—and **faithfulness hallucinations**—where content diverges from the provided context or user interactions. Understanding these types allowed us to assess their different impacts on game narratives and player experiences.

Building on this foundation, we proposed two key design strategies:

1. **Controlling Narrative Boundaries:** This strategy aims to limit the disruptive effects of hallucinations by restricting the flow of information and maintaining thematic focus. By controlling dialogue length and topic scope, game designers can limit deviations to within acceptable narrative limits, thereby preserving the game's coherence.

2. **Establishing an Irrational Worldview:** By creating game worlds with surreal or chaotic logic, hallucinations become a natural part of the game's narrative. This approach allows players to accept and even expect unusual content, integrating hallucinations seamlessly into the game's framework.

We validated these strategies through case studies of three diverse LLM-driven games: **"Yandere AI Girlfriend Simulator ~ With You Til The End," "1001 Nights,"** and **"Suck Up!"** These examples demonstrate how our strategies can be applied across different genres to transform hallucinations into narrative and gameplay strengths. Our case studies revealed that controlling narrative boundaries effectively maintains game coherence, while establishing an irrational worldview enhances player immersion by normalizing unexpected content.

This paper suggests re-considering the inherent unpredictability of LLMs as special features for game design. We thus suggest game designers create dynamic, open-ended systems that embrace chaos and unpredictability, rather than strictly rational, predefined, and closed-off structures. This approach fosters unexpected plot twists and enhances player agency, making the gaming experience more engaging and interactive.

However, this study also acknowledges limitations such as the scalability of these strategies across different game genres and the potential challenges in maintaining narrative consistency. Future research could explore automated methods for dynamically adjusting narrative boundaries and further investigate player responses to integrated hallucinations.

In summary, this study advances the application of LLMs in gaming by providing a scalable and adaptable framework for future research and development. By embracing the dual nature of hallucinations, game designers can achieve greater narrative innovation and deeper player engagement, ultimately expanding the creative possibilities within the gaming industry.

REFERENCES

Ada Eden. 2023. *1001 Nights*. Online Game. Ada Eden.

AdaEden1001. 2024. 'When Tweaking the King's Model Feedback in 1001 Nights'. X.

15 December 2024. https://x.com/AdaEden1001/status/1867703925249650779.

AlterStaff. 2024. *Yandere AI Girlfriend Simulator ~ With You Til The End*. Online Game. AlterStaff.

Caillois, Roger, and Meyer Barash. 2001. *Man, Play, and Games*. Urbana: University of Illinois Press.

CaseOh. 2024. *I felt evil playing this game (Suck Up*. Video. YouTube, 22 May.

https://www.youtube.com/watch?v=ZG1-8vZYt1k&t=1531s

Chen, Mark, Jerry Tworek, Heewoo Jun, Qiming Yuan, Henrique Ponde de Oliveira Pinto, Jared Kaplan, Harri Edwards, et al. 2021. 'Evaluating Large Language Models Trained on Code'. arXiv. https://doi.org/10.48550/arXiv.2107.03374.

Floridi, Luciano, and Massimo Chiriatti. 2020. 'GPT-3: Its Nature, Scope, Limits, and Consequences'. *Minds and Machines* 30 (4): 681–94. https://doi.org/10.1007/s11023-020-09548-1.

Foucault, Michel, David Cooper, and Michel Foucault. 2007. *Madness and Civilization: A History of Insanity in the Age of Reason*. Translated by Richard Howard. Routledge Classics. London: Routledge.

Gallotta, Roberto, Graham Todd, Marvin Zammit, Sam Earle, Antonios Liapis, Julian Togelius, and Georgios N. Yannakakis. 2024. 'Large Language Models and Games: A Survey and Roadmap'. *IEEE Transactions on Games*, 1–18. https://doi.org/10.1109/TG.2024.3461510.

Gupta, Akshat. 2023. 'Are ChatGPT and GPT-4 Good Poker Players? -- A Pre-Flop Analysis'. arXiv. https://doi.org/10.48550/arXiv.2308.12466.

Hu, Sihao, Tiansheng Huang, Fatih Ilhan, Selim Tekin, Gaowen Liu, Ramana Kompella, and Ling Liu. 2024. 'A Survey on Large Language Model-Based Game Agents'. arXiv. https://doi.org/10.48550/arXiv.2404.02039.

Huang, Lei, Weijiang Yu, Weitao Ma, Weihong Zhong, Zhangyin Feng, Haotian Wang, Qianglong Chen, et al. 2024. 'A Survey on Hallucination in Large Language

Models: Principles, Taxonomy, Challenges, and Open Questions'. *ACM Transactions on Information Systems*, November, 3703155. https://doi.org/10.1145/3703155.

Huizinga, Johan, and Johan Huizinga. 2009. *Homo Ludens: A Study of the Play-Element in Culture*. 30. [Nachdr.]. Boston: The Beacon Press.

Ji, Ziwei, Nayeon Lee, Rita Frieske, Tiezheng Yu, Dan Su, Yan Xu, Etsuko Ishii, Ye Jin Bang, Andrea Madotto, and Pascale Fung. 2023. 'Survey of Hallucination in

Natural Language Generation'. *ACM Computing Surveys* 55 (12): 1–38. https://doi.org/10.1145/3571730.

Kadavath, Saurav, Tom Conerly, Amanda Askell, Tom Henighan, Dawn Drain, Ethan Perez, Nicholas Schiefer, et al. 2022. 'Language Models (Mostly) Know What They Know'. arXiv. https://doi.org/10.48550/arXiv.2207.05221.

Kasneci, Enkelejda, Kathrin Sessler, Stefan Küchemann, Maria Bannert, Daryna Dementieva, Frank Fischer, Urs Gasser, et al. 2023. 'ChatGPT for Good? On Opportunities and Challenges of Large Language Models for Education'.

Learning and Individual Differences 103 (April):102274. https://doi.org/10.1016/j.lindif.2023.102274.

Lewis, Patrick, Ethan Perez, Aleksandra Piktus, Fabio Petroni, Vladimir Karpukhin, Naman Goyal, Heinrich Küttler, et al. 2021. 'Retrieval-Augmented Generation for Knowledge-Intensive NLP Tasks'. arXiv. https://doi.org/10.48550/arXiv.2005.11401.

Lin, Stephanie, Jacob Hilton, and Owain Evans. 2022. 'TruthfulQA: Measuring How Models Mimic Human Falsehoods'. arXiv. https://doi.org/10.48550/arXiv.2109.07958.

Liu, Fuxiao, Kevin Lin, Linjie Li, Jianfeng Wang, Yaser Yacoob, and Lijuan Wang. 2024. 'Mitigating Hallucination in Large Multi-Modal Models via Robust Instruction Tuning'. arXiv. https://doi.org/10.48550/arXiv.2306.14565.

Ma, Weiyu, Qirui Mi, Yongcheng Zeng, Xue Yan, Yuqiao Wu, Runji Lin, Haifeng Zhang, and Jun Wang. 2024. 'Large Language Models Play StarCraft II: Benchmarks and A Chain of Summarization Approach'. arXiv. https://doi.org/10.48550/arXiv.2312.11865.

Meta Fundamental AI Research Diplomacy Team (FAIR)†, Anton Bakhtin, Noam Brown, Emily Dinan, Gabriele Farina, Colin Flaherty, Daniel Fried, et al. 2022. 'Human-Level Play in the Game of *Diplomacy* by Combining Language Models with Strategic Reasoning'. *Science* 378 (6624): 1067–74. https://doi.org/10.1126/science.ade9097.

Neal Agarwal. 2024. *Infinite Craft*. Web browser game. Neal Agarwal.

OpenAI, Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, et al. 2024. 'GPT-4 Technical Report'. arXiv. https://doi.org/10.48550/arXiv.2303.08774.

Playroom. 2023. *Death by AI*. Web Browser Game. Playroom.

Proxima Enterprises. 2024. *Suck Up!* Online Game. Proxima Enterprises. Salen, Katie, and Eric Zimmerman. 2003. 'This Is Not a Game: Play in Cultural

Environments'. In *DiGRA Conference*. https://api.semanticscholar.org/CorpusID:8118681.

Tekinbaş, Katie Salen, and Eric Zimmerman. 2003. *Rules of Play: Game Design Fundamentals*. Cambridge, Mass: MIT Press.

Touvron, Hugo, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, et al. 2023. 'LLaMA: Open and Efficient Foundation Language Models'. arXiv. https://doi.org/10.48550/arXiv.2302.13971.

Wang, Guanzhi, Yuqi Xie, Yunfan Jiang, Ajay Mandlekar, Chaowei Xiao, Yuke Zhu, Linxi Fan, and Anima Anandkumar. 2023. 'Voyager: An Open-Ended Embodied Agent with Large Language Models'. arXiv. https://doi.org/10.48550/arXiv.2305.16291.

Wei, Jason, Yi Tay, Rishi Bommasani, Colin Raffel, Barret Zoph, Sebastian Borgeaud, Dani Yogatama, et al. 2022. 'Emergent Abilities of Large Language Models'. arXiv. https://doi.org/10.48550/arXiv.2206.07682.

Xu, Zelai, Chao Yu, Fei Fang, Yu Wang, and Yi Wu. 2024. 'Language Agents with Reinforcement Learning for Strategic Play in the Werewolf Game'. arXiv. https://doi.org/10.48550/arXiv.2310.18940.

Yuan, Haoqi, Chi Zhang, Hongcheng Wang, Feiyang Xie, Penglin Cai, Hao Dong, and Zongqing Lu. 2023. 'Skill Reinforcement Learning and Planning for Open- World Long-Horizon Tasks'. arXiv. https://doi.org/10.48550/arXiv.2303.16563.

ZA/UM. 2021. Disco Elysium: The Final Cut. PC Game. ZA/UM.

Zhao, Wayne Xin, Kun Zhou, Junyi Li, Tianyi Tang, Xiaolei Wang, Yupeng Hou, Yingqian Min, et al. 2024. 'A Survey of Large Language Models'. arXiv. https://doi.org/10.48550/arXiv.2303.18223.