Inclusive FPS Game Design using Reflexion-Based Adaptation System for Older Adults

ANONYMOUS AUTHOR(S)

First-person shooter (FPS) games often exclude older adults due to fast-paced controls, violent visuals, and cognitive demands. To address this, we propose a training-free adaptation system based on the Reflexion architecture, which uses multi-agent large language models (LLMs) to suggest between-session gameplay presets tailored to older adult players. The system aims to (1) offer actionable, an explainable design rationale without relying on gameplay data, and (2) improve playability through modular adjustments in UI, weapon design, aiming assistance, and a bundled adaptation called Narrative Thematic Style. Our method integrates three GPT agents that analyze performance and post-session feedback to iteratively generate preset configurations. Unlike systems that require large datasets, our system enables reasoning under data-scarce conditions and provides interpretable design logic. We implemented a Unity prototype and conducted play sessions with four older adults. The system identified key accessibility needs, such as clearer UI, simpler controls, and reduced violence. While quantitative changes were modest, participants reported greater emotional comfort and satisfaction with the adaptive experience. This study shows how Reflexion-based LLM agents can support inclusive game design for underrepresented players by enabling scalable, explainable, feedback-driven adaptation strategies.

CCS Concepts: • **Human-centered computing** → **Accessibility systems and tools**; • **Applied computing** → **Computer games**.

Additional Key Words and Phrases: Inclusive Game Design, Older Adults, Reflexion, Multi-LLM Agents

ACM Reference Format:

Anonymous Author(s). 2025. Inclusive FPS Game Design using Reflexion-Based Adaptation System for Older Adults. In *Proceedings of Make sure to enter the correct conference title from your rights confirmation email (Conference acronym 'XX).* ACM, New York, NY, USA, 9 pages. https://doi.org/XXXXXXX.XXXXXXX

1 Introduction

Digital games are increasingly recognized as meaningful sources of leisure, emotional stimulation, and social engagement for older adults [8, 12]. However, most commercial games are not designed with older adult players in mind and fail to account for their lived experiences, emotional needs, or comfort levels. Prior research has identified several barriers to older adult gameplay—particularly in fast-paced genres like FPS—including unfamiliar narrative conventions, discomfort with violent content, and difficulty navigating visually dense or unintuitive interfaces [7]. Despite these challenges, older adult players often express curiosity toward games in this genre, yet their experiences remain understudied and underrepresented in design practices.

One major obstacle to inclusive FPS design is the lack of gameplay data from older adults. Because they are rarely the target audience for these games, little is known about how specific adaptations—such as simplifying visuals, altering narrative tone, or softening violent imagery—impact their engagement or comfort. As a result, game designers are left without actionable data or rationale to guide accessibility improvements. This absence of player data is especially problematic for adaptation systems that rely on training or behavioral modeling, which assume the existence of large, representative datasets.

To address this gap, we adapt the **Reflexion** system [15]—originally proposed for agent self-improvement through verbal reinforcement—to the context of inclusive game adaptation. Our approach uses three collaborating LLM agents: an **Evaluator**, which assesses the success of prior adaptations; a **Self-Reflection agent**, which identifies failure patterns and generates insights; and an **Actor**, which outputs a new adaptation plan for the next session. These agents operate between sessions, using performance data and post-session feedback from older adult players to drive modular updates to gameplay—such as adjusting enemy difficulty, UI size, and narrative framing—without requiring any prior training data.

This paper introduces a training-free, Reflexion-based adaptation system for FPS games. The system allows designers to explore what kinds of changes may improve older adults' experiences in the absence of existing gameplay norms or datasets. By embedding reasoning agents into the design loop, we investigate whether Reflexion can provide both effective adaptations and interpretable design rationale.

We explore the following research questions:

- **RQ1:** Can a training-free Reflexion-based adaptation system provide designers with meaningful rationale and actionable insights for creating accessible FPS gameplay for older adults, even in the absence of large-scale user data?
- **RQ2:** To what extent can Reflexion-based, between-session adaptation improve older adults' FPS playability in terms of mental comfort, input precision, and emotional engagement?

We contribute: (1) a novel multi-agent adaptation system based on Reflexion that simulates iterative design feedback, (2) a working prototype using modular adaptation presets and post-game reflection, and (3) findings from older adult play sessions that reveal how this approach supports design reasoning in data-scarce accessibility contexts.

2 Related Work

2.1 Game Designing for Older Adults

Game design for older adults has received growing attention, with research highlighting the need for age-specific ergonomic adaptations to accommodate cognitive, sensory, and motor changes associated with aging. Simplified controls, larger interface elements, clear instructions, and customizable accessibility settings have been shown to enhance usability and reduce entry barriers [10]. Participatory design approaches have also been successfully implemented, involving older adults directly in the development process to improve engagement and usability [3] demonstrating that long-term co-design processes can lead to more accessible mechanics and clearer instructional feedback. Still, much of the existing work remains limited in scope, and there is considerable room to further explore how different genres, interaction styles, and design strategies can better support the diverse needs and preferences of older adult players.

2.2 Experience-Driven Game Adaptation

Traditional adaptive game systems primarily adjust difficulty and pacing in response to player behavior, using rule-based systems and biometric feedback [1, 5]. These approaches typically focus on observable performance metrics like success/failure or response time.

Ever since, supervised learning methods have expanded the design space by training models on labeled player data to predict engagement, enjoyment, or dropout risk [2, 9, 13]. For instance, facial expression analysis has been used to build

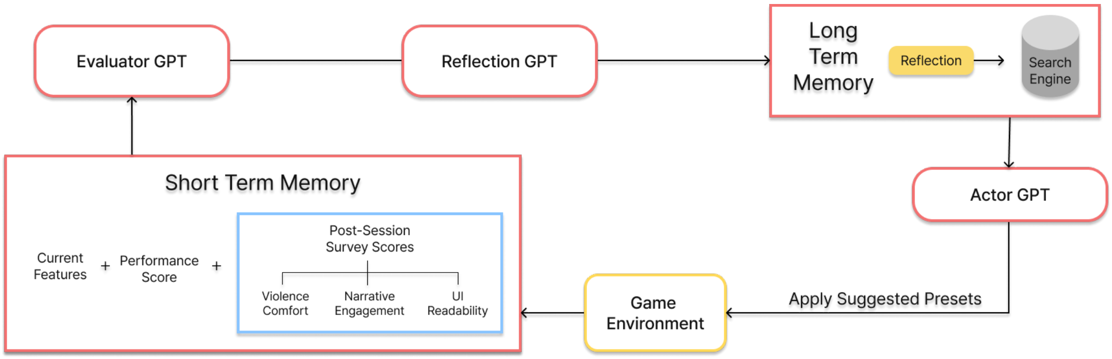

Fig. 1. Overview of Reflexion-Based Multi-Agent Adaptation Architecture

Random Forest classifiers that guide in-game difficulty adaptation [2], while linear regression has predicted gameplay preferences from player personality traits [9]. However, these methods depend heavily on large, labeled datasets—often derived from younger or experienced player populations—limiting their generalizability to underrepresented groups such as older adults.

Reinforcement learning (RL) has emerged as a promising solution for real-time adaptation without requiring labeled data. RL agents have been used to personalize AI behavior in fighting games [17], adapt level complexity in VR exergames [5], and even support therapeutic emotional regulation [4]. Yet RL-based systems are often narrowly focused on difficulty tuning and rarely offer explainable design recommendations for game developers.

2.3 Large Language Models and Adaptive Game Design

Recent work has explored the potential of LLMs in various domains of game design. In procedural content generation (PCG), LLMs have been used to generate puzzle levels and spatial maps via prompt tuning and feedback loops [11, 16]. Other systems combine LLMs with image generators to create entire game worlds, characters, and ambient settings from short textual descriptions [14]. In narrative design, tools like SCENCRAFT use LLMs to generate dialogue and scenes based on structured prompt templates [6]. While these efforts demonstrate the generative power of LLMs, they primarily focus on production-scale content creation rather than individualized gameplay adaptation. Few LLM-based systems engage with real-time player feedback or prioritize inclusivity for older adult players.

3 Methodology

3.1 System Overview

We present how multi-agent Reflexion-based adaptation system can support inclusive FPS game design for older adults through structured post-session feedback. Building on the Reflexion system [15], originally developed for language agents, we extend its key principles, modular agent roles, memory-based reasoning, and iterative reflection, into an interactive gaming context. Our system captures both real-time performance (task completion time) and post-session ratings (violence comfort, narrative engagement, UI readability), applying preset-based adaptations at session intervals rather than in real time, thus supporting inclusive design without requiring large-scale training data (see Figure 3).

The architecture integrates a modular Unity environment with a multi-agent LLM system consisting of three GPT agents (see Figure 1):

- **Evaluator GPT (GPT-4.1)** is invoked after each gameplay session to assess whether the previous configuration met the player's needs. It takes as input the player's real-time performance metrics and post-session Likert-scale feedback, and produces a binary success/failure label along with a natural language summary of the session outcome.
- **Reflection GPT (GPT-4.1)** is selectively triggered when a session is deemed unsuccessful. Drawing on the failed configuration, its outcomes, and similar historical sessions retrieved from memory, it generates explanatory insights and design recommendations, which are stored for future reference.
- **Actor GPT (GPT-4.1)** initiates the planning process for the upcoming session. Using the previous session's evaluation summary, optional reflection insights, and k-nearest memory samples retrieved from FAISS embeddings, it generates both a natural language rationale and a structured preset string (e.g., "12021") that defines the next configuration.

The preset string consists of predefined codes (0, 1, 2) corresponding to modular adaptation dimensions. These codes map directly to Unity prefabs and are used to reconfigure the game environment between sessions. Through this interaction cycle—evaluation, reflection, and planning—the agents collaboratively produce interpretable and personalized gameplay adjustments while maintaining session continuity.

3.2 Actor GPT: Preset-Driven Modular Design Output

Table 1. **Preset Codes by Adaptation Dimension**: Actor GPT modifies each game dimension by choosing from three preset codes (0, 1, 2) to customize the session configuration. These options include: *Difficulty Level*, which adjusts enemy health to modulate mechanical challenge; *UI Size*, which enhances visual readability; *Weapon Appearance*, which ranges from realistic to playful to reduce violent connotations; *Aim Assistance*, which enables or disables input precision aids; and *Narrative Thematic Style*, a bundled adaptation that coordinates enemy design, environment aesthetics, and cutscene tone to support cohesive emotional and thematic experiences tailored for older players.

Dimension	**Code 0**	**Code 1**	**Code 2**
Difficulty Level	Easy (2 HP)	Medium (6 HP)	Hard (10 HP)
UI Size	Small	Medium	Large
Weapon	Pistol	Water gun	Banana
Narrative Style	Realistic: human, construction, serious	Semi-Stylized: banana-man, city, mild humor	Comical: slime, forest, cartoonish
Aim Assist	Off	On	—

The system follows a preset-driven modular design strategy, where each digit in the preset string maps to a distinct adaptation dimension. Actor GPT generates these strings based on evaluation feedback and memory, encoding accessibility changes for the next session. Each digit corresponds to a Unity prefab module, allowing interpretable control over gameplay features (see Table 1 and Figure 2).

3.3 Signal Collection and Feedback Loop

The system operates through a continuous feedback loop driven by two primary input streams (see Table 2 for details):

- **Real-Time Performance Metrics** – Captured during gameplay to evaluate mechanical or cognitive difficulty. Each session includes a stage completion round, defined as the time elapsed from weapon pickup to successful enemy elimination, measured against predefined performance thresholds.

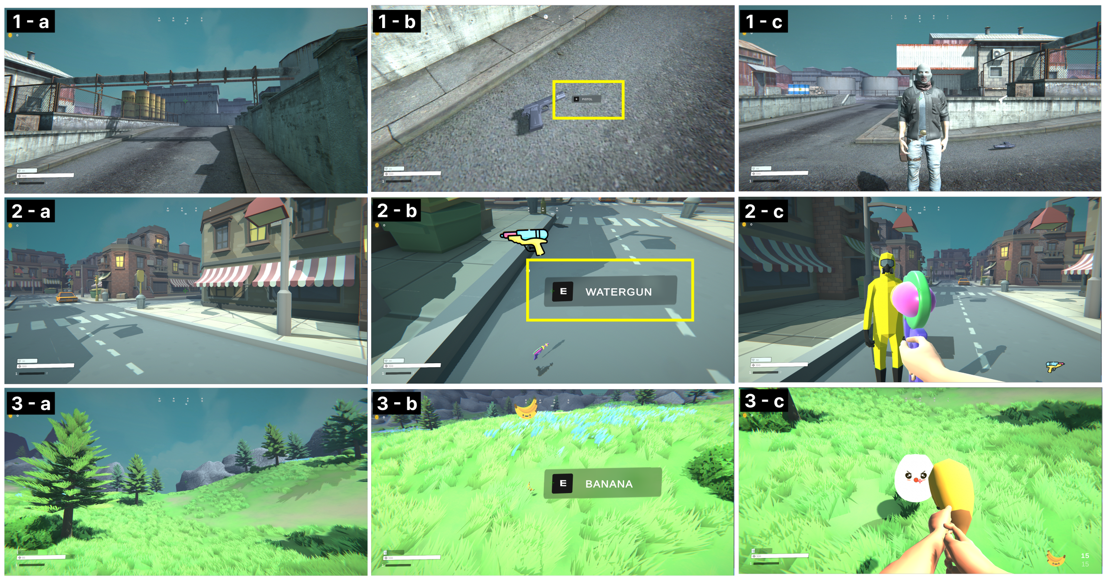

Fig. 2. **Modular Game Environment and Presets.** Each numbered row (1–3) corresponds to a Reflexion gameplay session, showing how modular adaptation dimensions were adjusted over time. Each column illustrates a different gameplay element: (a) the *in-game environment*, which changes from a realistic construction site (row 1), to a low-poly city (row 2), to a cartoon forest (row 3); (b) *Weapon Pickup UI Size*, indicated by a yellow outline that increases from small to large across rows; and (c) the *weapon and enemy appearance*, which shifts from a realistic pistol and human enemy (row 1), to a water gun and banana-man (row 2), to a banana weapon and comical slime enemy (row 3). Narrative Thematic Style is reflected across both (a) and (c), combining environmental context and enemy design. Each adaptation dimension is selected independently, and the figure visualizes how these elements evolved across sessions in response to player feedback.

Table 2. Objective and Subjective Evaluation Metrics

Type	Metric	Description and Scoring Criteria
Objective	Performance Score	Measures how quickly the player completes the task (seconds). **100:** under 10s; **80:** 10–15s; **60:** 15–20s; **30:** over 20s.
Subjective	Violence Comfort	Player's perceived discomfort with violent visuals or weapon realism. Rated from **1 (Not comfortable at all)** to **5 (Extremely Comfortable)**.
	Narrative Engagement	Degree of interest in the game's story, characters, and dialogue. Rated from **1 (Not engaging at all)** to **5 (Extremely engaging)**.
	UI Readability	Ease of reading text and navigating interface elements. Rated from **1 (Very difficult)** to **5 (Very easy)**.

- **Post-Session Evaluation Scores** – Collected using 5-point Likert scales to assess subjective perceptions of accessibility and narrative engagement. Players respond to structured questions designed to evaluate comfort, readability, and story interest.

These evaluation scores are processed by Evaluator GPT to determine whether the configuration met the player's needs, using the predefined scoring criteria. The outcomes guide Actor GPT's adaptation plans in future sessions.

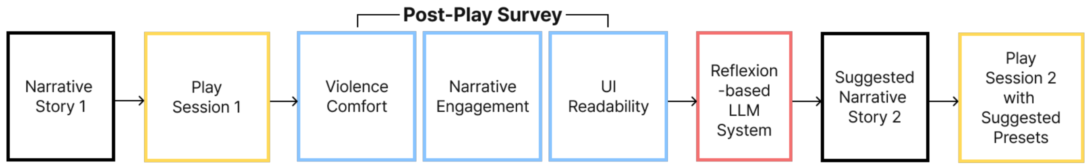

Fig. 3. **Feedback Loop: Data Collection, Evaluation, and Adaptation Cycle.** The system collects gameplay and post-play survey data from an initial session, analyzes it using a Reflexion-based LLM system, and generates tailored content and presets for the next session.

3.4 Evaluator GPT and Self-Reflection GPT

Evaluator GPT integrates real-time signals and post-session ratings to determine whether the current configuration met the player's needs. If marked unsuccessful, Self-Reflection GPT is invoked to analyze prior adaptation outcomes and identify patterns behind failures. It updates long-term memory with design recommendations and explanatory notes. These reflections serve as a foundation for improving future adaptations, reducing the recurrence of ineffective configurations.

3.5 Actor GPT with Memory-Driven Adaptation Planning

All session data—comprising performance metrics, evaluation scores, and adaptation outcomes—are embedded via FAISS for long-term storage and retrieval. Each memory entry is annotated with success/failure labels and insights generated by Self-Reflection GPT.

Actor GPT later performs k-sampling over these memory embeddings to retrieve relevant adaptation history. It uses this context to generate new configuration strings that align with past successes and avoid known failure patterns. This memory-driven approach ensures that adaptations are not only personalized but also historically grounded and progressively refined.

By completing this cycle of evaluation, reflection, and adaptation, the system offers a continuous, explainable, and trust-building gameplay experience tailored to older adults' evolving needs.

3.6 Participants and Experimental Procedure

We recruited four older adults aged 60 and above (2 males, 2 females; mean age = 62.75) to play games using a PC platform. All participants self-reported low PC proficiency levels and received brief control instructions from the interviewer before engaging in 5-6 gaming sessions each. Following the completion of all sessions, in-depth interviews lasting approximately 20 minutes were conducted with each participant to gather comprehensive qualitative data regarding their gaming experiences.

Table 3. Demographics of Participants

Participant	Sex	Age	PC Game Proficiency	Game Experience
1	Male	65	Low	Beginner
2	Female	63	Low	Beginner
3	Male	62	Low	Intermediate
4	Female	61	Low	Beginner

4 Results

4.1 Comparison of Reflexion's Reflection and User Feedback

To evaluate the effectiveness of the Reflexion-based system's adaptive strategies, we analyzed gameplay data and Self-Reflection agent outputs from four participants across four areas: performance, user interface (UI), violence comfort, and narrative engagement. After an initial session, all participants played under optimized settings, including reduced enemy health, minimized visual violence, enlarged UI elements, and auto-aim support. We then compared Reflexion's post-play reasoning with in-depth interview responses from the same participants.

4.1.1 Performance. Despite reduced enemy health and auto-aim, participants averaged a low performance score of 38. Reflexion suggested that core FPS mechanics—such as control schemes, reaction speed, and enemy recognition—posed persistent limitations. Interview responses echoed this: *"I didn't know how to pick up the water gun or banana... so I wandered around and wasted a lot of time." (Participant 1)*

4.1.2 UI Accessibility. All participants initially rated UI readability as 1 (very difficult). Following Reflexion's recommendation, weapon label text was enlarged. While one participant noted, *"I could tell the text... got larger." (Participant 2)*, most reported the change was too limited to improve the overall interface. Performance remained low, and the need to use both keyboard and mouse was still cited as a major obstacle. Reflexion concluded that while text enlargement aided visibility, it failed to reduce the overall cognitive load.

4.1.3 Violence Comfort. Participants recorded an average comfort score of 3.8, despite the use of abstract weapons and enemies with no blood effects. Reflexion noted that the nature of combat itself remained uncomfortable for some. One participant remarked, *"Instead of shooting a banana gun... maybe you could just throw the banana." (Participant 4)* This indicates a desire for deeper reimagining of interaction, not just visuals. Conversely, one participant's comfort score declined from 4 to 1, which Reflexion interpreted as desensitization—consistent with their report of minimal disturbance.

4.1.4 Narrative Engagement. Narrative engagement averaged 2.7 and remained low even after introducing new story and map designs. Reflexion inferred that difficulties with cognitive and controls overload hindered engagement. As one participant noted, *"The setting changed... I couldn't figure out what the story was." (Participant 1)* Another shared, *"Using both the keyboard and mouse... made it hard to pay attention." (Participant 2)* These comments suggest that cognitive and physical barriers disrupted participants' ability to follow the narrative, further compounded by tonal and contextual inconsistencies across sessions.

4.2 Key Findings by Research Questions

RQ1: These results demonstrate that Reflexion accurately identified effective and ineffective accessibility features for individual participants by analyzing applied accessibility features, performance scores and post-session evaluations. When features proved ineffective, the system provided targeted design recommendations that aligned with participants' reported difficulties and suggestions from interviews. Most participants achieved low performance scores despite UI size, auto-aim, and reduced enemy difficulty due to control complexity. Motor and cognitive challenges also limited narrative immersion improvements, indicating that older adults need simpler control schemes. Violence reduction required fundamental design changes to weapon interactions and core gameplay mechanics, beyond superficial modifications. These insights provide valuable guidance for designers creating accessible games for older adults, highlighting the need for comprehensive accessibility consideration.

RQ2: Despite modest post-session evaluation scores, participants held positive views in interviews about the adaptive system's potential to enhance both playability and satisfaction. In particular, those who previously avoided FPS games due to violence reported experiencing partial mental comfort when violence settings and thematic styles were immediately adjusted. While some suggested that fundamental changes to weapon interactions would be beneficial, they expressed optimism that this adaptive system could make FPS games more accessible to them. Beyond violence adaptation, participants emphasized the value of personalization. They noted that tailoring narratives and thematic styles to individual preferences could significantly improve their gaming satisfaction and engagement. These findings indicate that real-time personalized adaptive systems hold considerable promise for enhancing playability and satisfaction among older adult gamers.

5 Discussion

5.1 Key Insights and Implications for Designers

The most significant insight from our findings is that accessibility for older adults must encompass both physical and emotional comfort, not just mechanical assistance. While there has been prior research on adaptive systems that adjust gameplay difficulty [1, 2, 5], our results show that such approaches alone are not enough. Participants who previously avoided FPS games due to violence showed increased playability when the system adapted aesthetic and narrative elements alongside gameplay mechanics. This challenges conventional accessibility models that rely primarily on technical tuning, and highlights that psychological comfort and narrative familiarity are just as important. By providing individualized recommendations grounded in player feedback and reflective reasoning, the Reflexion system enables designers to understand why specific adaptations work—supporting more meaningful and effective design decisions than one-size-fits-all solutions.

5.2 Limitations and Future Work

Despite these contributions, the study presents several limitations. First, the system was constrained to predefined design modules, preventing more flexible adaptations. Second, the small sample size (n = 4) limits generalizability, lacking statistical power to validate broader claims about system effectiveness across diverse older adult populations. Nevertheless, as this study aimed to explore the feasibility of Reflexion-based adaptive systems using multi-LLM agents, these preliminary findings provide valuable groundwork for future research.

Future research should address these limitations through several key directions. First, developing more flexible design capabilities that modify gameplay mechanics beyond preset options. Second, improving Reflexion's reasoning by incorporating domain-specific knowledge from game design theory and age-related cognitive research. Third, conducting larger-scale studies to provide rigorous statistical validation across broader older adult populations.

6 Conclusion

Our Reflexion-based multi-LLM agent system demonstrates the feasibility of effective accessibility design for older adults—a previously underserved population in gaming—using minimal data to generate meaningful adaptations. Through enhanced design flexibility and broader, rigorous user validation, we aim to validate this approach and extend it as a foundational tool for diverse underrepresented gaming populations.

References

[1] Gustavo Andrade, Geber Ramalho, Hugo Santana, and Vincent Corruble. 2005. Extending Reinforcement Learning to Provide Dynamic Game Balancing. In *Proceedings of the 2005 ACM SIGCHI International Conference on Advances in Computer Entertainment Technology*. ACM, 149–154. doi:10.1145/1178477.1178504

[2] Kurt Blom, Matthias Lankes, and Martin Martsch. 2020. Modeling and Adjusting In-game Difficulty Based on Facial Expression Analysis. In *2020 IEEE Conference on Games (CoG)*. IEEE, 1–8. doi:10.1109/CoG47356.2020.9231609

[3] Eirik Brox, Efstathios I Konstantinidis, and Gunn Evertsen. 2017. User-centered design of serious games for older adults following 3 years of experience with exergames for seniors. *JMIR Serious Games* 5, 1 (2017), e1.

[4] Ronja Fuchs, Robin Gieseke, and Alexander Dockhorn. 2024. Personalized Dynamic Difficulty Adjustment – Imitation Learning Meets Reinforcement Learning. In *2024 IEEE Conference on Games (CoG)*. IEEE, IEEE, Osaka, Japan. https://arxiv.org/abs/2408.06818 Preprint available on arXiv.

[5] Thomas Huber, Sebastian Mertes, Svetlana Rangelova, Spetim Flutura, and Elisabeth André. 2021. Dynamic Difficulty Adjustment in Virtual Reality Exergames through Experience-driven Procedural Content Generation. In *2021 IEEE Symposium Series on Computational Intelligence (SSCI)*. 1–8. doi:10.1109/SSCI50451.2021.9660086

[6] Vikram Kumaran, Jonathan Rowe, Bradford Mott, and James Lester. 2023. SCENECRAFT: Automating Interactive Narrative Scene Generation in Digital Games with Large Language Models. In *Proceedings of the AAAI Conference on Artificial Intelligence and Interactive Digital Entertainment*, Vol. 19. 86–96.

[7] Byungjoo Lee, Young Yim Doh, Kyung Myun Lee, Junyoung Shin, and Minseok Do. 2022. Game Design Guide for Adults in Their 50s and Older. https://wikidocs.net/155599. Contributors include Seyeon Lee, Injung Lee, Seokbeom Park, Hajun Kim, Jinyoung Lee, Hyunwook Lee, Taeeun Kim, Hae In Kim, Jongchang Park, Kyunghyun Lee, and Hyuna Cho. Last updated: June 3, 2022.

[8] Eddie Makuch. 2021. *Time And Money Spent On Games In The US Jumps In 2021 Thanks In Part To Boomers*. https://www.gamespot.com/articles/time-and-money-spent-on-games-in-the-us-jumps-in-2021-thanks-in-part-to-boomers/1100-6496866/?ftag=CAD-01-10abi2f Accessed on 2025-05-02.

[9] Frank Nagle, Colleen M. Wolf, Jurek Kirakowski, and Christine Murray. 2016. Towards a System of Customized Video Game Mechanics Based on Player Personality: Relating the Big Five Personality Traits with Difficulty Adaptation in a First-person Shooter Game. In *2016 IEEE International Conference on Serious Games and Applications for Health (SeGAH)*. IEEE, 1–8. doi:10.1109/SeGAH.2016.7586295

[10] Henk H. Nap, Yvonne AW de Kort, and Wijnand A. IJsselsteijn. 2009. Digital Game Design for Elderly Users. *Proceedings of the 2009 ACM SIGACCESS Accessibility and Computing* 93 (2009), 17–22.

[11] Muhammad U. Nasir and Julian Togelius. 2023. Practical PCG Through Large Language Models. *arXiv preprint arXiv:2305.18243* (2023). https://arxiv.org/abs/2305.18243

[12] Brittne Nelson-Kakulla. 2019. *2020 Gaming Trends of the 50+*. Technical Report. AARP Research, Washington, DC. doi:10.26419/res.00328.001 Published December 2019.

[13] Jannik Pfau, Stefan Göbel, and Ralf Steinmetz. 2020. Enemy Within: Long-term Motivation Effects of Deep Player Behavior Models for Dynamic Difficulty Adjustment. In *2020 IEEE Conference on Games (CoG)*. IEEE, 127–134. doi:10.1109/CoG47356.2020.9231675

[14] Ahad Shams, Douglas Summers-Stay, Vsevolod Metelsky, Arpan Tripathi, and Karan Malhotra. 2023. Towards a Holodeck-style Simulation Game. *arXiv preprint arXiv:2308.13548* (2023). https://arxiv.org/abs/2308.13548

[15] Noah Shinn, Federico Cassano, Edward Berman, Ashwin Gopinath, Karthik Narasimhan, and Shunyu Yao. 2023. Reflexion: Language Agents with Verbal Reinforcement Learning. arXiv:2303.11366 [cs.AI]

[16] Graham Todd, Sam Earle, Muhammad Umair Nasir, Michael Cerny Green, and Julian Togelius. 2023. Level Generation Through Large Language Models. In *Proceedings of the 18th International Conference on the Foundations of Digital Games*. 1–8.

[17] H. Wang, Y. Gao, and X. Chen. 2010. DRL-dot: A reinforcement learning NPC team for playing domination games. *IEEE Transactions on Computational Intelligence and AI in Games* 2, 1 (2010), 17–26.