The Wizard Speaks in Tensors

Student ID: 18025652

Joel Falconer
SUPERVISOR: M J Hunter Brueggemann

Table Of Contents

Abstract

Introduction

Related Work

Methodology

Project Write Up

Evaluation

Discussion

Conclusion & Future Work

Bibliography

Abstract

There is a long-established precedent for the evolution of immersive role-playing experiences, from the early tabletop systems like dungeons and dragons through to the first digital RPGS and into the modern-day digital role-playing games, this project builds the trajectory into the next stage of roleplaying. In this project I aimed to build a prototype demonstrating a fully reactive LLM-powered NPC companion. With it I planned to explore if intelligent system design using dynamic memory + narrative rules can make the NPC feel consistent, coherent, and emotionally grounded to see whether players experience increased engagement and emotional resonance when the NPC recalls events and adapts dialogue and If players do resonate with the character what roll they see AI playing in the future of roleplaying interactive experiences in games. To build my prototype I used Unity (Game engine) for building the CRPG Epoch Zero: Data Heist, Custom systems: Game Events Manager, memory and lore Libraries, payload assembler, dialogue UI for AI interface and a shallow combat slice. I used a fine-tuned Mistral 7B model that I merged and deployed on hugging face with its own inference endpoint. I also created a memory, lore and greater context schema for RAG and rule blocks to further ground the model. In the game you play as Zero a hacker/operative who is partnered with a former loader bot who has gained sentience called C4554NDR4 (Cassandra) the two share a sibling like bond and try to keep each other safe in the city of Carnus, a colony on Jupiter's moon Europa. The demo is designed to focus on the basics of a CRPG: Quests, combat and dialogue and see how these pillars can ground an AI NPC. I have built a Memory-aware LLM pipeline: Events happen in game, they are turned into memory entries and added to a memory library, then through a search algorithm said memory is added to a payload which is passed to the model at inference. The character's personality is formed from a dataset of seven thousand lines that encode emotion, persona, and machine-human dynamics for narrative flavour. The model responds in real-time via a unity coroutine and Hugging Face API. Early playtests showed strong perceived presence and emotional engagement. Issues observed: response repetition, occasional rule drift and temperature sensitivity causing dialogue drift, latency (note: this was an issue while using a development server on Google Colab). These thus far showed the potential and limitations of LLM NPCs. This shows how an LLM grounded by intelligent system design can enable responsive NPC companions. And this project provides a framework for future adaptive storytelling.

Introduction

In recent years there has been a rise in interest for AI in video games, whether it be plain text to speech AI voices or be AI driven NPCs (Non player characters), there is a clear focus on

implementing this technology in responsive environments. There has always been a cultural fascination with fictional characters even before interactive media. Since we first put pen to paper audiences have found themselves forming parasocial connections with fictional characters in film and literature. In 1974 with the release of the first edition of Dungeons and dragons marked the first mainstream system to formalize roleplaying mechanics and stats with improvised storytelling. This was the first step away from static authored storytelling. Then again in the late 70s and early 80s the first wave of digital CRPGS was released with titles such as Akalabeth (1979), Ultima (1981) and Wizardry (1981) these titles built upon the foundation created by DND using the same stats, mechanics and roleplaying systems and were the framework for future digital titles to come like KOTOR(2003), The Witcher 3(2015) and Baldur's Gate 3(2023) across these titles marked key waypoints in the evolution of cinematic storytelling, reactivity, and branching narrative design in RPGs. All these titles were pre authored dialogue experiences which means despite wide branching narratives, still limited reactivity due to the constraints of the technology. However, with the development of LLMs marks the potential rise of truly emergent story telling with the possibility of dynamic memory-aware NPCs.

My project Epoch Zero is a CRPG vertical slice set on Jupiter's Icy moon Europa. You play as Zero, a young engineer sent by anonymous brokers to infiltrate and steal corporate data. Alongside you is Cassandra or C4554NDR4, an LLM-driven companion whose dialogue, personality and emotional context evolve through a custom memory system. As players choose between multiple approaches to a heist, Cassandra stores key narrative events and retrieves them through an embedded RAG system to shape uniquely contextual, dynamic responses.

This work matters because generative AI is rapidly entering every creative field including games, yet its role remains undefined. For decades, games have offered rich worlds and unforgettable characters, but those interactions have always been one directional. Epoch Zero asks: What happens when characters talk back? How does real-time, context -aware responsiveness reshape narrative, exploration, and emotional connection? Does it deepen immersion or risk new forms of isolation in a world already struggling with it? The project invites players to grapple with these emerging questions firsthand.

Players will engage with the piece through a traditional mouse and keyboard set up. The vertical slice is intentionally short and highly repayable and designed for experimentation. Players can try different narrative paths, evaluate Cassandra's memory, and observe how her dialogue shifts in response to their choices. The aim is not just to complete the heist, but to explore whether this AI-driven interaction feels meaningful, disruptive, novel, or unnecessary. The player's own reactions: curiosity, attachment, discomfort, fascination are part of the outcome the work seeks to surface.

In contributing to the field, Epoch Zero demonstrates that generative AI is not a creative replacement but a subsystem. The model only becomes compelling when supported by strong systematic design (memory architecture, event logic, retrieval algorithms, and authored narratives). It argues for responsible system integration of AI and asks the question Can players emotionally resonate with an AI simulation of a fictional character.

Related works

Baldur's gate three is a CRPG (Character Role Playing Game) set in the forgotten realms universe, the canon dungeons and dragon's setting. It was released on the third of august 2023 by Larion studios. In 2022 I played the early access build of the game and after its full release I have put over one thousand hours into it and continue to play it to this day not just for academic research. I am familiar with this title and have first hand experience with its companion narrative driven system design.

What stood out to me with this game is its characters, specifically its companions. In my own playthroughs I spent hours in dialogue with them, with branching paths and consequences, there are many permutations of what relationships you as the player can have with them. I've many times began entire playthroughs with specific companions in mind and what dynamics I aimed to have with them whether it be platonic, romantic, or antagonistic. This is because in addition to the main plot of saving the world each character is given their own ark and storyline. It is because of these systems that Baldur's gate has succeeded as much as it has.

This design was deliberate as the developers believed it was the core of the game's ability to keep player retention forty plus hours into a game. "The companions are the heart of the game. Your relationship with them is what will make or break the story." – Sven Vincke This is true as the through line for each playthrough is what happens with your party members and concludes with wrapping up their individual story arcs. Critically this game's companions were the centre of praise for most mainstream outlets. From IGN's review of the game "Baldur's gate 3 thrives because of its companions – some of the best written characters the genre has ever seen." – (Hafer, 2023) another from PC Gamer "The companions steal the Show. Their writing, performances, and reactivity are astonishing" – (Brevig, 2023). Critics repeatedly praise the characters reactivity and performances as responsible for the success of the game. Aside from critical praise players were found to spend hours talking with their companions. Larian Studios reports a median playtime of 62 hours for PC players, indicating that many invested over one hundred hours in the game. "The median time played for BG3 on steam is 62 hours and 55 minutes… which means those of you clocking in over 60 hours are just scraping average" – (West J,2023)

Baldur's gate three stands currently as the modern benchmark for CRPGs and its established stand out feature, its companions and their reactivity, was integral to its success. Reviews and player behaviour indicate that much of the game's playtime is spent interacting with these characters. LLMs excel at the key features these characters have been praised for: reactive dialogue, contextual recall, varied responses. These map directly onto what CRPG companions are designed to do so introducing an LLM into a companion role is only a natural extension of the genre's existing design patterns. Baldur's gate 3 already serves as a reference point for success in the medium using authored characters, so introducing LLMs into a CRPG like Baldur's gate would be an excellent test .This is why my vertical slice uses a similar format intentionally as a control environment with the model being the unknown variable introduced.

The purpose of this would be to evaluate if the LLM-driven companion would enhance or disrupt the player experience.

when I began looking into this subject, I came across Professor Sherry Turkle, A sociologist, and licensed clinical psychologist at MIT. Her focus is primarily on how digital tools and platforms from early personal computers to artificial intelligence affects us psychologically and socially. The work that I was particularly interested with for this study was Alone Together (2011) despite it being written in a time before AI it could not be more relevant.

In this work she makes 2 arguments that were pertinent to my work, the first being about projection and companionship, she argues that humans are very vulnerable that we are lonely and crave intimacy and emotional resonance, yet we stray further from said intimacy with other human beings. However, in the digital age she speaks about the 'Sociable Robot' which can range from a Furbee doll to a Tamagotchi, these become vessels for simulated companionship which doesn't come with the complexities or demands of true friendship (Turkle,2011) "Technology is seductive when what it offers meets our human vulnerabilities. And as it turns out, we are very vulnerable indeed. We are lonely but fearful of intimacy. Digital connections and the sociable robot may offer the illusion of companionship without the demands of friendship." This leads into an ethical dilemma, just as we have seen people become emotionally and para socially attached to fictional characters from static fiction, we have also observed similar behaviour with early computer programs and toys due to these machines offering simple responsiveness creating the illusion of connection.

The second argument is the reason we tend to offload emotional weight on to machines. Machines Cannot reject us;(Turkle,2011) argues that "We expect more from technology and less from each other" they do not negotiate with us the terms of our relationship with them. When you play Baldur's Gate 3, or any other CRPG, when a character disagrees or shows discomfort this isn't the same as a partner or friend who is discontent with what we are doing in a relationship it instead simulates just enough consequences to feel emotionally charged within the game but not enough to force players to reflect upon on or negotiate the real emotional labour or complexity a human relationship would demand. (Turkle,2012) observes that "human relationships are rich and they are messy and they are demanding. And we clean them up with technology."

Turkle's concerns about social robots comes way before the rise of popular LLMs. Her arguments refer to a far less complicated forms of sociable robot. LLMs are far more capable of providing emotional plausibility and simulation. This is the ethical concern when designing a framework like this what Turkle described as "illusions of companionship" have become harder for users to distinguish. This project's model is specifically designed to be emotionally resonate, to recall a shared history with the user and maintaining a stable identity. This makes her far closer to a safe relational space that Turkle warned about. The question becomes when does NPC immersion become a substitute for emotional connection. The purpose for including this work in this thesis is to understand the ethical stakes behind such a system. This project is not meant to damn this technology but to surface issues related to isolation and artificial intimacy.

Park et al.'s Generative Agents: Interactive Simulacra of Human Behaviour (2023) presents a multi agent simulation that sets out to see if LLMs can store memory, reflect on them and then act autonomously.

There are four systems in this project that influenced my design. The first being the Memory stream, functionally this works as continuous storage of every experience, with each being time stamped and in natural language. This forms the foundations of the agent's episodic memory. With this memory the agent uses the second system reflection, where it periodically condenses memories and produces short top-level insights ("I enjoy talking to x," "I should practice music more"). That rolls into planning, which uses reflections plus the recent events that were logged to generate both short term and long-term goals. These goals are then used to create schedules ("At 4pm I will talk to x"). This then is comprised into a full behaviour loop where agents act according to their plans, those actions then produce new memories. This cycle continues giving agents continuity and personality.

This work demonstrates that LLM-Driven NPCs cannot rely on powerful models alone they require scaffolding through (Memory, Reflection, planning) to remain coherent. This project uses aa similar loop with Memory, retrieval, and contextual reasoning to maintain coherence and memory over time. While the Sanford project used a multi agent village ecology this project uses one NPC as a companion, it is the same idea instead within a CRPG dialogue system. The purpose of using this study is that it provides an already proven architecture for how a system for agent behaviour would work and it shaped this project conceptually.

Methodology

When I began I started with Gpt2 small and later Gpt2 Large because I initially wanted local inference however through much iteration these legacy models proved to be computationally inadequate for a conversational agent, it could catch the cadence of the character but it would actively spew poetic nonsense, not game worthy dialogue, the latency would be too high for real time interaction as well as my 3050 RTX GPU would struggle to keep a decent framerate as well as run generation. In the end I chose Mistral-7B-Instruct + Lora because even at small scale it has a strong performance meaning that when it comes to finetuning the costs are feasible. And due to Mistral-7B being pretrained on dialogue it already can produce quality dialogue and with its modern instruction tuned architecture it would interpret my RAG schema as clear instructions. Because of the size of models, I opted for cloud inference as for the stated reasons above local inference is infeasible. With a dedicated endpoint it off loads computing power during gameplay and ensures interactive latency. Because of this the model is now decoupled from hardware constraints.

I chose unity for my game engine as the projects goals is not graphical fidelity it is interaction design. Unity is fast for prototyping UI design, triggers, and basic RPG systems. Unity also has built in coroutines which is optimal for asynchronous model calls. I chose a CRPG structure because the genre relies heavily on companion dialogue, branching narratives and player immersion. This makes them a natural testbed for LLM driven character interaction. The demo

would be short so players would be able to quickly loop through the different paths and experiment with the model's ability keep memories of what had transpired in their playthroughs.

When it came to fine tuning the model I constructed the data set in layers to address each behavioural requirement. The first being the foundational voice which is the characters tone, cadence, emotional resonance. The second being the personality layer which includes her moral stance, emotional baseline, and core beliefs. The third was the relational dynamics such as the older sibling dynamic with the protagonist (Zero) as well as the characters machine / human contrast with those around her. The fourth would be a series of specialised modes that would trigger with different phrasing from player prompts. These modes were tactical reasoning if she was asked about quest data or combat tips, emotional grounding so she could have emotionally grounded dialogue if the player's prompt was phrased in an emotional context, philosophical reflection , whenever the character is asked about their beliefs or given existential queries , wit and humour to provide levity to a situation or to be used as a segway from one tone to another and world building and lore so the model has a baseline understanding of where and who she is. The reason for these layers is to prevent tonal collapse, so the character has a strict hierarchy of modes to fall back on to avoid drift.

The reason for the structured context is that LLMs perform far better with structured relevant context, without this scaffolding models tend to drift or hallucinate and cannot maintain character identity or current canon. This structured context was inspired by the engineering principle of retrieval augmented generation and the Stanford "AI agents" paper. The system is designed around four contextual pillars : A transcript for short term memory that will keep conversational continuity, a memory library that allows key events to persist across gameplay and ensure the model sticks to a single canon to enable player-model relationship building, Lore injection to keep stable world facts to reduce hallucination risk and preserve world continuity and objective state context so the model understands player progression. I had to make decision between a dynamic context that essentially would have the entire context shift every new input or a sticky context that would preserve a context over multiple turns and can only be changed if a different memory was more pertinent then the original, I opted for the latter so conversations could last longer and discussion of a single topic could get deeper. This would be achieved with a high-level scoring mechanic. I also included a governance layer in the schema; this would be a block of rules that would give the model constraint. Rules like "Make sure to reference memory and lore as is do not embellish" or "answer in 3 sentences or less". This would be to regulate token count and add another guard rail against drift.

The methodology is shaped around evaluating whether an LLM companion can sustain emotional engagement, maintain narrative continuity and preserve a stable character identity throughout play. to investigate these aims the system incorporates:

- Layerd fine-tuning to maintain characterisation.
- A multi-component context architecture to preserve relevance and world knowledge.
- A CRPG interaction environment to evaluate reactivity under narrative pressure.
- And a behavioural ruleset to constrain the model and prevent drift.

Build Write-up

As referenced in the methodology, the project began with GPT-2 however these legacy models were abandoned for these reasons; after 4000 lines I could no longer train it locally which was the reason I wanted to use a legacy model and even when moved to Google Colab to train it after 4000 lines the model showed no improvement as the validation loss plateaued and still produced poor dialogue quality. The model was then migrated to mistral-7B-Instruct-V.0.3. This model was an instant success as it immediately could produce game ready dialogue, this is due to mistral already being pretrained on dialogue, so the dataset provided was simply training the model on who the character is not how to speak. Mistral performed well when served remotely through an inference endpoint. The data set followed the same layered structure as stated in the methodology with each one thousand sample block training the model on a different aspect of the characters personality. Across each cycle of training the model, this process produced seven thousand lines. These lines would be a player prompt followed by the expected completion from the model with character tags for both the protagonist (Zero) and the Companion (C4554NDR4). Once the model was evaluated, using Mistral, the RAG context schema as mentioned in the methodology was introduced and proved successful in injecting context into the model's responses. After this the Lora weights were merged with the base model to create a single safe tensors package. During integration with unity the model was assessed using a google Colab flask server but later would be deployed on Hugging face's platform.

To simulate agent like memory and situational awareness a modular context system was developed:

- Transcript system: a rolling window of recent dialogue.
- Memory Library: a collection of memories tied to major events.
- Lore Library: a collection of static world facts and definitions
- Game state: a class object holding all stages of the current quest.
- Rules Block: behaviour constraints for the model.
- Sticky context: dominant memory that persists.

The Assembler script is a JSON assembler that constructs the final context packet. The payload includes:

- Transcript (last 4-5 lines)
- Selected memory
- Selected lore
- Current objective
- Rules
- Player input

A unity coroutine formats the request into the required messages array and parameters block for Hugging face inference. Then it sends the request via UnityWebRequest with the Bearer token and JSON body. This also manages the clean up when displaying the model's response on the in-game UI.

To find relevant lore and memories I designed a retrieval algorithm. The player's input is split into "Uni tokens" and "Bi tokens" the former for single word tags the latter for 2-word tags. Each memory entry or lore entry has a list of tags, for each match points are added to the relevance score of each memory entry or lore entry.

Point contributions:

- Uni tokens + 1
- Bi tokens + 1

Every matched tag is added to a unified token set array; this array is then compared against the canonical data base. This data base contains 3 lists: Names, Locations and Factions depending on which list the token matches the weight of the score added changes.

Canonical weight:

- Name + 3
- Location + 2
- Faction + 2

So, each entries relevance score is the initial token score + the canon bonus score. All entries are ordered in descending order by their relevance score. For lore, the top two entries are selected. For memory, the singular top entry is chosen to be matched against the sticky context, if sticky context is empty and the top entries score is higher than the selection threshold (3), it will become the current context. If sticky context is not empty the top score will be compared against current sticky context's score if its greater it will replace the current context if not it is discarded.

The game follows a tried-and-true CRPG architecture. For movement characters use NavMesh components including the player as it is a point and click traversal system using ray cast hits. The companion character is programmed to follow the player character with light weight steering and obstacle avoidance.

The enemy characters would use the same NavMesh movement system with minimal state logic. Each enemy would have accessed to an array containing the player and companion characters called "partyMembers" if the enemy were hostile it would search for party members in its detection radius if a member were found the enemy will move into range for an attack. While not hostile they were interactable, if clicked on it would open a dialogue panel for pre authored dialogue. This low complexity design supports the needs of combat without detracting from the focus of the LLMs integration.

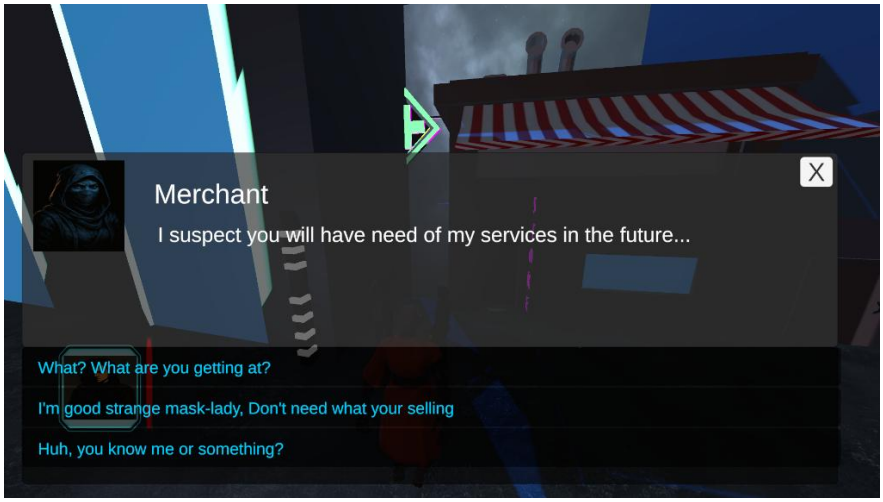
The inventory system is quite simple, you have storage for any item you pick up across the game, and you have equipped slots for equipment you wish to use such as weapons and armour. Items are broken down into two categories: Equipment and quest items. These items use Scriptable Object-based stat definitions. Item pick up is also tied to the memory system with quest items having their own trigger that will update the model's memory.

Quest items:

- Skeleton Key
- Abell Pass card.

Equipment:

- Zero’s Pistol
- Abell Body Armour



The dialogue system with regular NPCs uses a dialogue data structure using scriptable objects defining:

- NPC lines
- Player dialogue options
- Branching outcomes
- Memory triggers
- Quest progression flags

The dialogue interface also includes:

- NPC portraits
- Dialogue box

Dialogue for Cassandra is different; she has her own dedicated inference interface which provides a clean simple UI that shows the following information.

- Current context for memory and lore
- Transcript
- Input field and button.

- Portrait of Cassandra
- Dialogue box to show Cassandra's response.

The game uses a simplified DND 5th edition stat system using basic stats like:

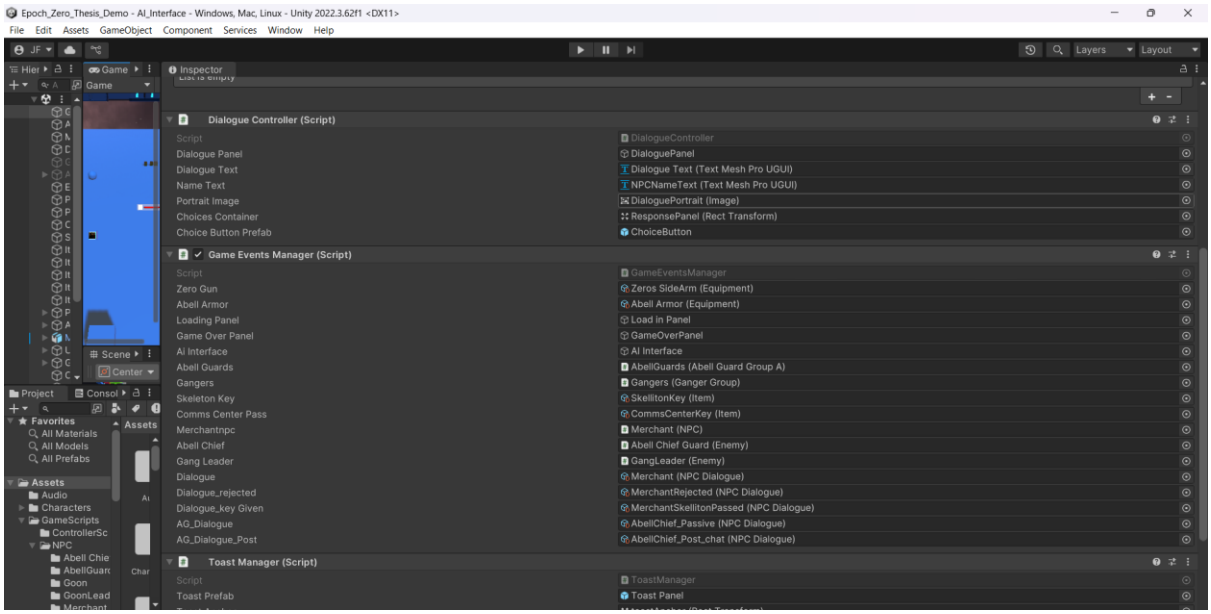
- Current health
- Max health
- Armour class
- Damage
- Attack
- Range

These are linked to:

- Player's health bar using Unity's "Image.fillAmount."
- Enemy health bars that appear when targeted
- Floating damage

Combat is intentionally simple, it follows the DND loop of characters rolling some dice to hit, if their roll is higher then their target's armour class(integer) the target takes damage if not then the attack misses. This attack roll can be augmented using certain equipment giving combatants a better chance to hit. Unlike DND it is not turn based instead characters have an attack cool down before they can make another roll. for the player, the loop is simple if the cool down is finished, they click right on their mouse while hovering over their intended target and roll for an attack. This plays the attack animation, and the attack button is then put on cool down for 2 seconds to prevent spam.

The Game Manager coordinates all game events handling:



- Dialogue based memory additions.
- Enemy death memory triggers
- World traversal memory triggers
- Updating dialogue data for NPCS
- Setting groups of enemies to hostile
- Ending demo terminal interaction

These events use and interact with:

- Memory Library
- Lore Library

Once the core systems were functioning, I used assets from two unity Environment packages and generated place holder character models using Rodin/Hyper3D AI for world polish. I then did a sound pass, added an ambient track for tone, and adjusted the directional light to a blue hue to reflect Europa's icy surface. I added punch and gunshot SFX using Unity's AudioSource, with 3D spatialisation.

For final integration I moved the model to a permanent endpoint on Hugging face, once the LLM was evaluated end to end on a permanent server I created an executable for the project ready for play testing.

Evaluation

For evaluation I deployed the build as a stand alone executable. I selected seven participants from a gaming server on discord. This meant that all participants were familiar with the genre they also all have played dungeons and dragons and participate in role playing games as a hobby on a regular basis. All participants would download the build and conduct two full play throughs to experience both branching paths. They were instructed to check in with Cassandra after every development and share their thoughts on the mission. Afterwards two participants were selected for an informal interview where I would obtain qualitive data.

Participant one

Immersion:

- Immersion was maintained.
- Occasional repeated responses caused brief breaks, attributed to model/implementation rather than character inconsistency.

Contextual Awareness

- Showed strong awareness of recent events.
- Struggled to recall earlier events in the session.

Team Functionality

- Felt like an actual party member, not just an advisor.
- Provided decisive, justified opinions when asked how to proceed.

Stand out Intelligent Behaviour

- Suggested hacking a forcefield to bypass a barrier even though this action was not possible.
- Participant viewed this as potential emergent problem-solving.

Overall experience / Immersion impact

- Enhanced immersion overall

- Participant noted immediate situational awareness for example the current quest objective, this increased the feeling of companionship.

Future Potential in CRPGS

- Strong enthusiasm for full-scale implementation
- Felt LLM companions represent a major step forward in narrative immersion.

Participant two

Immersion

- No major immersion breaks reported.
- The companion referring to itself as an AI did not break immersion because the participant interpreted it as being consistent due to it being a robot character.
- There was no looping behaviour reported.

Contextual Awareness

- Reported strong immediate awareness of what just happened.
- When asked about recent events or objectives they reported timely and appropriate responses?
- Reported examples of the companion predicting what the next steps of the objective may be, suggesting forward awareness.
- However, there were moments of overstating capability: stating they could open a forcefield but not yet having the required item.

Team functionality

- Companion was useful for navigation and early game advice.
- Participant framed companion as guide through the game.

Standout Intelligent Behaviours

- Responded naturally to an unscripted joke, showed humour, then pivoted back to mission context.
- Claimed it could open a door but later acknowledged limitations. Perceived as an interesting example of awareness of overstating capabilities.

Overall experience / Immersion impact

- Participant noted that traditional pre authored characters eventually become repetitive due to finite recorded dialogue.
- The ability to speak freely and still receive contextual responses is a major improvement in that aspect.
- Overall felt further immersed.

Future Potential in CRPGS

- Participant is interested but very sceptical.
- Curios how an ai companion would maintain coherence over a longer narrative.
- Also highlighted the potential for unprompted natural conversation as a major advantage for investment in characters.

Discussion

From the feedback given, both participants claimed that their immersion was not broken, that reactivity felt surprisingly natural and that Cassandra stayed within the premise of being Zero's robot companion. However, where the limits seemed to appear were occasional repetitious loops. The model felt responsive in the moment, but players felt the character was not necessarily deeply remembering. From this we can see that the persona layers responsible for short term coherence worked as well as the transcript and sticky context benefitted Cassandra's performance here greatly. However, due to a lack of a more robust memory stream like the one mentioned in the simulacra Stanford paper the model struggled to retain older memories that were not tied to key branching events. To answer the project's aim of evaluating whether players could achieve emotional resonance the answer is only situationally and is heavily dependant on the scaffolding around the model.

Both players initiated roleplay and did not require UI prompts to get them to engage with Cassandra; They asked for feelings, they asked for advice and sought tactical options. Because of Cassandra's tactical layer as mentioned in the methodology, she gave direct tactical insight that shaped decisions like a squad mate. Players treated her as a teammate, someone who contextualised events, interpreted objectives, reaffirmed quest information, and guided early navigation. But emotional roleplay remained shallow, players never described a bond forming, Cassandra was seen as a useful gameplay tool or a reactive support character, even a co-player rather than a companion with developing emotional depth. This is likely because players could engage in a form of resonance and roleplay with the character due to her always responding in a safe manor without any emotional cost. However, this created engagement not intimacy. Because Cassandra had reactive intelligence, but she lacked long term planning or evolving motivations. Because of this they did not see Cassandra having her own ark or inner life that they could truly invest and roleplay emotionally with. To answer the second aim, what role do they see AI having in the future of gaming, systems like this have the potential to one day produce characters players will bond with over long narratives but only if the model is integrated with a system that can give it the ability to reflect on its self, the player and the world to then create its own motivation and ending.

Conclusion

In conclusion both participants demonstrated that players are willing and able to engage with AI-driven companions. Because Cassandra was reactive, contextually aware and had a stable voice players treated her as a legitimate world entity. This supports the core hypothesis whether and AI driven NPC could get players to meaningfully engage with it in moment-to-moment interaction within a CRPG.

Resonance is possible as both players projected personality on to Cassandra and enjoyed the responsiveness. However, resonance remained situational and not sustained. Following Turkle's argument we are likely to bond with systems in someway when they respond safely and consistently, however as the simulacra research showed reactivity alone does not equal narrative agency.

The reason she fell short of narrative depth is that while she could sound engaged, she could not form evolving motivations, pursue long term goals, create internal conflict or growth and she could not build towards an ending. Players picked this up immediately not as a flaw but a limitation of the technology. Reactivity alone can not replace narrative arcs. CRPG companion stories are effective because they are written with structured growth, climaxes, and closure. An LLM cannot achieve this on its own because it does not truly remember, want or change. What they are good for is what players described Cassandra as a useful tactical partner, a support character, a reactive guide, and an immersive co-presence.

For this project to move forward the model would need a more in-depth memory consolidation system beyond simple RAG. Like the Stanford system, I would need to implement a memory stream loop that would generate logs about micro and macro events that would then generate goals and values. This new purposed system would have to be fully integrated with game systems beyond just the objective; Systems such as: Inventory, player stats, map layouts etc. essentially moving from LLM character to full agent.

This project demonstrates that LLM companions work as players are happy to engage and even if shallowly roleplay with them. But unfortunately, they do not yet achieve the same emotional depth or narrative relevance as pre authored RPG characters. The future lies not in replacing prewritten characters but instead a hybrid approach of prewritten macro arcs built up with AI driven micro arcs that are made with the use of an integrated memory stream and agent system.

Bibliography

Hafer, L. (2023) *Baldur's Gate 3 Review*. IGN. Available at: https://www.ign.com/articles/baldurs-gate-3-review (Accessed: 14 November 2023).

Brevig, M. (2023) *Baldur's Gate 3 players are spending an absurd amount of time in Act 1*. PC Gamer. Available at: https://www.pcgamer.com/baldurs-gate-3-players-are-spending-an-absurd-amount-of-time-in-act-1/ (Accessed: 14 November 2023).

West, J. (2023) *Baldur's Gate 3 dev says players with over 60 hours in the RPG "are just scraping average."* GamesRadar. Available at: https://www.gamesradar.com/games/baldurs-gate-3-dev-says-players-with-over-60-hours-in-the-rpg-are-just-scraping-average-so-its-time-to-get-those-rookie-numbers-up-but-with-nearly-2-000-hours-myself-im-not-sure-how-to-feel/ [Accessed: 21 November 2025].

Turkle, S. (2011) *Alone together: why we expect more from technology and less from each other*. New York: Basic Books. (Accessed: 20 November 2025).

Larian Studios (2023) *Baldur's Gate 3* [video game]. Ghent: Larian Studios. (Accessed: 20 November 2025).

Park, J.S., O'Brien, J.C., Cai, C.J., Morris, M.R., Liang, P., and Bernstein, M.S. (2023) 'Generative agents: interactive simulacra of human behaviour', *Proceedings of the 36th Annual ACM Symposium on User Interface Software and Technology (UIST '23)*. Available at: https://arxiv.org/abs/2304.03442 (Accessed: 20 November 2025)