The Turing *Quest*: Can Transformers Make Good NPCs?

Qi chen Gao
Brock University
1812 Sir Isaac Brock Way
St. Catharines, ON, Canada
qgao@brocku.ca

Ali Emami
Brock University
1812 Sir Isaac Brock Way
St. Catharines, ON, Canada
aemami@brocku.ca

Abstract

In this paper, we investigate the potential of using large pre-trained language models to generate non-playable character (NPC) scripts in video games. We introduce a novel pipeline that automatically constructs believable NPC scripts for various game genres and specifications using Transformer-based models. Moreover, we develop a self-diagnosis method, inspired by prior research, that is tailored to essential NPC characteristics such as coherence, believability, and variety in dialogue. To evaluate our approach, we propose a new benchmark, *The Turing Quest*, which demonstrates that our pipeline, when applied to GPT-3, generates NPC scripts across diverse game genres and contexts that can successfully deceive judges into believing they were written by humans. Our findings hold significant implications for the gaming industry and its global community, as the current reliance on manually-curated scripts is resource-intensive and can limit the immersiveness and enjoyment of players.

1 Introduction

Over the past decade, there has been a growing interest in applying deep learning models to Natural Language Generation (NLG) for open-domain dialogue systems and conversational agents. In parallel, the gaming industry has been striving to create more immersive experiences for players by enhancing their interactions with non-playable characters (NPCs). However, the potential of utilizing state-of-the-art deep learning models, such as Transformer-based models, to create NPC scripts remains largely unexplored.

Pre-trained Transformer-based language models (PLMs) like OpenAI's GPT-3 (Brown et al., 2020) and ChatGPT (Schulman et al., 2022) have demonstrated impressive conversational abilities (Milne-Ives et al., 2020). In certain contexts, the text generated by these models can be nearly indistinguishable from human-written text (M Alshater, 2022) without the aid of external tools or watermarks (Gambini et al., 2022). The use of these models in real-world applications has been expanding in areas such as customer service automation (Xu et al., 2017) (Zou et al., 2021), educational conversational agents (Molnár and Szüts, 2018), and mental health dialogue systems (Abd-Alrazaq et al., 2019).

Despite their growing prevalence, the effectiveness and generalization capabilities of PLMs in various contexts remain uncertain. One such uncharted domain is the creation of "non-playable characters" or NPCs in video games.

When comparing chatbots to NPCs, the latter can be considered as a narrative-driven variant of goal-oriented chatbots. However, NPCs and chatbots serve different purposes and operate in distinct environments. Generating NPC scripts presents unique challenges, as the dialogue must be consistent with the game's plot, genre, and the NPC's character to maintain player immersion and suspension of disbelief (Kerr and Szafron, 2009). According to Lee and Heeter (2015), NPC believability hinges on "*the size and nature of the cognitive gap between the [NPC that] players experience and the [NPC] they expect*". Players anticipate NPCs with individualized and possibly dynamic traits, which should be reflected in their dialogue. While incorporating personality into dialogue systems is well-studied (Qian et al., 2017) (Smestad and Volden, 2019) (de Haan et al., 2018), the challenge of generating goal-oriented, believable NPC scripts that align with a game's narrative and thematic elements, while preserving player immersion, remains substantial.

The ability to automatically generate contextually appropriate dialogue for a specified character could have an effect on the design paradigms of future video games. While manually scripted narratives and plot points will continue to hold their value, developers could augment player immersion

Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics - Student Research Workshop, pages 93–103
July 10-12, 2023 ©2023 Association for Computational Linguistics

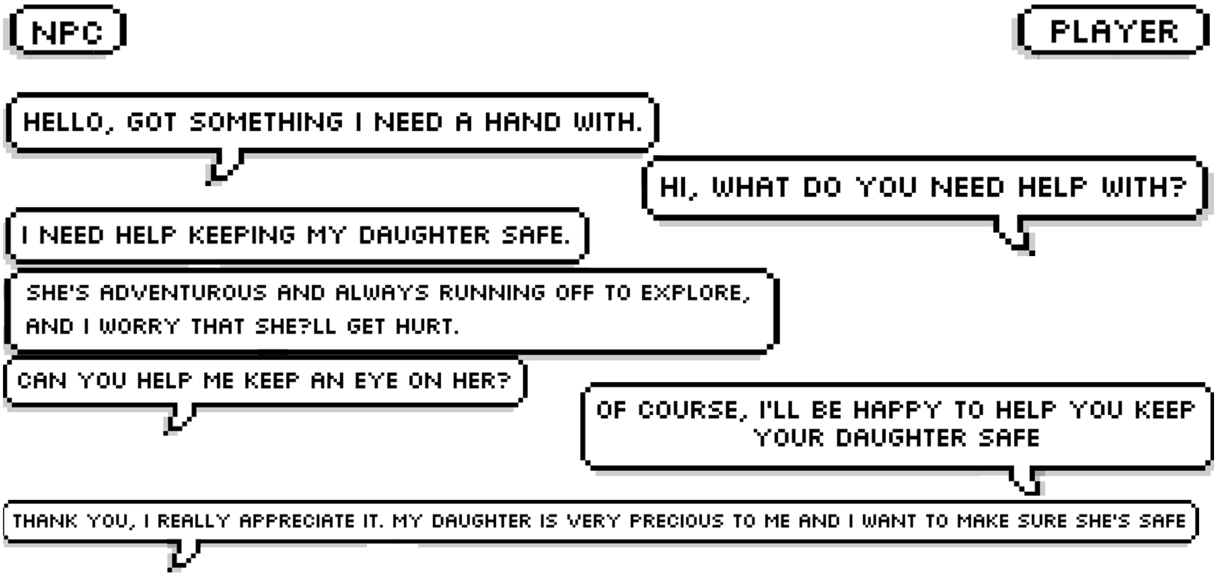

NPC
PLAYER
HELLO, GOT SOMETHING I NEED A HAND WITH.
HI, WHAT DO YOU NEED HELP WITH?
I NEED HELP KEEPING MY DAUGHTER SAFE.
SHE'S ADVENTUROUS AND ALWAYS RUNNING OFF TO EXPLORE, AND I WORRY THAT SHE?LL GET HURT.
CAN YOU HELP ME KEEP AN EYE ON HER?
OF COURSE, I'LL BE HAPPY TO HELP YOU KEEP YOUR DAUGHTER SAFE
THANK YOU, I REALLY APPRECIATE IT. MY DAUGHTER IS VERY PRECIOUS TO ME AND I WANT TO MAKE SURE SHE'S SAFE

Figure 1: A sample output of our NPC construction pipeline.

by allowing an array of NPCs to dynamically respond to a player's in-game progression.

Traditionally, game design involves scripted dialogues only for NPCs that contribute directly to a quest or story line, thereby limiting the extent of player interaction. It is not often possible for a player to initiate a conversation with a companion about an ongoing quest or solicit their views, creating an impression that, from an NPC's perspective, the player's existence is confined to the quests they undertake.

Simply implementing an interactive companion system necessitates writing dialogues for every quest for all possible companions—a labor-intensive task. Expanding this system to encompass a majority of a game's NPCs would further compound these challenges, increasing the amount of labour to an unreasonable degree. The vast amount of dialogue required for each narrative stage would significantly exceed typical time and resource constraints of most developers. Despite the potential enrichment of the player experience, the practicality of creating such an immersive, dialogue-rich environment using solely human-authored dialogue in game development remains questionable.

In this study, we investigate the application of Transformer-based models like GPT-3 to the task of creating NPCs and generating believable scripts. To this end, we develop an NPC construction pipeline capable of generating dialogue based on the NPC's attributes alone. Our pipeline comprises three key modules: a) a *Feature Characterization Schema* that classifies NPCs based on personality traits and world descriptions, b) an *Automatic Prompt Creation* process that employs the schema to generate tailored prompts for conditioning language models, and c) a *Dialogue Generation* phase that uses the customized prompts to generate scripts with Transformer-based PLMs. Figure 1 provides an example of dialogue generated through this pipeline. We also devise and automate an evaluation metric for NPC dialogue quality, drawing inspiration from related literature (Brown et al., 2020). Lastly, we propose the Turing Quest: a test using human judges to assess the believability and quality of generated NPC scripts.

2 Related Work

In recent years, there has been a growing interest in dialogue systems and conversational agents. However, the exploration of dialogue generation for NPCs in video games, despite their similarities to chatbots, remains limited. Although most video games in the past decade include NPC dialogue, research on automating its creation using Artificial Intelligence (AI) is still in its infancy.

NPC Dialogue generation. In the early 2000s, efforts in NLP to create better NPC dialogue relied on hand-crafted algorithms and manually authored grammars (Schlünder and Klabunde, 2013) (Ryan et al., 2016). Schlünder and Klabunde (2013) succeeded in generating greetings that players perceived as more polite and appropriate than in-game greetings. However, their rule-based method relied on labor-intensive, discrete human-defined steps that were difficult to scale into full branching conversations. With recent advancements in goal-oriented chatbots utilizing machine learning techniques such as reinforcement learning (Liu et al., 2020) and dialogue generation through deep reinforcement learning (Li et al., 2016) (Li, 2020), automating NPC dialogue generation becomes increasingly feasible.

The introduction of AI into games has led to the application of various AI techniques and algorithms to enhance gameplay experiences through improved bots (Nareyek, 2004) and adaptive experiences (Raifer et al., 2022). There has been significant research into using machine learning to create bots that provide challenging and entertaining opponents for players (Håkansson and Fröberg, 2021). However, this trend of applying machine learning to different game design tasks does not extend to dialogue generation for NPCs.

Although pre-trained language models such as GPT-3 continue to expand their applicability, generalization remains an unsolved problem. While PLMs like GPT-3 have shown natural language generation capabilities (Topal et al., 2021), research into NLG with Transformer-based models trained on NPC dialogue has revealed that the generated dialogue "compared rather poorly to human-written [dialogue]" in terms of purpose and coherence (Kalbiyev, 2022). Nevertheless, generalization difficulty for LMs is not unique to NPC dialogue (Ye et al., 2021). We hypothesize that NPC dialogue is not merely another generalization problem but a distinct task. This hypothesis is supported by the inadequacy of chatbot evaluation metrics (Peras, 2018) when applied to NPC dialogue.

NPC Dialogue Metrics. Metrics proposed for chatbots do not directly translate to suitable metrics for NPC dialogue. While chatbot success is often determined by how "human" they sound and their ability to maintain a conversation with a human (Turing, 1950), NPC dialogue is always directed and goal-oriented. Generating dialogue for NPCs presents unique challenges compared to text generation in fictional settings. The generated dialogue must be consistent with the game world and the NPC's specific traits and personality, and it should ensure coherence and contextual relevance in relation to the player's input. No test equivalent to the Turing test or its alternatives, such as the Winograd schema (WSC) (Winograd, 1972; Levesque et al., 2011) exists specifically for NPC dialogue. To our knowledge, there is no standard metric to evaluate the quality of generated NPC dialogue. One suggested metric for NPC dialogue is "coherence, relevance, human-likeness, and fittingness" (Kalbiyev, 2022). While coherence, relevance, and human-likeness can be applied to chatbots, fittingness—defined by Kalbiyev (2022) as how well the response fits the game world—is unique to NPCs.

3 NPC Construction Pipeline

The objective of the NPC construction pipeline is to automatically generate coherent, contextually appropriate, and engaging utterances for an NPC, given the dialogue history between the NPC and a player, as well as the contextual information about the NPC and the game. The pipeline consists of three modules, which serve to a) characterize the NPC according to a generalized representation schema that captures crucial information about the NPC's role, personality, and game context, b) generate short prompts based on the characterization, providing contextually relevant pretexts for the language model (LM), and c) generate utterances based on these prompts using an LM optimized for NPC dialogue generation.

3.1 Module 1: Feature Characterization Schema

The first module in the pipeline involves developing a schema that characterizes a given NPC according to a number of game- and NPC-relevant features. Identifying the most concise set of features needed to define any NPC is a challenging task, as NPCs not only exhibit vastly different personalities but can also serve different purposes for the player and the game world. For example, in the action role-playing game, "The Elder Scrolls V: Skyrim" (Bethesda Game Studios, 2011), the NPC *Balgruuf the Greater* is a Jarl, i.e., a king or ruler who assigns quests to the player to maintain peace. In contrast, a character like *KL-E-0* from "Fallout 4" (Bethesda Game Studios, 2015), a robot arms dealer in a post-nuclear apocalyptic world, has little concern for peace. Based on (Warpefelt, 2016), NPCs should possess both a ludic function and a narrative framing for their actions to be coherent and believable. That is, an NPC should fulfill a gameplay or mechanical purpose—i.e., a ludic function—while advancing the narrative through their actions.

To develop a characterization of NPCs that captures their differences across various games and genres, we should consider several important features, such as their relationship and role with respect to the player (e.g., buying and selling, providing quests, etc.) and their individual personality and values. Taking into account narrative purpose, ludic purposes, and the personality and characteristic differences of NPCs, we propose five game-specific features to characterize and distinguish NPCs:

	Narrative	Ludic function
World Desc.	✓	
NPC Role		✓
NPC Personality	✓	
Game State	✓	✓
NPC Objective	✓	✓

Table 1: The features and their purpose(s).

Each of these five features either fulfills a ludic function or contributes to the game's narrative, and in some cases, a feature serves both purposes. This schema enables us to classify NPCs based on their in-game mechanics (Hunicke et al., 2004) while also capturing their role in the game's story. By incorporating these features into the NPC construction pipeline, we can create NPCs that not only adhere to the context and constraints of the game world but also exhibit distinct and engaging personalities, which can significantly enhance players' immersion and overall gaming experience.

World Description. A world description provides a summary of the story thus far, including information about the game world and its unique characteristics. Without this information, actions, thoughts, and utterances may be incoherent or unfitting, as they lack awareness of the setting and genre. This may result in dialogue or actions that conflict with the player's expectations. For instance, if Balgruuf from the previous example, originating from a fantasy adventure game, were placed in a sci-fi horror set in space, his actions, appearance, and dialogue would clash with the rest of the game. NPCs become "essentially incomprehensible if they are not framed according to the narrative" (Warpefelt, 2016). Ignoring information related to the setting, genre, and themes present in the NPC's world may affect the believability and fittingness of the NPC. More importantly, the narrative dissonance generated could shatter the *willful suspension of disbelief*—coined by Samuel Taylor Coleridge (1971)—and break the player's immersion in the game's world and story.

Role. Each unique NPC is created to fulfill a purpose. Continuing from the previous example, Balgruuf primarily functions as a *quest-giver*—facilitating the player's progression through the main quest line and occasionally offering side quests to enrich the narrative experience. Omitting his role would fail to represent a critical function of his character. Defining the role of an NPC, whether as a vendor, quest giver, or storyteller, etc., is thus crucial. We selected these roles based on the typology of NPCs and the NPC model proposed in (Warpefelt, 2016). We adapted the types of NPCs from (Warpefelt, 2016) and simplified the set of NPC types to those that would feasibly have a conversation with the player while also merging entries that were similar in their roles. This resulted in eight types of NPCs, six neutral or friendly roles, and two non-friendly roles, as shown below, in Table 2.

Metatype	Role
Functional	Vendor Service Provider Questgiver
Providers	Story teller
Friendly	Ally Companion
Adversaries	Enemy Villain

Table 2: Adapted NPC types.

The role an NPC occupies influences their expected dialogue. Although these roles are not mutually exclusive within a single NPC (e.g., some NPCs can be vendors at times while providing a quest at another time), at any given point during a dialogue with a player, the NPC occupies only one of these roles.

Personality. To describe any given NPC, it is necessary to elaborate on their personality and unique characteristics that distinguish them from other characters. These characteristics include physical attributes and appearances, psychological and personality traits such as the strength of the *OCEAN* personality traits proposed in (Digman, 1990), likes and dislikes, etc. This feature focuses on the details of the NPC's character, such as their occupation, beliefs, and other related details. NPCs are characters at their core, making it essential to incorporate these details into their depiction.

Game State. This describes the progression of the game and changes to the NPC's location. The NPC's dialogue may change based on the objectives completed by the player and the current state of the in-game world. The addition of this feature allows us to focus on the NPC during any single time frame during the course of the game. This enables better classification of dynamic NPCs that change over the course of the game and react to the player's actions. This feature also allows specifying details such as the current location of the NPCs and the scope of information the NPC possesses. Game state serves both a narrative and ludic purpose; for example, a shopkeeper may offer more goods depending on the player's actions, and the NPC's location also aids in framing their actions and dialogue, as a vendor may only offer certain goods in specific towns.

Objective. The NPC Objective is the purpose of the NPC apart from the player. According to Dennett Daniel (1981), *personhood* consists of six different themes: Rationality, Intentionality, Stance, Reciprocity, Communication, and Consciousness. Providing an NPC with a *role* satisfies intentionality, as each action should be motivated by what the NPC was designed to achieve. However, giving them goals and aspirations allows the NPC to have a *stance* and perhaps even *consciousness* (Kalbiyev, 2022). If a blacksmith's objective is to raise enough money for their family, they should act and speak accordingly. Their actions and dialogue should not solely reflect their personality but also their objective. This feature allows the schema to capture complex and dynamic NPCs with intricate values and goals not fully represented by their *role* or *personality*. The addition of this feature enables the NPC to have a greater purpose than merely serving as an outlet for exposition or facilitating a game function.

With these features, we propose that each unique NPC can be encapsulated and represented wholly, as shown in figure 2. Each one of these features is independent of one another, allowing for modularity when designing NPCs. However, clashing combinations may still exist regardless of the modular nature of this schema.

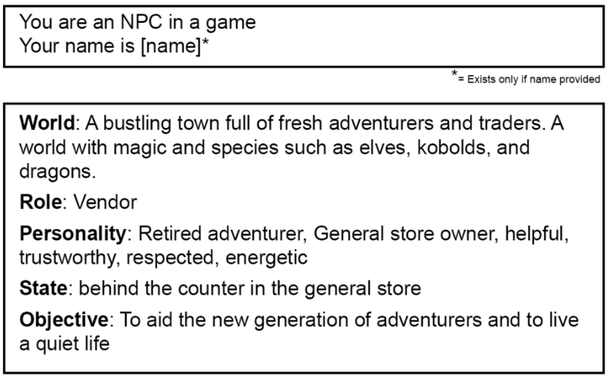
World	A fantasy world of Dragons and magic; Skyrim
Role	Questgiver
Personality	Nord, Jarl of Whiterun, Loyal, Noble, Blonde, reasonable
State	Sitting on throne in dragonsreach. Contemplating the war and recent reports of dragons
Goal	The safety and prosperity of the people of whiterun and a solution to the looming dragon threat.

Figure 2: Completed features for "Balgruuf the Greater".

3.2 Module 2: Prompt Creation

Prompt creation was designed with the feature representation schema in mind. Providing the LM with sufficient information about an NPC is crucial to ensure that the generated dialogue remains consistent with the character's identity. These requirements are akin to the challenges faced by the feature representation schema. Consequently, the prompt creation module integrates the various features present in the schema and uses them as a prompt. The first line of each prompt begins with the sentence "You are an NPC in a game", followed by optional details such as a name, some details about the world that the NPC inhabits, the role of the NPC, basic personal characteristics, their current state (e.g., sitting outside thinking about their daughter), and finally their goal(s). Most of these categories are optional, except for the NPC type (i.e., their *role*), which must always be present. By incorporating these features, the prompt creation module empowers users to guide the LM in generating diverse NPCs with individualized personalities, allowing for greater customization without the need for prior fine-tuning or training.

NPC Header. Utilizing this prompt creation method, we created the NPC header, a representative example is depicted in figure 3. This header plays a pivotal role in dialogue generation by providing essential information about the character. For our needs, we also created a player header using the same information used in the NPC header, guiding the LM to mimic a player's behavior and facilitate automated dialogue generation. The generated player dialogue is less creative and more prone to repetition compared to human-written dia-



You are an NPC in a game
Your name is [name]*

*= Exists only if name provided

World: A bustling town full of fresh adventurers and traders. A world with magic and species such as elves, kobolds, and dragons.

Role: Vendor

Personality: Retired adventurer, General store owner, helpful, trustworthy, respected, energetic

State: behind the counter in the general store

Objective: To aid the new generation of adventurers and to live a quiet life

Figure 3: Example of an NPC header.

logue. This issue is beyond the scope of this paper, as our focus lies on NPC dialogue generation.

3.3 Module 3: Dialogue Generation

Dialogue generation was executed automatically and iteratively. The prompt was structured as a combination of the header and the current dialogue history. The header section is continually swapped depending on which agent's dialogue—NPC or player—is currently being generated. By placing the header at the top of the prompt and swapping it for the active agent, PLMs can generate dialogue that is coherent with the current speaker and their traits.

First Sentences. In early development-stage results, GPT-3 demonstrated difficulty in generating effective first sentences. Combined with the inherent challenge of generating human-like responses, this led to a significant drop in the overall quality of dialogue—often resulting in both NPC and player generating blank lines or constantly repeating the same responses. A workaround was developed by employing a small set of hand-written first sentences based on the genre and NPC type. This workaround allowed the conversation to avoid immediate repetition while minimizing interference with dialogue generation.

Repetition. In our preliminary testing, we found that PLMs struggle to avoid repetition when the player dialogue is similar to a past query or sentence. This often caused the NPC's response to be similar or even identical to its previous response. To circumvent this issue, we implemented a dynamic frequency penalty. The dynamic frequency penalty incrementally increases when the NPC or player generates a response that already exists in the conversation. After detecting a repetition and incrementing the frequency penalty, the LM attempts to regenerate with the same prompt, excluding the repeated sentence. This process occurs up to three times or until a new sentence is generated before resetting the frequency penalty to the original value before any increments. This technique significantly reduced overall repetitions and drastically decreased the occurrence of loops appearing early in the conversation.

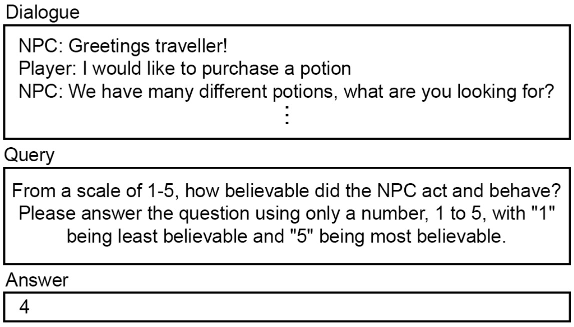
4 Evaluation

To assess the performance of the NPC construction pipeline and the resulting generated dialogue, we designed a comprehensive evaluation metric that examines dialogue quality based on coherency, believability, degree of repetition, alignment of the NPC's dialogue with their role, and fittingness of the NPC's dialogue within their world. These categories draw from and adapt Kalbiyev (2022)'s metric for evaluating video game dialogue. Each metric is assigned a score between one and five, with the sum of these scores indicating the overall quality of the dialogue.

Self-diagnosis harnesses the capacity of Transformer-based language models to detect patterns within text and their few-shot learning performance to enable rapid, automated evaluation of dialogue without prior fine-tuning. We conducted a human evaluation of 66 different NPC scripts to assess the accuracy and reliability of our self-diagnosis approach. After each conversation was evaluated and scored, we found a correlation between parameters and their average score. By including our full NPC header, we were able to generate dialogue of higher quality. We then conducted a single-blind test where human judges were asked to determine whether an NPC script was generated by AI or written manually by a human.

4.1 Self-Diagnosis

We investigated the ability of pretrained language models, such as GPT-3, to understand, evaluate, and diagnose dialogue when given a specific nontrivial query (e.g., "whether an NPC behaved coherently"). Schick et al. (2021) demonstrate that PLMs can identify socially undesirable attributes in text, such as racism and violence. We propose that this self-diagnosis capability is not only applicable to socially undesirable attributes but also enables PLMs to self-diagnose a broader and more general set of attributes, themes, and behaviors without further fine-tuning. For simple questions, such as if a genre was clearly distinguishable in text, PLMs perform accurately in a zero-shot environment without examples and further guidance. This behavior is supported by Sanh et al. (2022). However, this performance does not hold when dealing with more complicated and potentially subjective questions.



Dialogue

NPC: Greetings traveller!
Player: I would like to purchase a potion
NPC: We have many different potions, what are you looking for?
⋮

Query

From a scale of 1-5, how believable did the NPC act and behave? Please answer the question using only a number, 1 to 5, with "1" being least believable and "5" being most believable.

Answer

4

Figure 4: Prompt structure of self-diagnosis.

Our self-diagnosis approach consists of providing examples of different scoring dialogue for each metric that needed further clarification. By scoring dialogue", we mean, for example, giving the LM a prompt like "What a perfect score looks like" or "What a 3 should look like". In preliminary tests, we found that simply inputting a script and posing a question led to relatively reliable results; however, the output occasionally did not align with human responses or logic. By formulating the question more precisely and asking for a numeric response rather than a free-form sentence response, we were able to obtain a numeric answer more accurately. To account for potential variability in the responses, we set the temperature to 0 for each test, yielding a deterministic model devoid of stochastic behavior. We leveraged the PLM's few-shot learning abilities by adding three examples of different scoring sample dialogue before the prompt. This approach aligns scores obtained through self-diagnosis more closely with human scores on queries that a PLM would otherwise have difficulties with.

4.2 The Turing Quest

To evaluate the performance of our NPC Construction pipeline and the degree to which the resulting generated dialogue appears human-written, we propose a test tailored to NPC dialogue—the *Turing Quest*. Inspired by the Turing test (Turing, 1950), the goal of this test is to determine whether a generated NPC script can be distinguished from human-written dialogue by human judges. A script passes the Turing Quest if the judge deems it human-written, and fails if perceived as AI-generated. Conducting this test on multiple NPC script samples helps assess the proficiency of state-of-the-art PLMs in generating convincing NPC dialogue.

The Turing Quest is a self-administered questionnaire. For each script, it asks the judge to determine if the NPC's dialogue is written by a human or an AI. Since the scope of this test is to determine the believability of an NPC's dialogue, the player's dialogue can be manually written by a human.

For our test, six NPC scripts were evaluated by 12 individual judges. Four of the six scripts were generated by GPT-3, one was manually written, and the final script was sampled from the game *Skyrim*. Our test group comprised twelve people familiar with video games and NPCs. From the responses of our judges, we determined the average passing rate was 64.58% for all AI-generated scripts. The best performing generated script had a pass rate of 75%. Interestingly, 75% of judges believed that the dialogue sampled from Skyrim was AI-generated and 50% thought the same for the manually written script. This could highlight the expectations of players regarding the current state and abilities of LMs and conversational agents. These findings provide strong empirical evidence that our pipeline, when applied to PLMs, is capable of producing NPC scripts that resemble and perhaps even surpass human-written NPC dialogue.

5 Experiments and Results

5.1 Parameter Search and Model Selection

We conducted a comprehensive random grid parameter search to identify the optimal model and parameters for generating high-quality NPC dialogue. Three key parameters influenced the quality and score of the generated dialogue: the language model, temperature setting, and the integration of our NPC construction pipeline prompt.

Utilizing different versions of GPT-3 (OpenAI's text-davinci-002, text-curie-001, and text-babbage-001 models) and a range of temperatures (0 to 1, incremented by 0.1), we compared the quality of dialogue generated with our full prompt and a minimal version without the world description, NPC Personality, game state, and NPC objective sections. We repeated the experiment with another NPC role to ensure generalizability[1].

[1] The code to reproduce all of our experimental results are available at https://github.com/FieryAced/-NPC-Dialogue-Generation.

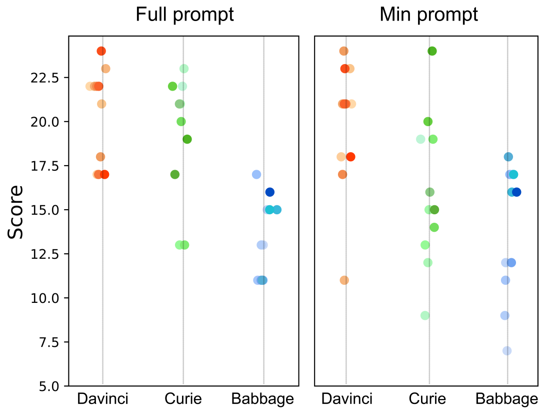


Figure 5: Evaluation Scores of varying models and temperatures.

Our analysis revealed a significant decline in quality from the text-davinci-002 to text-curie-001 models, and an even more pronounced decrease between text-curie-001 and text-babbage-001. This is consistent with recent research which has shown that larger and more complex models, such as GPT-3's text-davinci-002 model, have the ability to learn and generalize more complex patterns from larger and more diverse datasets, resulting in better performance across a wide range of natural language processing tasks (Brown et al., 2020).

Furthermore, the recently proposed InstructGPT framework by Ouyang et al. (2022) allows for targeted fine-tuning of pre-trained language models to better suit the task at hand. This approach involves providing additional instructions during fine-tuning, such as providing task-specific prompts or data augmentation techniques, which results in improved performance for downstream tasks. With the success of InstructGPT, it is becoming increasingly clear that language models can be further optimized for specific use-cases by adjusting their architecture or fine-tuning process. Thus, it is reasonable to assume that newer and more advanced models, such as text-davinci-003, should generally perform better than their predecessors. Finally, our analysis shows that full-prompt models outperformed minimal prompt ones, with an average 4.06 point higher score, demonstrating the effectiveness of our prompting method.

A Pearson correlation test (excluding the atypical data point with a temperature of 0) showed a positive correlation between temperature and score, $r(8) = .7055, p = .022646$. Higher temperature values yielded better results, with the highest average scores at temperatures of 0.9 and 0.8.

Based on these findings, we recommend using advanced Transformer-based LMs like OpenAI's GPT-3 "text-davinci-002" at a temperature around 0.9, along with our NPC construction pipeline, for optimal NPC script generation.

5.2 Results

Self-Diagnosis: To assess the reliability of the self-diagnosis module, we manually evaluated 66 NPC scripts using the same metrics applied in self-diagnosis. A Pearson correlation test showed a strong positive correlation between self-diagnosed and human-evaluated scores, $r(64) = .8092, p < .00001$. This demonstrates the module's consistency and correlation with human evaluation scores.

Turing Quest Results: Our NPC construction pipeline, when using the recommended parameters, generates dialogue that not only passes as human-written but also scores highly on the evaluation metric. On average, our generated dialogue was thought to be hand-written 64.58% of the time with the best performing script passing as human written 75% of the time. The generated NPC scripts exhibit goal-oriented behavior and adherence to the in-game world and genre, maintaining player immersion. The Turing Quest results further confirm the high quality of the generated dialogue.

6 Conclusion

We developed a novel pipeline capable of automatically generating NPC scripts comparable or of superior quality to human-written NPC dialogue using Transformer-based PLMs. We then created a self-diagnosis module which provides a method to evaluate and compare the quality of NPC dialogue quantitatively. Finally, our proposal of the Turing Quest allows us to determine the capabilities of a language model when applied to the task of NPC dialogue generation and whether a script passes as human-written. While the NPC construction pipeline allows for modularity even in between responses, that aspect was not explored in depth in this paper. We will explore dialogue generation for dynamic NPCs with evolving roles or attributes in future research.

Limitations

The dialogue generated for the player exhibits a higher degree of repetition and has a tendency towards looping. This limitation exists as we did not focus on generating player dialogue as that is a different problem of its own. To account for this limitation, both the self-diagnosis and the Turing Quest only evaluate the NPC's dialogue.

Currently, the maximum context window for the dialogue history portion is limited by the max tokens of a given model minus the tokens required for the NPC header. Despite being a rare occurrence, it is possible that the dialogue history becomes so long that the model may not be able to generate any responses as there is no more remaining space. We did not experience this problem; however, a workaround would be to discard the oldest dialogue history entry as needed. This approach however may cause the NPC to lose out on information that it would otherwise be able to leverage in dialogue.

Ethics Statement

The presence of bias within NPC models/systems poses a significant risk particularly as the demographic of young individuals, still in the age of development, who enjoy playing video games continues to expand. In 2006, 92% of children in the ages of 2-17 had played video games (Doğan, 2006). 97% of players under the age of of 18 play more that an hour of games daily (Granic et al., 2014). According to recent statistics, the global demographic of active video game players is projected to increase over 5% year-over-year (Doğan, 2006), reaching over 3 billion active players worldwide in 2023[2]. This means, in the future, video games will reach more young children and adolescents. If the presence of bias is not addressed, it could subconsciously normalize problematic behaviours seen in games in children as humans are a product of both nature and nurture (Plomin and Asbury, 2005). This in turn may lead to more biases being overlooked or ignored by the next generation of researchers, creating a vicious cycle.

Acknowledgements

This work was supported by the Natural Sciences and Engineering Research Council of Canada and by the New Frontiers in Research Fund.

[2] https://www.statista.com/statistics/748044/number-video-gamers-world

References

Alaa A Abd-Alrazaq, Mohannad Alajlani, Ali Abdallah Alalwan, Bridgette M Bewick, Peter Gardner, and Mowafa Househ. 2019. An overview of the features of chatbots in mental health: A scoping review. *International Journal of Medical Informatics*, 132:103978.

Bethesda Game Studios. 2011. The elder scrolls v: Skyrim.

Bethesda Game Studios. 2015. Fallout 4.

Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel Ziegler, Jeffrey Wu, Clemens Winter, Chris Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. 2020. Language models are few-shot learners. In *Advances in Neural Information Processing Systems*, volume 33, pages 1877–1901. Curran Associates, Inc.

Hayco de Haan, Joop Snijder, Christof van Nimwegen, and Robbert Jan Beun. 2018. Chatbot personality and customer satisfaction. *Info Support Research*.

C Dennett Daniel. 1981. Conditions of personhood. *The Identities of Persons*, 175.

John M Digman. 1990. Personality structure: Emergence of the five-factor model. *Annual review of psychology*, 41(1):417–440.

Filiz Öztütüncü Doğan. 2006. Video games and children: violence in video games. In *New/Yeni Symposium Journal*, volume 44, pages 161–164.

Margherita Gambini, Tiziano Fagni, Fabrizio Falchi, and Maurizio Tesconi. 2022. On pushing deepfake tweet detection capabilities to the limits. In *14th ACM Web Science Conference 2022*, WebSci '22, page 154–163, New York, NY, USA. Association for Computing Machinery.

Isabela Granic, Adam Lobel, and Rutger CME Engels. 2014. The benefits of playing video games. *American psychologist*, 69(1):66.

Carl Håkansson and Johan Fröberg. 2021. Application of machine learning to construct advanced npc behaviors in unity 3d.

Robin Hunicke, Marc Leblanc, and Robert Zubek. 2004. Mda: A formal approach to game design and game research. *AAAI Workshop - Technical Report*, 1.

A Kalbiyev. 2022. Affective dialogue generation for video games. Master's thesis, University of Twente.

Christopher Kerr and Duane Szafron. 2009. Supporting dialogue generation for story-based games. In *Proceedings of the AAAI Conference on Artificial Intelligence and Interactive Digital Entertainment*, volume 5, pages 154–160.

Michael Sangyeob Lee and Carrie Heeter. 2015. Cognitive intervention and reconciliation: Npc believability in single-player rpgs. *International Journal of Role-Playing*, 5:47–65.

Hector Levesque, Ernest Davis, and Leora Morgenstern. 2011. The winograd schema challenge. In *AAAI Spring Symposium: Logical Formalizations of Commonsense Reasoning*.

Jiwei Li, Will Monroe, Alan Ritter, Michel Galley, Jianfeng Gao, and Dan Jurafsky. 2016. Deep reinforcement learning for dialogue generation. *arXiv preprint arXiv:1606.01541*.

Piji Li. 2020. An empirical investigation of pre-trained transformer language models for open-domain dialogue generation. *CoRR*, abs/2003.04195.

Jianfeng Liu, Feiyang Pan, and Ling Luo. 2020. Gochat: Goal-oriented chatbots with hierarchical reinforcement learning. In *Proceedings of the 43rd International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '20, page 1793–1796, New York, NY, USA. Association for Computing Machinery.

Muneer M Alshater. 2022. Exploring the role of artificial intelligence in enhancing academic performance: A case study of chatgpt. *Available at SSRN*.

Madison Milne-Ives, Caroline de Cock, Ernest Lim, Melissa Harper Shehadeh, Nick de Pennington, Guy Mole, Eduardo Normando, and Edward Meinert. 2020. The effectiveness of artificial intelligence conversational agents in health care: Systematic review. *J Med Internet Res*, 22(10):e20346.

György Molnár and Zoltán Szüts. 2018. The role of chatbots in formal education. In *2018 IEEE 16th International Symposium on Intelligent Systems and Informatics (SISY)*, pages 000197–000202. IEEE.

Alexander Nareyek. 2004. Ai in computer games: Smarter games are making for a better user experience. what does the future hold? *Queue*, 1(10):58–65.

Long Ouyang, Jeffrey Wu, Xu Jiang, Diogo Almeida, Carroll Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, et al. 2022. Training language models to follow instructions with human feedback. *Advances in Neural Information Processing Systems*, 35:27730–27744.

Dijana Peras. 2018. Chatbot evaluation metrics. *Economic and Social Development: Book of Proceedings*, pages 89–97.

Robert Plomin and Kathryn Asbury. 2005. Nature and nurture: Genetic and environmental influences on behavior. *The Annals of the American Academy of Political and Social Science*, 600(1):86–98.

Qiao Qian, Minlie Huang, Haizhou Zhao, Jingfang Xu, and Xiaoyan Zhu. 2017. Assigning personality/identity to a chatting machine for coherent conversation generation. *CoRR*, abs/1706.02861.

Maya Raifer, Guy Rotman, Reut Apel, Moshe Tennenholtz, and Roi Reichart. 2022. Designing an automatic agent for repeated language–based persuasion games. *Transactions of the Association for Computational Linguistics*, 10:307–324.

James Ryan, Michael Mateas, and Noah Wardrip-Fruin. 2016. Characters who speak their minds: Dialogue generation in talk of the town. In *Twelfth Artificial Intelligence and Interactive Digital Entertainment Conference*.

Samuel Taylor Coleridge. 1971. *Biographia Literaria, 1817*.

Victor Sanh, Albert Webson, Colin Raffel, Stephen H. Bach, Lintang Sutawika, Zaid Alyafeai, Antoine Chaffin, Arnaud Stiegler, Teven Le Scao, Arun Raja, Manan Dey, M Saiful Bari, Canwen Xu, Urmish Thakker, Shanya Sharma Sharma, Eliza Szczechla, Taewoon Kim, Gunjan Chhablani, Nihal Nayak, Debajyoti Datta, Jonathan Chang, Mike Tian-Jian Jiang, Han Wang, Matteo Manica, Sheng Shen, Zheng Xin Yong, Harshit Pandey, Rachel Bawden, Thomas Wang, Trishala Neeraj, Jos Rozen, Abheesht Sharma, Andrea Santilli, Thibault Fevry, Jason Alan Fries, Ryan Teehan, Tali Bers, Stella Biderman, Leo Gao, Thomas Wolf, and Alexander M. Rush. 2022. Multitask prompted training enables zero-shot task generalization.

Timo Schick, Sahana Udupa, and Hinrich Schütze. 2021. Self-Diagnosis and Self-Debiasing: A Proposal for Reducing Corpus-Based Bias in NLP. *Transactions of the Association for Computational Linguistics*, 9:1408–1424.

Björn Schlünder and Ralf Klabunde. 2013. Greetings generation in video role playing games. In *Proceedings of the 14th European Workshop on Natural Language Generation*, pages 167–171, Sofia, Bulgaria. Association for Computational Linguistics.

J Schulman, B Zoph, C Kim, J Hilton, J Menick, J Weng, JFC Uribe, L Fedus, L Metz, M Pokorny, et al. 2022. Chatgpt: Optimizing language models for dialogue.

Tuva Lunde Smestad and Frode Volden. 2019. Chatbot personalities matters. In *International conference on internet science*, pages 170–181. Springer.

M. Onat Topal, Anil Bas, and Imke van Heerden. 2021. Exploring transformers in natural language generation: Gpt, bert, and xlnet. *CoRR*, abs/2102.08036.

A. M. Turing. 1950. I.—COMPUTING MACHINERY AND INTELLIGENCE. *Mind*, LIX(236):433–460.

Henrik Warpefelt. 2016. *The Non-Player Character: Exploring the believability of NPC presentation and behavior*. Ph.D. thesis, Stockholm University.

Terry Winograd. 1972. Understanding natural language. *Cognitive Psychology*, 3(1):1–191.

Anbang Xu, Zhe Liu, Yufan Guo, Vibha Sinha, and Rama Akkiraju. 2017. A new chatbot for customer service on social media. In *Proceedings of the 2017 CHI conference on human factors in computing systems*, pages 3506–3510.

Qinyuan Ye, Bill Yuchen Lin, and Xiang Ren. 2021. Crossfit: A few-shot learning challenge for cross-task generalization in nlp.

Yicheng Zou, Lujun Zhao, Yangyang Kang, Jun Lin, Minlong Peng, Zhuoren Jiang, Changlong Sun, Qi Zhang, Xuanjing Huang, and Xiaozhong Liu. 2021. Topic-oriented spoken dialogue summarization for customer service with saliency-aware topic modeling. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 35, pages 14665–14673.