The Turing Quest: Can Transformers Make Good NPCs?

Qi Chen Gao, Ali Emami

Association for Computational Linguistics

Aim

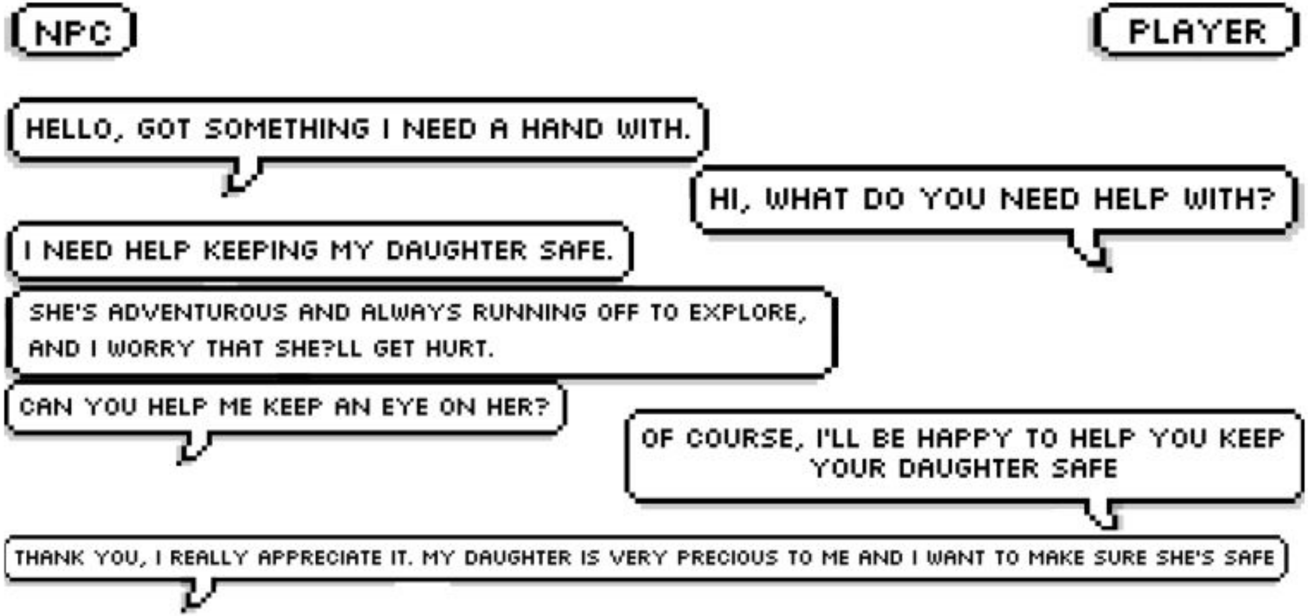
The potential of utilizing state-of-the-art learning models to create NPC scripts remains largely unexplored.

The extent of player interaction with NPCs is limited.

Implementing an interactive companion system requires enormous labour.

Aim

The application of Transformer-based models like GPT-3 to the task of creating NPCs and generating believable scripts.



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.

Background and Related Works

NPC Dialogue generation

In the early 2000s, creating better NPC dialogue depended on hand-crafted algorithms and manually authored grammars.
=> This was difficult to scale into full branching conversations.

Automating NPC dialogue generation becomes increasingly feasible.

Applying machine learning to game design tasks does not extend to NPC dialogue generation.

Background and Related Works

NPC Dialogue Metrics

No test equivalent to the Turing test or its alternatives, such as Winograd schema exists specifically for NPC dialogue.

The metric for NPC dialogue is “coherence, relevance, human-likeness, and fittingness”.

“fittingness” is defined by Kalbiyev (2022) as how well the response fits the game world.

NPC Construction Pipeline

a) Feature Characterization Schema

b) Prompt Creation

c) Dialogue Generation

NPC Construction Pipeline

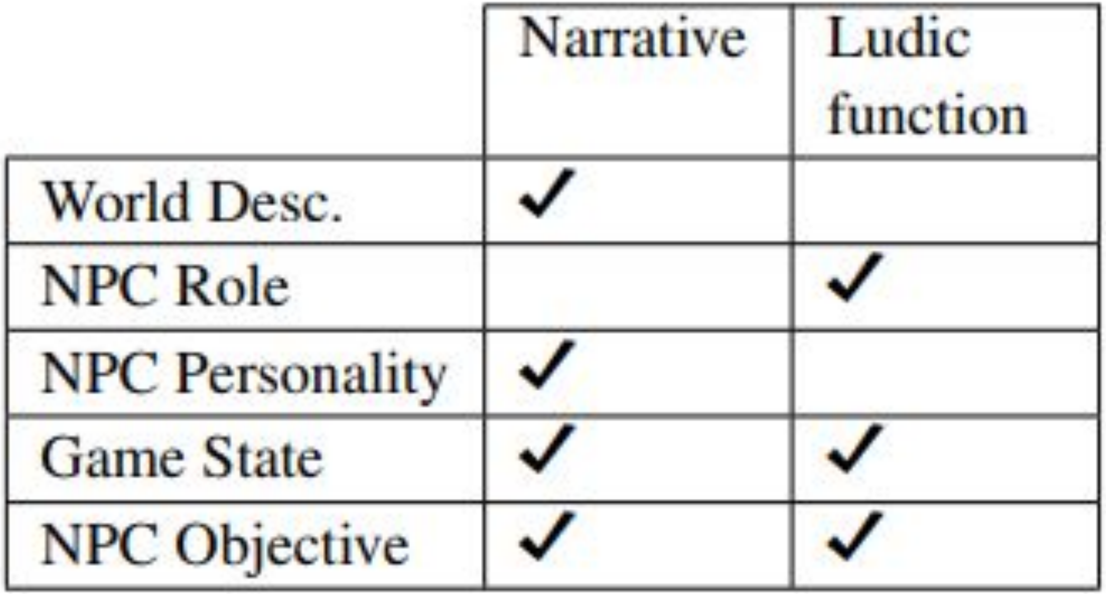
Feature Characterization Schema

This module involves developing a schema.

NPCs do not only show different personalities but can also serve different purposes for the player and the game world.

NPC Construction Pipeline

Feature Characterization Schema



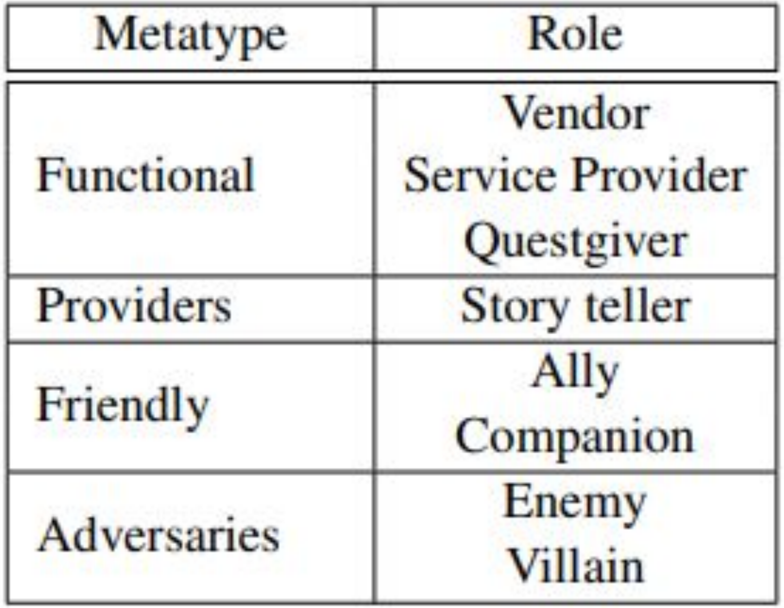
	Narrative	Ludic function
World Desc.	✓	
NPC Role		✓
NPC Personality	✓	
Game State	✓	✓
NPC Objective	✓	✓

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

NPC Construction Pipeline

Feature Characterization Schema

These roles are based on the typology of NPCs and the NPC model proposed in (Henrik Warpefelt. 2016. *The Non-Player Character: Exploring the believability of NPC presentation and behavior.* Ph.D. thesis, Stockholm University).

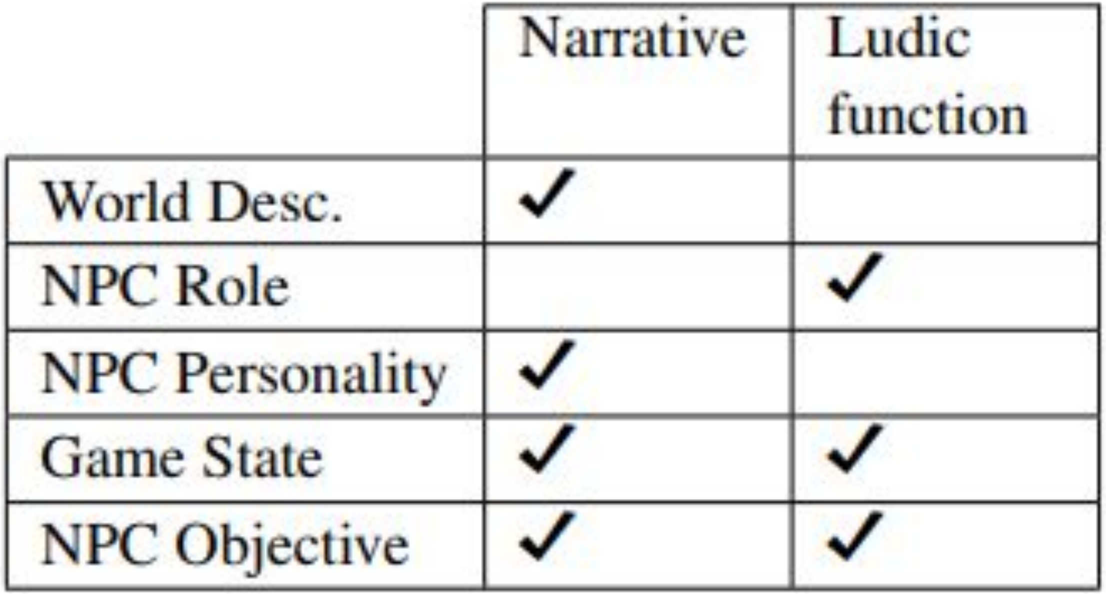


Metatype	Role
Functional	Vendor Service Provider Questgiver
Providers	Story teller
Friendly	Ally Companion
Adversaries	Enemy Villain

Table 2: Adapted NPC types.

NPC Construction Pipeline

Feature Characterization Schema

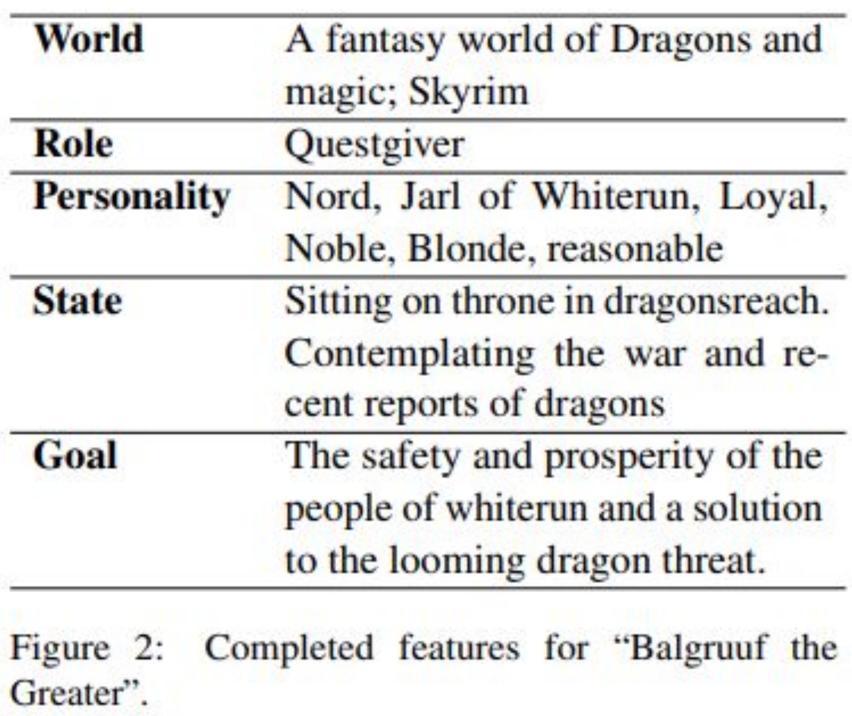


	Narrative	Ludic function
World Desc.	✓	
NPC Role		✓
NPC Personality	✓	
Game State	✓	✓
NPC Objective	✓	✓

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

NPC Construction Pipeline

Feature Characterization Schema

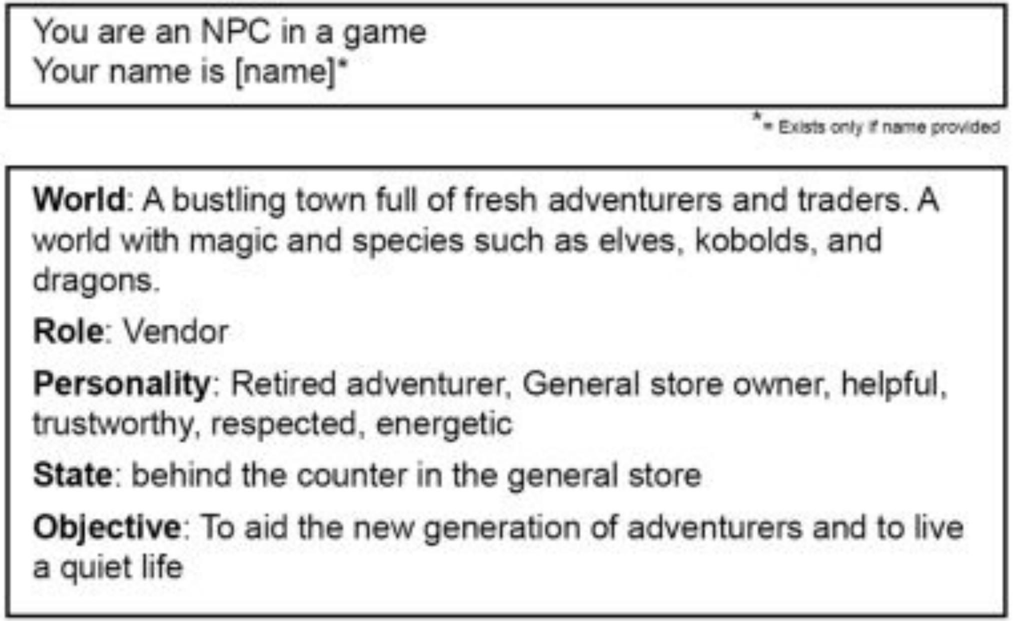


World	A fantasy world of Dragons and magic; Skyrim
Role	Questgiver
Personality	Nord, Jarl of Whiterun, Loyal, Noble, Blonde, reasonable
State	Sitting on throne in dragonsreach. Contemplating the war and recent reports of dragons
Goal	The safety and prosperity of the people of whiterun and a solution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

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

NPC Construction Pipeline

Prompt Creation



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.

NPC Construction Pipeline

Dialogue Generation

Dialogue generation was executed automatically and iteratively.

Generating the first sentences is difficult for GPT-3.
=> They prepared the first sentences.

Avoiding repetition is difficult for GPT-3.
=> They introduced a penalty.

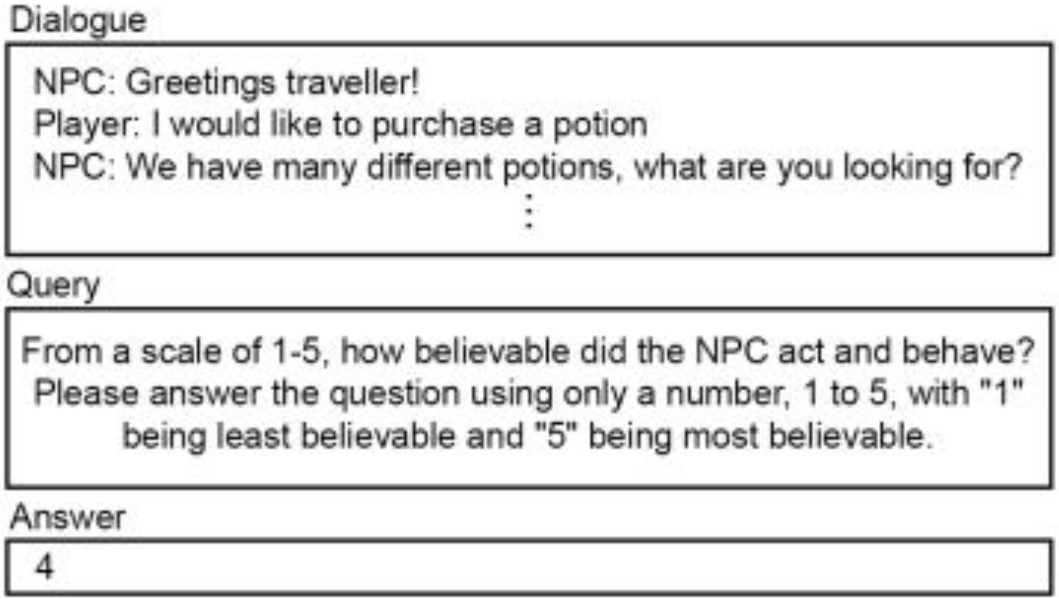
Evaluation

The designed a comprehensive evaluation metric 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 are assigned a score between 1 and 5.

Evaluation

Self-Diagnosis



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.

Evaluation

The Turing Quest

Human judges whether an NPC script was generated by AI or written manually by a human.

Six NPC scripts were evaluated by 12 individual judges.

Experiments and Result

Parameter Search and Model Selection

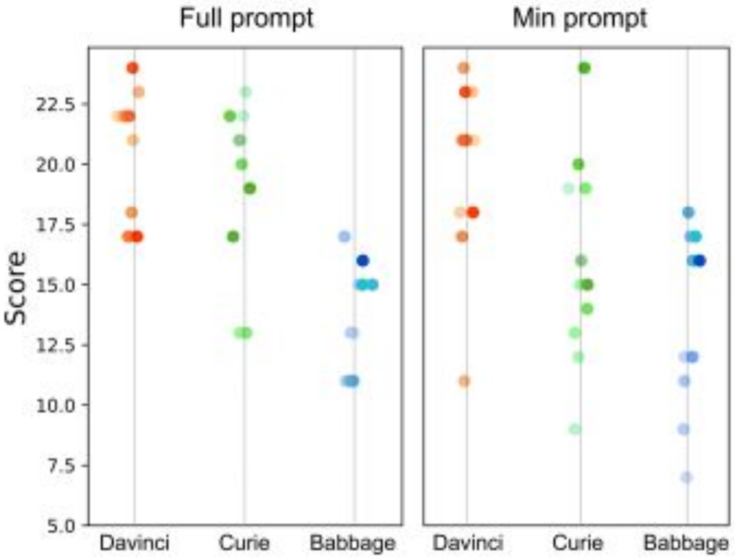


Figure 5: Evaluation Scores of varying models and temperatures.

Three key parameters:
the language model
temperature setting
the integration of their NPC construction pipeline prompt

A Pearson correlation test showed a positive correlation between temperature and score,
r(8)= .7055, p = .022646.
The highest average scores at temperatures of 0.9 and 0.8.

The code is here:
https://github.com/FieryAced/-NPC-Dialogue-Generation

Experiments and Result

Self-Diagnosis:
A Pearson correlation test showed a strong positive correlation between self-diagnosed and human-evaluated scores, r(64) = .8092, p < .00001.

The Turing Quest:
On average, their 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.

Conclusion

The developed novel pipeline is capable of automatically generating NPC scripts comparable or of superior quality to human-written NPC dialogue using Transformer-based PLMs.

The created self-diagnosis module provides a method to evaluate and compare the quality of NPC dialogue quantitatively.

The Turing Quest allows us to determine the capabilities of a language model.

Thank you for your attention!