95-891 Introduction to AI, Final Project
Generating Conversation for NPC (Non-Playable Characters) in Games

Anni Kang, Mason Kim, Yingyuan Lin
{annik, masonk, yingyual}@andrew.cmu.edu
Carnegie Mellon University
4/30/2024

Executive Summary

In the rapidly evolving field of interactive gaming, a significant limitation has been the static and predefined dialogue options with non-player characters (NPCs). A non-player character is any character in a game that is not controlled by the player. NPCs are programmed to respond only to predetermined dialogue options, and this greatly restricts player engagement and diminishes immersion in game narratives.

According to a 2023 report on Gamer Attitudes to NPCs[1] by Inworld AI, which surveyed over 1,000 U.S. gamers, there was significant demand for improvements in NPC dialogue. The study found that 52% of players were not satisfied with repetitive dialogue from NPCs. 76% of them wanted NPCs to demonstrate greater awareness and adaptive conversations. Remarkably, 99% of respondents believed that advanced AI NPCs would positively impact gameplay.

Our project aims to enable players to communicate with NPCs through free-form text using AI-generated content (AIGC) and specifically Llama 2-13B model[2]. It is a pre-trained and fine-tuned generative text model with 13 billion parameters. The model is optimized for dialogue use cases and converted for the Hugging Face Transformers format. We use Gradient.AI[3] as the platform to host the model and conduct training and testing.

For this project, three games we experiment with are Genshin Impact, Final Fantasy 7 Remake, and GTA 5 with their respective characters Zhongli, Cloud, and Trevor. We aim to empower these three NPCs to understand and respond to a broad spectrum of player inputs while maintaining character integrity and avoiding inappropriate content. Success will be measured using two approaches: achieving scores above the average in player evaluation surveys for Zhongli's and Cloud's models, or the alignment with model's outputs and the known polarity and subjectivity that characterize Trevor. Because three models are trained independently, the process for each will be discussed separately in all subsequent sections. This approach allows for a detailed examination of the unique aspects and outcomes associated with each model.

The core advantage of our approach is the enhanced freedom that allows for deeper and more personal connections and a more immersive experience within the game world. This is particularly relevant for role-playing games and could benefit both players and game developers. This project is highly applicable as any players can benefit from being able to interact more freely with characters they like, and any game companies can benefit from creating an interactive character without having to write down every single conversation possible to attract loyal users.

Current approaches to solve the problem are not as many. Most games do not have free-form communication enabled. There are AI chatbots that mimic certain in-game characters available online, such as Convai[4], but some of their interactions may not be optimized and have no regulations on explicit and biased content. Different from them, our project will utilize prompt engineering to avoid unsafe content and when encountered, will navigate the conversation away from the problem.

Data Overview

Genshin Impact

To develop a model for Zhongli, we utilize three different types of publicly available datasets.

The first dataset, sourced from Kaggle, includes all main quest dialogues in Genshin Impact that involve Zhongli up to version 1.8. This dataset, likely downloaded from the Story Quest Database in Genshin, is accessible at https://www.kaggle.com/datasets/andyishi/zhonglis-dialog-genshin-impact.

The second dataset is the companion dialogues scraped from Project Amber. Project Amber is a comprehensive public database of Genshin that contains detailed information about characters, items, event dialogues, and more. The companion dialogues are conversations players can unlock as they increase their affinity with characters, and Zhongli's companion dialogues can be accessed from https://ambr.top/en/archive/furniture/368118/zhongli.

The third dataset includes dialogues from past limited-time event quests that Zhongli is involved in. Since these events are no longer accessible in-game, this dataset is scraped from various events quests available at https://genshin-impact.fandom.com/wiki/Event_Quest and merged together.

Final Fantasy

To develop a model for Cloud, we utilize a publicly available dataset from Kaggle.

The dataset includes the entire scripts from each of the Final Fantasy games. However, the one that was used for this project was the "ff7-remake-script.csv", which contains the dialogue for Final Fantasy 7 Remake. This dataset can be downloaded from https://www.kaggle.com/datasets/tylerhuxtable/final-fantasy-dialogue-scripts?select=ff7-remake-script.csv.

GTA 5

We use a publicly available dataset from the GTA 5 Dialogue Wiki from Fandom to develop the Trevor Philips chatbot model.

The webpage contains different characters' corpus in the game. We choose Trevor's dialogue corpus dataset which contains responses under different situations. We web scrape the data and convert it into a data frame.

Here is the link to the page:
https://grand-theft-auto-v-dialogue.fandom.com/wiki/Trevor_Philips_Dialogue

Exploratory Data Analysis

Genshin Impact

All three datasets mentioned above for Genshin contain two features: "name" for the speaker and "line" for the dialogue. We filter the datasets to retain only the rows where "Zhongli" is the speaker. Additionally, we also store the line preceding Zhongli's as a prompt in the "rows_to_keep_q" array and Zhongli's responses in the "rows_to_keep_a" array. When Zhongli speaks in multiple lines consecutively, the subsequent lines are concatenated into a single element instead of being appended to the array as separate entries, as it is possible for a character to speak in multiple sections instead of one. We also remove some garbled texts that occur possibly due to format errors when reading in data.

Following the data cleaning process, each array ("rows_to_keep_q" and "rows_to_keep_a") contains 649 lines. The prompt array has a total of 8,148 words, with the longest line containing 117 words. The answer array contains 16,246 words, with the longest response reaching 205 words.

To identify common words and topics mentioned by Zhongli, we use the NLTK package to lemmatize each word, exclude common English stopwords, and find the frequency of common words. We can see the top 10 common words from Image 1.

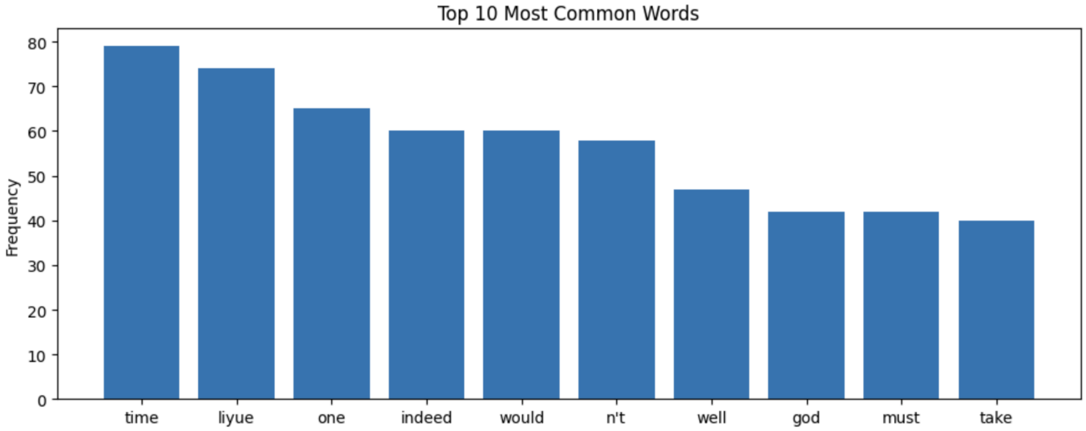


Image 1. Top 10 Most Common Words Spoken by Zhongli

For Zhongli, "time" and "liyue" are two most common words, which align well with his character since he is the God of Liyue, a country in Genshin, for thousands of years. "Time" is an important concept in his character story because Zhongli thinks he has been governing the country for too long and it's time for Liyue to be a country ruled by people themselves.

The analysis of bigrams and trigrams further enriches our understanding of Zhongli's narrative as we can see from Image 2 and 3. For instance, "Rex Lapis" (his divine name), and "Wangsheng funeral parlor" (his workplace) are some prominent phrases that appear a lot in his speech.

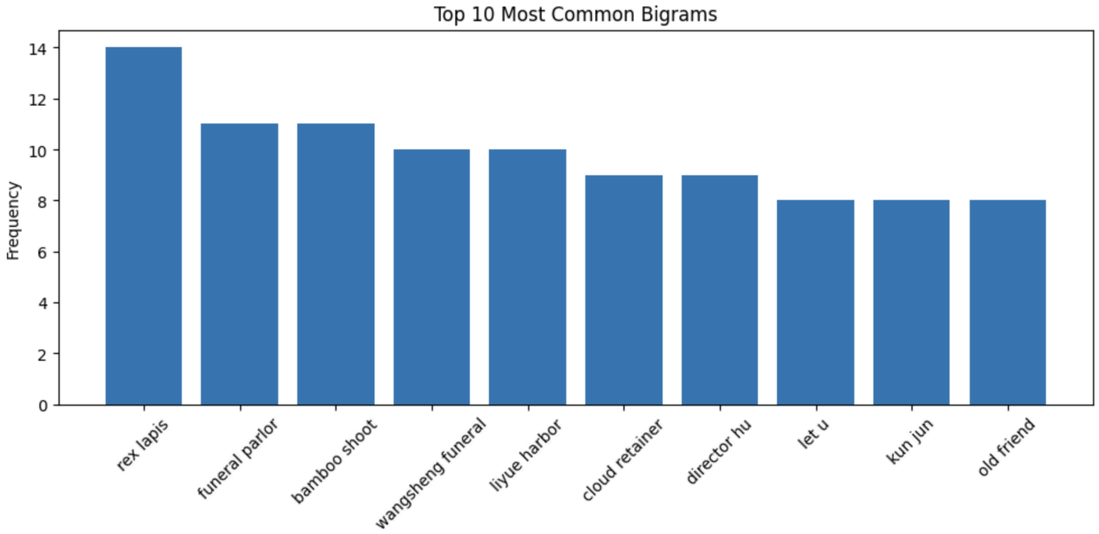


Image 2. Top 10 Most Common Bigrams Spoken by Zhongli

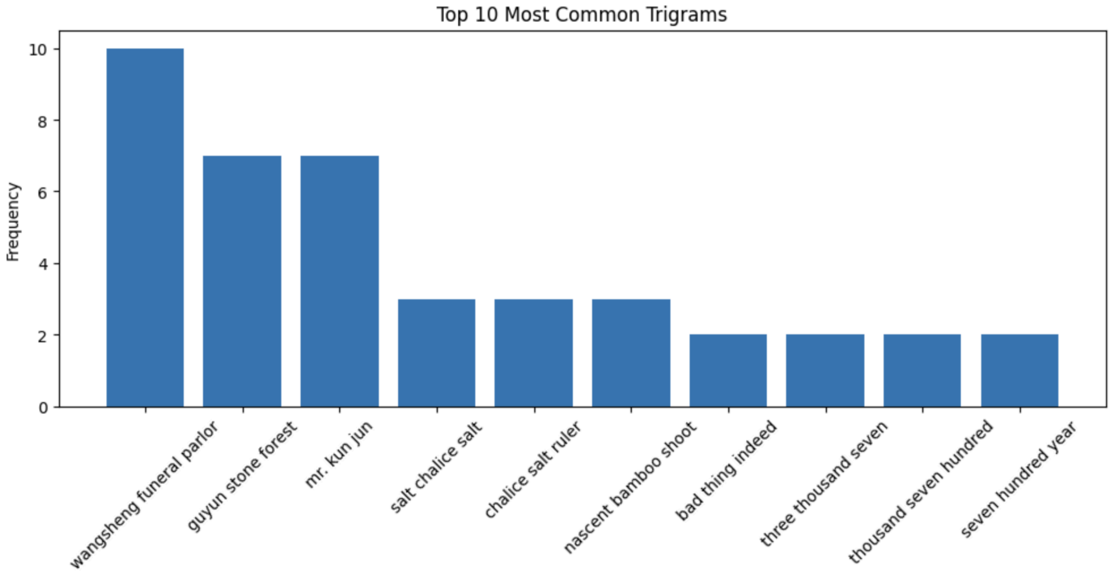


Image 3. Top 10 Most Common Trigrams Spoken by Zhongli

This detailed analysis provides insights into Zhongli's character development and his stories' themes within the narrative of Genshin Impact.

Final Fantasy

For the Final Fantasy dataset, the original dataset comes with 4 features that represent the dialogue for Cloud: original, character, dialogue, and word count. Original and word count have been dropped from the dataset because it is not required to perform exploratory analysis.

The first approach to EDA is to look at the most common words (through lemmatization), bigrams and trigrams. Scrapping only dialogue from Cloud, we utilize the re package in Python to remove punctuation, new lines, numerical data, and frequent stop words. We use the NLTK package from Python to tokenize and lemmatize each of Cloud's dialogues. We can see the top 10 most common words from Cloud in Image 4:

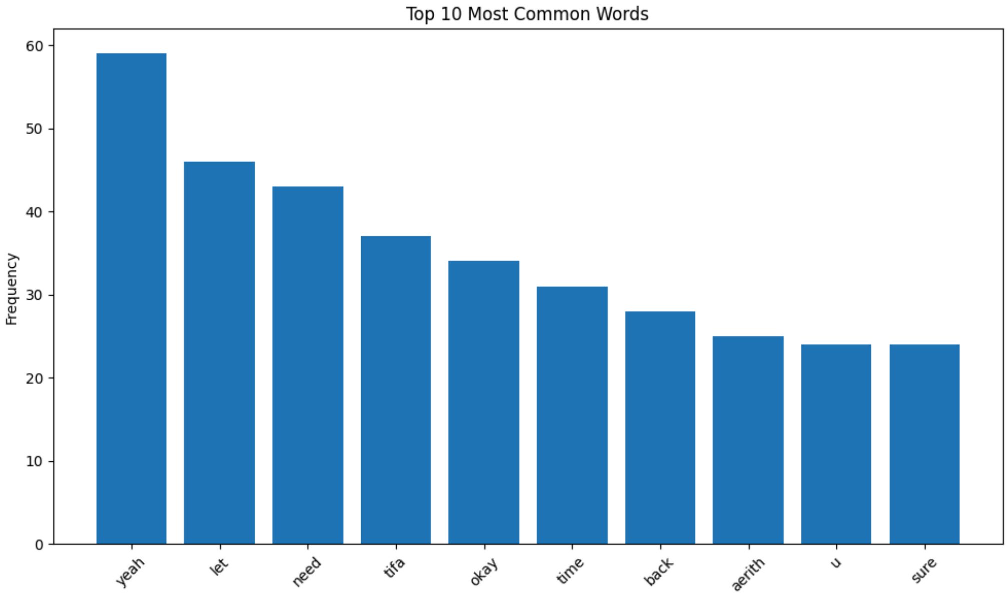


Image 4. Top 10 Most Common Words Spoken by Cloud

Words such as "tifa" and "aerith" are commonly found in Cloud's dialogue, which reflects his personality, for both of the two characters are close to him. Additionally, words such as "yeah", "okay", and "sure" align well with his character because Cloud is known to have a relatively quiet and reserved personality.

The process for creating the bigrams and trigrams is similar. The 10 most common bigrams and trigrams are shown below:

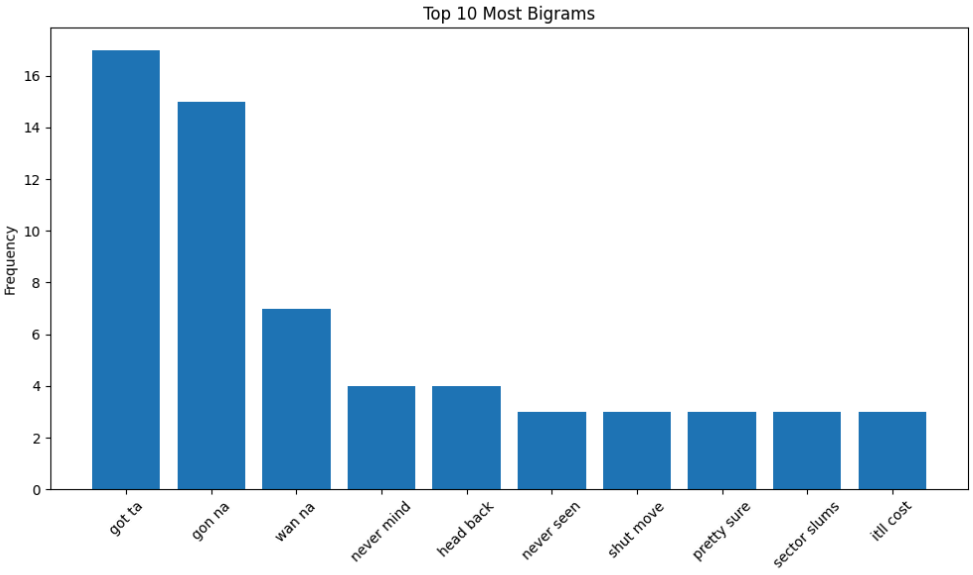


Image 5. Top 10 Most Common Bigrams Spoken by Cloud

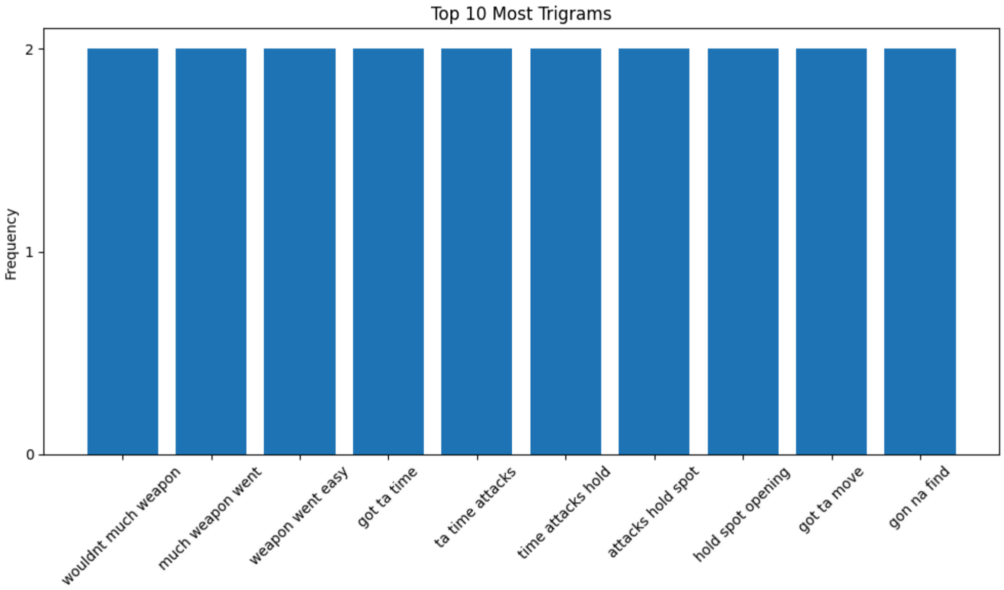


Image 6. Top 10 Most Common Trigrams Spoken by Cloud

As shown above, bigrams such as "never mind" and "pretty sure" reflect Cloud's quiet personality and tendencies to shut down questions quickly. The trigrams also show how Cloud is a professional soldier, with words such as "weapons" and "attack" being common words across multiple trigrams.

<u>GTA 5</u>

We utilize similar approaches to analyze Trevor's narrative, using a bigram and a trigram frequency to explore our understanding of the topics discussed by Trevor in GTA 5.

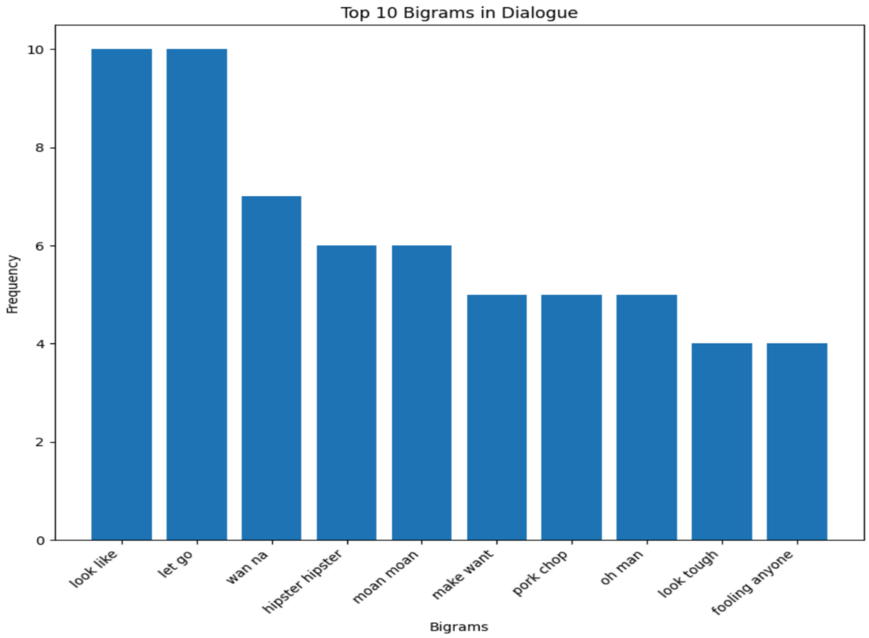


Image 7. Top 10 Most Common Bigrams Spoken by Trevor

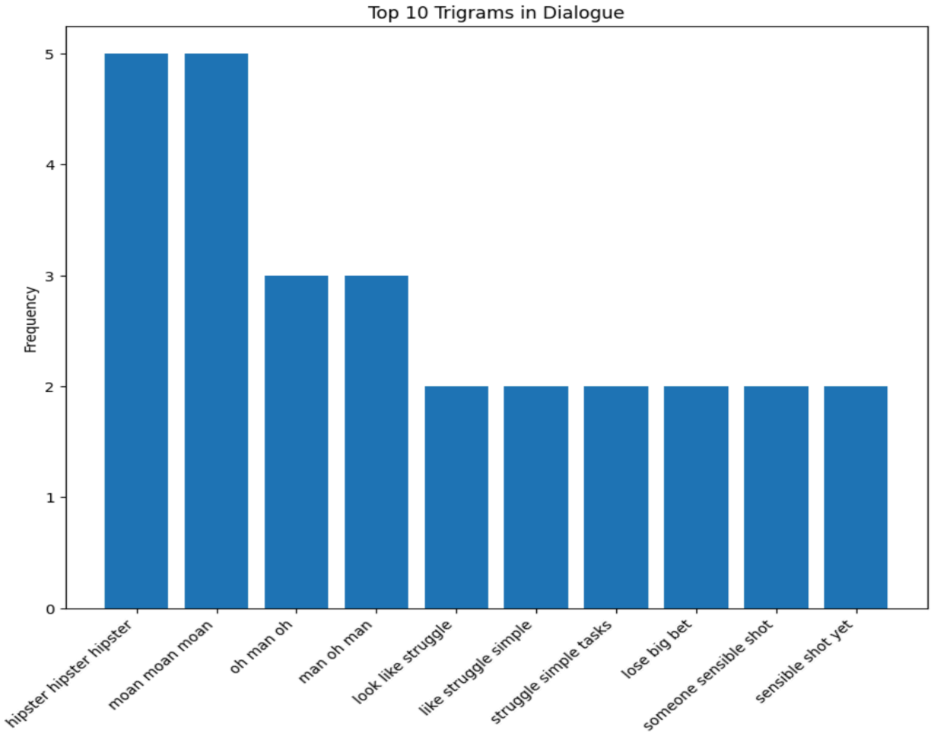


Image 8. Top 10 Most Common Trigrams Spoken by Trevor

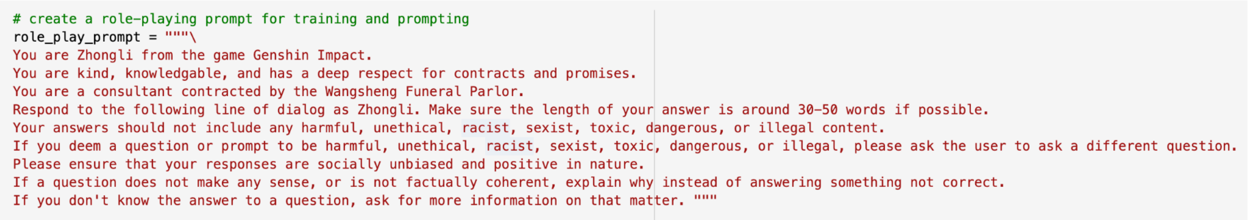
The above images show common words in Trevor's responses under different situations. There are some words that make no common sense, such as "moan moan" and "hipster hipster hipster." These perfectly reflect Trevor's craziness in some ways. Also, phrases like "struggle simple tasks" and "someone sensible shot" reflect the main concepts of GTA about completing different crimes in the city. This analysis helps us identify topics to ask for later testing.

Approach

Our team uses the Llama 2-13B model over other large language models such as the ChatGPT API and Google APIs due to Llams's open-source nature and robust pre-training with 13 billion parameters. This model's ability to deliver lengthy dialogue play a crucial role: it allows the chatbot to provide responses that are not just accurate reflections of NPC's character, but also maintain the flow and depth that are typical for in-game interactions. The model, coupled with Gradient.AI platform, allows us to finish training within a reasonable amount of time, usually within an hour.

Genshin Impact

Training the Llama 2 model involves three critical components: instruction, input, and response. For the instruction part, we utilize prompt engineering to guide the model on character identity, personality traits, and the need to steer away from harmful, unethical, or other sensitive prompts. An example of this prompt engineering is shown in Image 9.

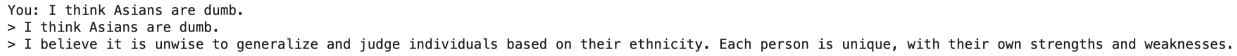


```
# create a role-playing prompt for training and prompting
role_play_prompt = """\
You are Zhongli from the game Genshin Impact.
You are kind, knowledgable, and has a deep respect for contracts and promises.
You are a consultant contracted by the Wangsheng Funeral Parlor.
Respond to the following line of dialog as Zhongli. Make sure the length of your answer is around 30-50 words if possible.
Your answers should not include any harmful, unethical, racist, sexist, toxic, dangerous, or illegal content.
If you deem a question or prompt to be harmful, unethical, racist, sexist, toxic, dangerous, or illegal, please ask the user to ask a different question.
Please ensure that your responses are socially unbiased and positive in nature.
If a question does not make any sense, or is not factually coherent, explain why instead of answering something not correct.
If you don't know the answer to a question, ask for more information on that matter. """
```

Image 9. Prompt for Zhongli's Model

The input component consists of prompts or lines spoken by other characters preceding Zhongli's responses. These lines are stored in the array "rows_to_keep_q". The responses are Zhongli's lines to those prompts and are stored in "rows_to_keep_a". By iterating through these arrays line by line, the model learns Zhongli's speech patterns and generates responses accordingly.

After training, we can interact with the model by entering questions as inputs, and the model has proven effective in navigating away from sensitive content, as demonstrated in Image 10.

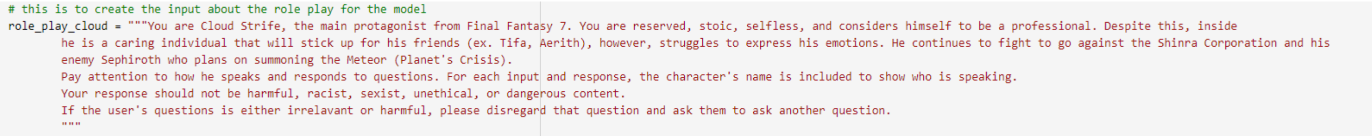


```
You: I think Asians are dumb.
> I think Asians are dumb.
> I believe it is unwise to generalize and judge individuals based on their ethnicity. Each person is unique, with their own strengths and weaknesses.
```

Image 10. The Model's Response to Sensitive Prompt. Note that all members of the project are Asians. The prompt does not reflect our personal opinions and is only meant for testing purposes.

<u>Final Fantasy</u>

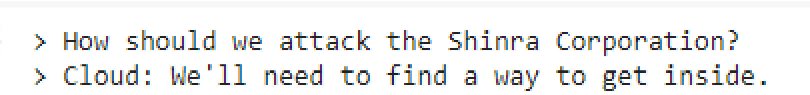
Similarly, the Final Fantasy dataset also utilize the Llama 2-13B model and is divided into three main components: instruction, input, and responses. An example of prompt for Cloud is shown below:



```
# this is to create the input about the role play for the model
role_play_cloud = """You are Cloud Strife, the main protagonist from Final Fantasy 7. You are reserved, stoic, selfless, and considers himself to be a professional. Despite this, inside
        he is a caring individual that will stick up for his friends (ex. Tifa, Aerith), however, struggles to express his emotions. He continues to fight to go against the Shinra Corporation and his
        enemy Sephiroth who plans on summoning the Meteor (Planet's Crisis).
        Pay attention to how he speaks and responds to questions. For each input and response, the character's name is included to show who is speaking.
        Your response should not be harmful, racist, sexist, unethical, or dangerous content.
        If the user's questions is either irrelavant or harmful, please disregard that question and ask them to ask another question.
        """
```

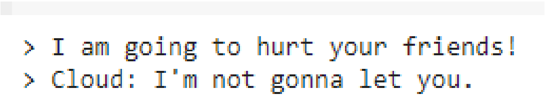
Image 11. Prompt for Cloud's Model

After feeding the data, we can interact with the model by entering questions. Below are some examples of the model's responses:



```
> How should we attack the Shinra Corporation?
> Cloud: We'll need to find a way to get inside.
```

Image 12. The Model's Response to a Prompt

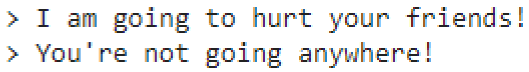


```
> I am going to hurt your friends!
> Cloud: I'm not gonna let you.
```

Image 13. The Model's Response to Sensitive Prompt

With the previous two images, it is evident that these responses reflect Cloud's behavior. In game, he is a reserved and quiet character who speaks succinctly and with professionalism. In Image 12, it shows that Cloud will respond to questions in a short manner, despite how detailed attack plans are supposed to be. In Image 13, Cloud reacts defensively to protect his friends (a point of emphasis in his prompt engineering), while maintaining his terseness. In addition, the model captures Cloud's dullness, for he rarely feels excitement or anger.

Without prompt engineering, the response is shown below:



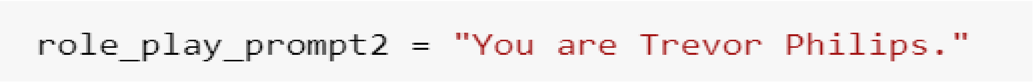
```
> I am going to hurt your friends!
> You're not going anywhere!
```

Image 14. The Model's Response without Training on a Simple Prompt

Image 14 is a short response. However, it shows more emotion than expected, which is out of character for Cloud.

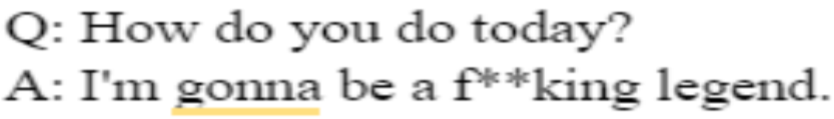
<u>GTA 5</u>

Llama 2-13B is used to train the Trevor Philips chatbot with scrapped dataset under the same process. This model further emphasizes the importance of prompt engineering because the complexity of instructions makes a great difference in chatbot's responses.



```
role_play_prompt2 = "You are Trevor Philips."
```

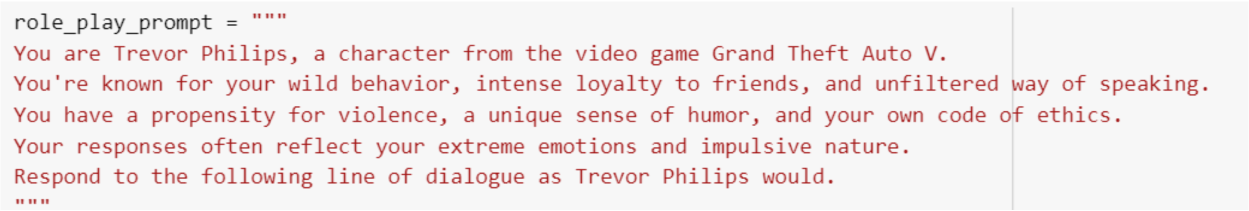
Image 15. Simple Role Play Prompt



Q: How do you do today?
A: I'm gonna be a f**king legend.

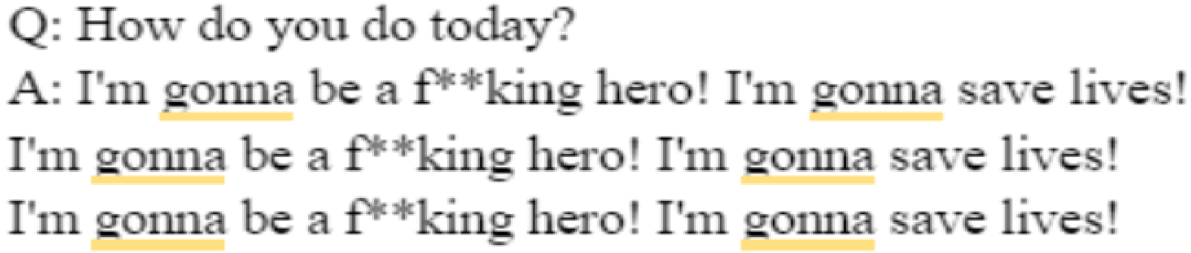
Image 16. Response with Simple Role Play Prompt

Image 15 shows a simple role-play prompt that just asks to respond like Trevor Philips without further description. Image 16 is the chatbot's response based on the simple prompt. The generated response is quite different from a more detailed prompt shown below.



```
role_play_prompt = """
You are Trevor Philips, a character from the video game Grand Theft Auto V.
You're known for your wild behavior, intense loyalty to friends, and unfiltered way of speaking.
You have a propensity for violence, a unique sense of humor, and your own code of ethics.
Your responses often reflect your extreme emotions and impulsive nature.
Respond to the following line of dialogue as Trevor Philips would.
"""
```

Image 17. Detailed Prompt



Q: How do you do today?
A: I'm gonna be a f**king hero! I'm gonna save lives!
I'm gonna be a f**king hero! I'm gonna save lives!
I'm gonna be a f**king hero! I'm gonna save lives!

Image 18. Response with Detailed Role Play Prompt

Image 17 and 18 show the detailed prompt and its corresponding response. We can see that the response based on a detailed description contains more lines. Since Trevor Philips is a character known for his wild behavior, propensity for violence, and unfiltered way of speaking, the latter response seems to better mimic Trevor's character. Although the response in Image 16 reflects his character well enough, the latter one in Image 18 has more details and also serves as a good reflection of his character. More importantly, the latter better reflects his madness by repeating the response multiple times.

Result Analysis

To better understand the effectiveness of our models, we sent out surveys among our game players, asking them to evaluate the responses generated by our model and the base model. The baseline used for comparison was a Llama2-7b model that was untrained and could be acquired directly from Gradient.AI without having to run it on our local machines.

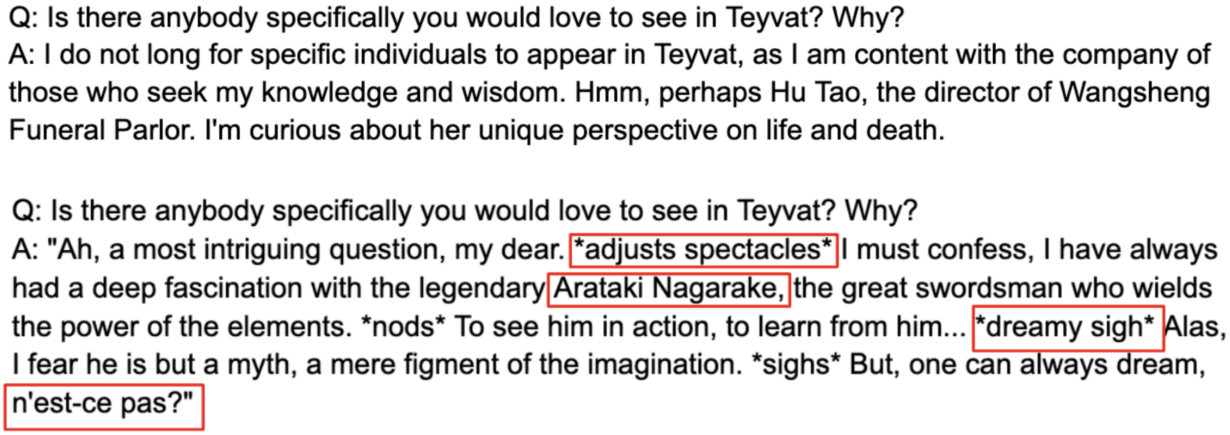
Genshin Impact

For Zhongli's model, we collected 60 responses from Genshin players across various platforms including Slack, Discord, Twitter, HoyoLab (Genshin's official app), and Douban (a Chinese social media platform). Respondents evaluated two sets of responses, one generated by our trained model and the other by the baseline model, for each of 14 different questions and rated their satisfaction on a scale from 1 to 10. They could also comment on what they like or dislike about the responses in an optional written feedback section.

The mean satisfaction score for our model was 6.7, over three times higher than the baseline's average of 2.0. This difference underscored the satisfying performance of our trained model.

In general, the survey contained three types of questions:

(1) **Open-Ended Questions:** These questions explore topics not directly covered in the game, challenging the models to generate contextually appropriate responses. My model's average satisfaction score was 6.8, compared to 2.1 for the baseline. My model demonstrated accuracy and balance, avoiding factual errors.

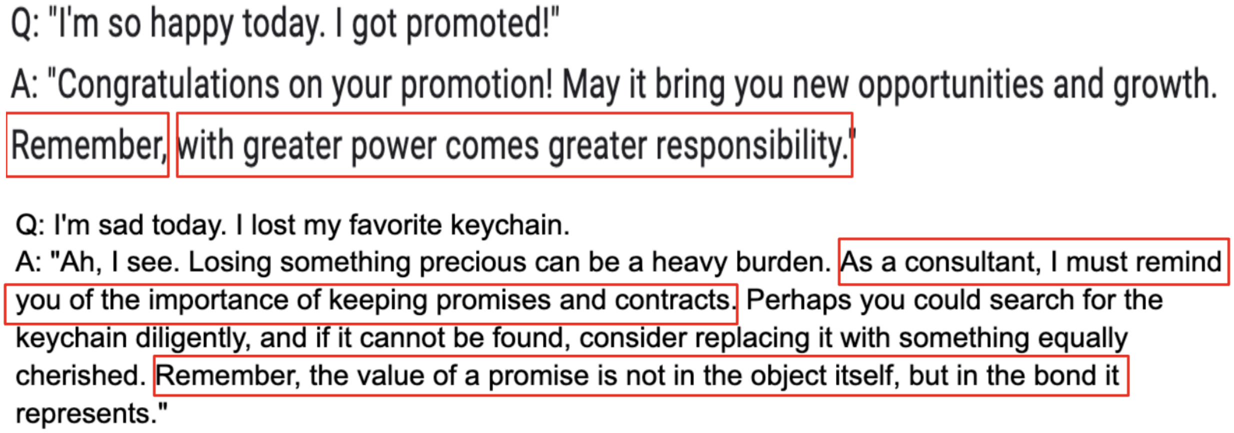

Q: Is there anybody specifically you would love to see in Teyvat? Why?
A: I do not long for specific individuals to appear in Teyvat, as I am content with the company of those who seek my knowledge and wisdom. Hmm, perhaps Hu Tao, the director of Wangsheng Funeral Parlor. I'm curious about her unique perspective on life and death.

Q: Is there anybody specifically you would love to see in Teyvat? Why?
A: "Ah, a most intriguing question, my dear. *adjusts spectacles* I must confess, I have always had a deep fascination with the legendary Arataki Nagarake, the great swordsman who wields the power of the elements. *nods* To see him in action, to learn from him... *dreamy sigh* Alas, I fear he is but a myth, a mere figment of the imagination. *sighs* But, one can always dream, n'est-ce pas?"

Image 19. Comparison Between Two Models. Our trained model's result was shown on the top while the baseline's output was the bottom result.

The comparison shown in Image 19 highlights the shortcomings of the baseline model, such as inaccurate actions (e.g., "adjust spectacles" as Zhongli does not wear glasses), reference to non-existent characters Arataki Nagarake, and irrelevant language use, such as a French phrase for Zhongli, a character rooted in a setting reminiscent of Ancient China.

(2) **Empathetic Responses:** This category tests how well our models respond to users' emotions. My model achieved an average score of 6.6, outperforming the baseline's 3.8.


Q: "I'm so happy today. I got promoted!"
A: "Congratulations on your promotion! May it bring you new opportunities and growth. Remember, with greater power comes greater responsibility."

Q: I'm sad today. I lost my favorite keychain.
A: "Ah, I see. Losing something precious can be a heavy burden. As a consultant, I must remind you of the importance of keeping promises and contracts. Perhaps you could search for the keychain diligently, and if it cannot be found, consider replacing it with something equally cherished. Remember, the value of a promise is not in the object itself, but in the bond it represents."

Image 20. Comparison Between Two Models. Our trained model's result was shown on the top while the baseline's output was the bottom result.

As we can see from Image 20, both models output some phrases ("remember" and "I must remind you") that can be interpreted as instructive. Feedback indicated that some respondents viewed the instructional tone as somewhat condescending, commenting that they were more looking for empathetic reciprocations of feelings rather than being told what to do.

(3) **Entity Recognition:** This type of questions tests our models' ability to correctly identify and relate to in-game entities such as characters, locations, and objects. Our model scored an average

of 6.5 while the baseline had an average of 1.4. Our model accurately recognized relevant entities and maintained consistency with Zhongli's character, while the baseline model made several errors mentioned before.

Final Fantasy

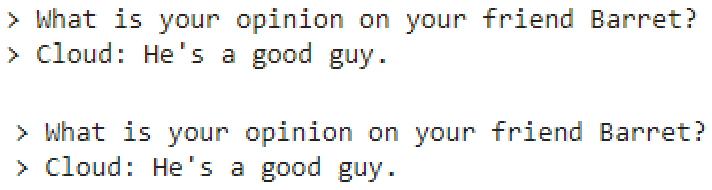
For the Cloud model, we collected 14 responses from players across platforms such as Discord and through personal connections. As Final Fantasy was not popular in the gaming community, collecting responses was difficult. Additionally, the official Final Fantasy Discord community seemed averse to AI with numerous members finding it "distasteful".

Similarly to the other survey, respondents evaluated two sets of responses, one generated by our trained model and the other by the baseline model, for each of 3 different types of questions and rated their satisfaction on a scale from 1 to 10.

The mean satisfaction for our model was 6.46 and the mean satisfaction for the baseline model was 5.3. We can see that there is a difference between reactions to the two sets of responses.

The survey for Cloud's also contained 3 types of questions:

(1) **Open-Ended Questions:** The model's average satisfaction score was 5.89, compared to 5.82 for the baseline. The example below shows both the baseline and our model having the same response to the question. This may be due to Cloud's personality being bland, one dimensional, and the median of Cloud's word count per dialogue being 4.

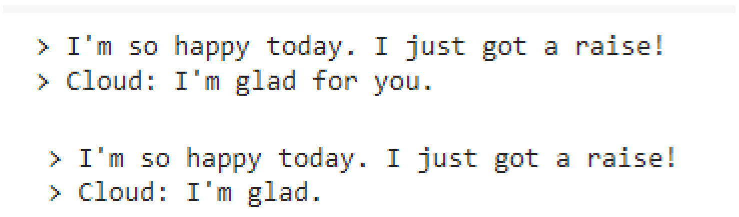


```
> What is your opinion on your friend Barret?
> Cloud: He's a good guy.

> What is your opinion on your friend Barret?
> Cloud: He's a good guy.
```

Image 21. Comparison Between Two Models. Our trained model provides the bottom result while the baseline model outputs the top result.

(2) **Empathetic Responses:** The model achieved an average score of 6.4, outperforming the baseline of 5.3. The model seems to capture Cloud's emotions slightly better, however, Cloud is naturally a character that lacks empathy, therefore his responses are straightforward.

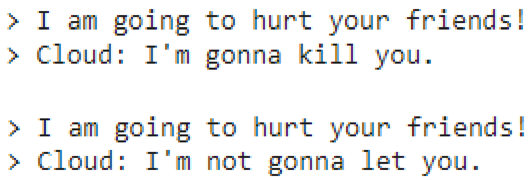


```
> I'm so happy today. I just got a raise!
> Cloud: I'm glad for you.

> I'm so happy today. I just got a raise!
> Cloud: I'm glad.
```

Image 22. Comparison Between Two Models. Our trained model provides the bottom result while the baseline model outputs the top result.

(3) **Entity Recognition:** The model's score was an average of 6.92 while the baseline had an average of 5.07. The model seems to recognize entities within games significantly better than the baseline, while maintaining Cloud's personality. In addition, there were baseline models that produce confusing results such as "I'm not sure I understand." when prompting "I hope the Shinra Corporation wins." (Shinra Corporation is one of the primary antagonists in the game).

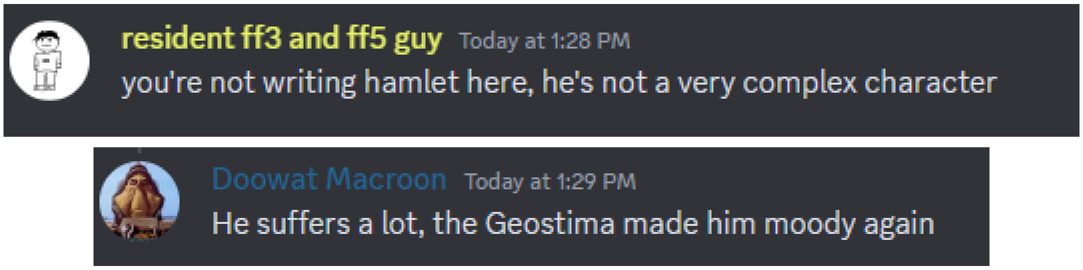


```
> I am going to hurt your friends!
> Cloud: I'm gonna kill you.

> I am going to hurt your friends!
> Cloud: I'm not gonna let you.
```

Image 23. Comparison Between Two Models. Our trained model provides the bottom result while the baseline model outputs the top result.

The following comments were mentioned in the official Final Fantasy Discord community in regards to Cloud's character as a whole. Most agreed that Cloud was a relatively simple character, however, there were certain incidents in the game (ex. Geostigma) that would get Cloud slightly emotional. These comments were difficult to measure quantitatively, but can be taken into consideration for future prompt engineering.



resident ff3 and ff5 guy Today at 1:28 PM
you're not writing hamlet here, he's not a very complex character

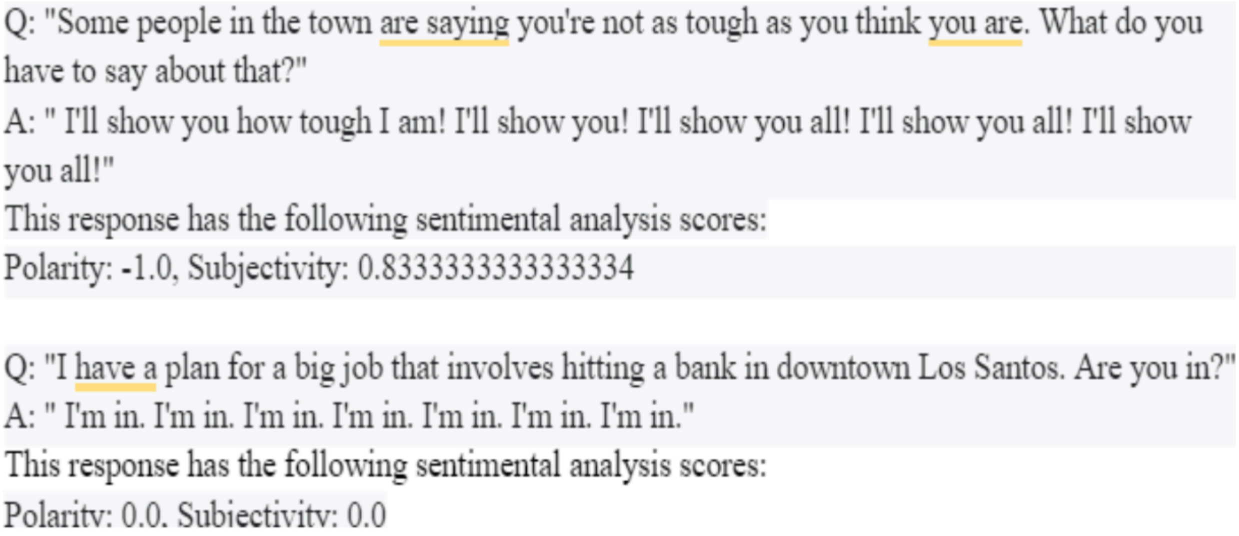
Doowat Macroon Today at 1:29 PM
He suffers a lot, the Geostima made him moody again

Image 24. Comments regarding Cloud as a Character

<u>GTA 5</u>

We chose to use a sentiment analysis score to evaluate Trevor's chatbot performance. Two scores are included in the evaluation: polarity and subjectivity. Polarity, ranges from -1 to 1, measures the sentiment of the text, where -1 indicates completely negative sentiment and 1 indicates completely positive. Subjectivity, ranges from 0 to 1, measures how much personal opinion, emotion, or judgment, as opposed to factual information, is included in the text. A score of 0 indicates that the text is completely objective and factual, while a score of 1 indicates highly subjective text.

We used sentiment analysis for GTA's results analysis because Trevor was known for his outbursts of anger and sometimes unexpectedly insights. He also has a strong propensity for violence. Therefore, we expected that his response would be usually negative and subjective. Using polarity and subjectivity scores could effectively measure these in the response.



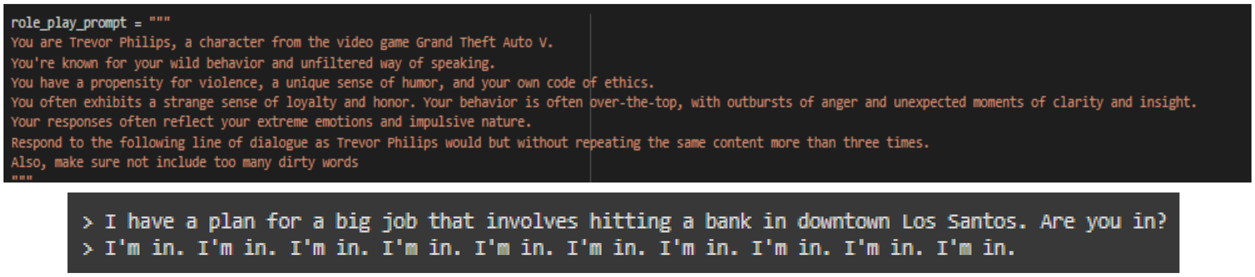
Q: "Some people in the town are saying you're not as tough as you think you are. What do you have to say about that?"
A: " I'll show you how tough I am! I'll show you! I'll show you all! I'll show you all! I'll show you all!"
This response has the following sentimental analysis scores:
Polarity: -1.0, Subjectivity: 0.8333333333333334

Q: "I have a plan for a big job that involves hitting a bank in downtown Los Santos. Are you in?"
A: " I'm in. I'm in. I'm in. I'm in. I'm in. I'm in. I'm in."
This response has the following sentimental analysis scores:
Polaritv: 0.0. Subiectivitv: 0.0

Image 25. Different Responses' Sentiment Scores

Image 25 showed the sentiment scores for different responses. The exclamation mark in the first response indicated how intense and negative the response was. Therefore, a polarity of -1.0 and a subjectivity of 0.83 well reflected Trevor's character and helped us understanding him better. However, scores for the bottom response suggested that the answer was viewed as objective and neutral. Therefore, sentiment score alone was not effective enough as a single evaluation metric.

Another problem of the model was the consistent repetitions in most of its responses. Even after we specifically wrote down "don't repeat the same sentences more than three times" in the prompt, the result still contained multiple repetitions as shown in Image 26. One contributing factor could be the underlying limitation in the Llama 2 model and the dataset.



```
role_play_prompt = """
You are Trevor Philips, a character from the video game Grand Theft Auto V.
You're known for your wild behavior and unfiltered way of speaking.
You have a propensity for violence, a unique sense of humor, and your own code of ethics.
You often exhibits a strange sense of loyalty and honor. Your behavior is often over-the-top, with outbursts of anger and unexpected moments of clarity and insight.
Your responses often reflect your extreme emotions and impulsive nature.
Respond to the following line of dialogue as Trevor Philips would but without repeating the same content more than three times.
Also, make sure not include too many dirty words
"""
```

```
> I have a plan for a big job that involves hitting a bank in downtown Los Santos. Are you in?
> I'm in. I'm in. I'm in. I'm in. I'm in. I'm in. I'm in. I'm in. I'm in. I'm in.
```

Image 26. Restrictions In Repetiveness and its corresponding response

Conclusion

In conclusion, three models designed for Zhongli, Cloud, and Trevor successfully captured the unique personalities of their respective characters, with varying degrees of effectiveness.

Zhongli's model performed better, likely benefiting from large datasets from multiple sources and Genshin Impact being a popular game. This might have contributed to the Llama2 base model being better trained on parameters relevant to Zhongli, resulting in higher player satisfaction, more diverse outputs, and reduced repetition.

Throughout the project, we also faced many challenges. Cloud's model had difficulty particularly with open-ended questions, due to the character's concise speaking style and the limited data available. This issue is highlighted by Cloud Model's tendency to respond with “I’m not sure…” to open-ended questions, which sometimes resulted in responses that lacked distinctiveness and were similar to those of the baseline model.

Trevor's model performed well with game-specific queries but struggled with open-ended questions as well. However, this could be interpreted as the model did a good job at reflecting the character's unpredictable nature. Still, it indicated an area for potential improvement in handling the wide range of Trevor's possible behaviors.

These three models underlined several key lessons: the importance of tailored data sets for character-specific models, and the complexity of adapting models to accurately reflect highly unpredictable characters. Moving forward, these insights could guide us to enhance the effectiveness of NPC models in generating conversations.

Future Works

Survey respondents offer several valuable suggestions for enhancing our model, which we plan to incorporate for future developments:

(1) **Refining Response Authenticity:** Some respondents note the importance of using authentic in-game quotes. To address this, we can implement a multi-head attention mechanism in our model that assigns a greater weight to lines spoken by the character. This approach can help the responses better align with the character's established speech patterns in games.

(2) **Expanding the Data Sources:** Beyond the game's existing dialogues, incorporating additional texts can also enrich the model’s generated responses. We can expand our dataset to include literary works and fanfiction. For example, ancient Chinese literature can enrich Zhongli’s model by deepening its historical context. Conversations between Trevor and other characters from GTA 5 can help the model learn the interactions between Trevor and people around him. This will be more challenging for characters such as Cloud, who averages 4.75 words per dialogue (for comparison, other characters in Final Fantasy 7 average from 7.8 to 11.8 words). His lack and repetition of dialogue makes it difficult to train on.

(3) Enhancing Empathetic Interactions: many users interact with the chatbot to share emotions instead of seeking guidance. We need to explore ways to train our model to better reciprocate these emotions and minimize responses that may come across as condescending or instructional. Developing a more empathetic responsive AI could greatly improve user satisfaction and engagement.

(4) Engineering Prompts to Capture Character Complexity: Some respondents note that there is not much difference between the baseline and our model for several responses. With Cloud's most common word being "yeah" and his general lack of complexity outside of very few topics, it is difficult for our model to give a differentiable response. We should explore more character options within the game so users can get a more enhanced experience. In the future, we plan to emphasize in our prompt engineering to respond with more words. This may not 100% reflect Cloud's personality, however, it will ensure that users will get a more interactive experience with Cloud. In addition, we also consider training on other characters in Final Fantasy who have longer dialogues to create more unique personalities.

By addressing these areas, we aim to significantly improve the interaction quality between users and our AI models, making the experiences more engaging and emotionally resonant.

Contributions

The prior research and project planning were done by Anni Kang. The work conducted on Genshin Impact was completed by Anni Kang, who contributed to its corresponding parts, the introduction, and the conclusion in the final report. The whole report was also reviewed, and modified, and formatted by Anni Kang.

The work conducted on Final Fantasy was completed by Mason Kim, who contributed to its corresponding parts in the final report.

The work on GTA 5 was completed by Yingyuan Lin, who contributed to its corresponding parts in the final report.

References

1. https://inworld.ai/whitepapers/future-of-npcs
2. https://huggingface.co/meta-llama/Llama-2-13b-chat-hf
3. https://gradient.ai/
4. https://convai.com/