COSE474-2024F: Final Project Proposal
"Contextually Aware Real-time NPC Dialogue Generating LLMs for Enhanced Game Immersion"

2021170964 Kyoungbin Park

1. Introduction

In recent years, subculture games such as Genshin Impact and Wuthering Waves have gained significant popularity. Concurrently, the gaming industry has witnessed the increasing integration of AI technologies, sometimes leading to unfortunate instances of developer layoffs.

The pervasive influence of AI in game development, coupled with the substantial market for games, underscores the critical importance of AI integration in this field. Games like Genshin Impact and Wuthering Waves heavily rely on storytelling and narrative elements. The degree of player immersion in the story and empathy with characters can significantly impact a company's monthly revenue. This context has sparked interest in exploring how current pre-trained Large Language Models (LLMs) can generate and maintain dialogues that align with these complex game narratives.

2. Problem definition & challenges

In subculture games, interactions typically involve NPCs (Non-Player Characters) initiating dialogue, followed by players selecting from a predetermined set of 1-3 response options to progress the story. While different dialogue choices may not substantially alter the overall narrative direction, it is crucial that all options remain within the bounds of the intended storyline.

The core objective of this project is to implement a system where players can freely input their desired responses to NPC dialogue, with the NPCs providing real-time, tailored replies. Successfully addressing this challenge could significantly enhance player immersion and interest in gameplay, as players experience more dynamic and personalized interactions within the game world.)

The primary challenges of this project include:

- **Consistency**: Maintaining character-specific speech patterns and personality traits throughout dialogues.
- **Appropriateness**: Avoiding controversial or explicit content in generated responses.
- **Coherence**: Ensuring LLM-generated responses remain relevant to the story, preventing topic drift and adhering to the narrative intentions of the game's writers.

To address these challenges, it will be necessary to develop finely-tuned LLM models for each NPC character, incorporating appropriate story elements, roles, and personality traits.

3. Related Works

In recent years, dialogue generation in large-scale AI models has seen significant advances. This section outlines several key studies and methodologies in this field, which inform the approach for this project.

Enhancing Dialogue Generation Through Iterative Prompt with Difference Analogy Learning (Park et al., 2024) introduces the **DataDialogue** dataset and utilizes **In-Context Learning** and the **Iterative Prompt Method**. This methodology aims to refine prompts iteratively, enhancing the quality of dialogue generation. The evaluation metrics employed in this work include **BLEU**, **ROUGE-L**, **CIDEr**, and **Dist metrics**, which provide a comprehensive assessment of the generated dialogue's coherence, fluency, and relevance.

Evaluating Factual Consistency of Texts with Semantic Role Labeling (Fan et al., 2023) focuses on ensuring factual consistency in generated texts. The core methodology involves applying **Semantic Role Labeling (SRL)** to evaluate how accurately generated texts capture the roles of entities and their actions, providing a framework for ensuring the factual integrity of AI-generated dialogues.

Large Language Models Meet Harry Potter: A Bilingual Dataset for Aligning Dialogue Agents with Characters (Chen et al., 2023) introduces the **Harry Potter Dialogue Dataset** for training AI agents to align with character-specific dialogues in a fictional setting. The

primary methodology involves **Attributes and Relations Construction**, which aims to create a dialogue structure that remains true to the narrative while allowing for natural interactions between characters and players.

Learning to Memorize Entailment and Discourse Relations for Persona-Consistent Dialogues (Chen et al., 2023) proposes a learning methodology aimed at memorizing entailment and discourse relations. This approach ensures that persona-consistent dialogues are generated, meaning that the characters maintain a coherent and stable personality throughout interactions, a key feature in narrative-driven games.

The Effect of Context-Aware LLM-Based NPC Dialogues on Player Engagement in Role-Playing Video Games (Csepregi, 2023) introduces a context-aware system for **NPC dialogue generation**. The integration of this system with large language models aims to enhance player engagement by ensuring that non-playable characters (NPCs) interact in a contextually appropriate manner within the narrative framework, thus increasing immersion and engagement.

4. Datasets

In this project, **Harry-Potter-Dialogue-Dataset** is used for fine-tuning LLM Models and standard for evaluating their purposes.

Harry Potter Dialogue is a dialogue dataset that integrates with scene, attributes and relations which are dynamically changed as the storyline goes on, which is deliberately designed to be used for researches on more human-like conversational systems in practice. For example, virtual assistant, NPC in games, etc. Moreover, HPD can both support dialogue generation and retrieval tasks.

It provides information about each character's 13 attributes such ad Gender, Age, Belongings, Hobby and Spells. Information about relations between characters is also given, which lets LLM to create more appropriate dialogues regarding to the context of the full story.

5. State-of-the-art methods and baselines

5.1. Methodologies

Model Comparison
To identify the most effective model for dialogue generation, we will evaluate multiple models such as Llama3.2 and Gemma. The comparison will focus on performance metrics and how well each model handles dialogue generation. The key question is: which model performs best, and how significantly does it outperform the others?

Baseline Performance
Establishing a baseline for model performance is essential for evaluating improvements. The performance of our baseline models will be assessed systematically, and if any limitations arise, we will conduct an analysis to understand why these weaknesses exist.

In-Context Learning and Iterative Prompting
In-context learning will be leveraged by providing models with relevant contextual information to improve dialogue coherence. Additionally, we will employ the Iterative Prompt Method, where model prompts are refined iteratively to enhance the quality of generated responses.

Fine-Tuning of LLMs with Distinct Personalities
We will fine-tune several large language models (LLMs), each tailored to embody distinct personalities. Each character will have their own LLM trained via their own attributes and relations data along with thier scripts along the story provided by the dataset. The goal is to evaluate how well these personality-driven models can maintain consistency and relevance in character dialogues.

5.2. Evaluation

Semantic Similarity
To assess how well the generated responses match the ground truth, we will use the Solar Embedding Model to compute semantic similarity between the generated dialogues and the reference answers.

Semantic Role Labeling
Semantic Role Labeling (SRL) will be applied to evaluate the roles of entities and their actions in the generated responses, ensuring that they align with the narrative structure.

Evaluation Metrics
We will use a variety of established evaluation metrics to measure dialogue quality:

- **METEOR** (Metric for Evaluation of Translation with Explicit ORdering)
- **ROUGE-L** (Recall-Oriented Understudy for Gisting Evaluation)
- **CIDEr** (Consensus-based Image Description Evaluation)
- **BLEU** (Bilingual Evaluation Understudy)
- **Perplexity**: a common metric for assessing the fluency of language models.

Personality Consistency Evaluation
For personality-based models, we will employ the **Big Five**

Inventory (BFI) Test and **LIWC (Linguistic Inquiry and Word Count)** software to evaluate how well the generated dialogues reflect the intended personality traits.

6. Schedule

Write a brief schedule/timeline of your project.

Week 1: Data Preprocessing

Week 2: Construct Attributes and Relation Tables

Week 3: Create Dialogue Generation System

Week 4 - 5: Fine-Tuning the LLM Model

Week 6: Evaluation

References

[1] Christopher Clarke, Yuzhao Heng, Lingjia Tang, Jason Mars. *PEFT-U: Parameter-Efficient Fine-Tuning for User Personalization.* Computer Science & Engineering, University of Michigan Ann Arbor, MI, 2024.

[2] Jeiyoon Park, Chanjun Park, Heuiseok Lim. *Enhancing Dialogue Generation Through Iterative Prompt with Difference Analogy Learning.* 2024 9th International Conference on Computer and Communication Systems (ICCCS), 2024.

[3] Jing Fan, Dennis Aumiller, Michael Gertz. *Evaluating Factual Consistency of Texts with Semantic Role Labeling.* Institute of Computer Science, Heidelberg University, 2023.

[4] Lajos Matyas Csepregi. *The Effect of Context Aware LLM Based NPC Dialogues on Player Engagement in Role Playing Video Games.* Department of Architecture, Design and Media Technology, Aalborg University, 2023.

[5] Nuo Chen, Yan Wang, Haiyun Jiang, Deng Cai, Yuhan Li, Ziyang Chen, Longyue Wang, Jia Li. *Large Language Models Meet Harry Potter: A Bilingual Dataset for Aligning Dialogue Agents with Characters.* Tencent AI Lab, Hong Kong University of Science and Technology, 2023.

[6] Ruijun Chen, Jin Wang, Liang-Chih Yu, Xuejie Zhang. *Learning to Memorize Entailment and Discourse Relations for Persona-Consistent Dialogues.* School of Information Science and Engineering, Yunnan University, Department of Information Management, Yuan Ze University, Taiwan, 2023.