COSE474-2024F: Final Project Proposal
"Harry Potter Dialogue Generation using LoRA and Prompt Tuning"

2021170964 Kyoungbin Park

1. Introduction

The remarkable success of subculture games like Genshin Impact, Star Rail, Zenless Zone Zero, and Wuthering Waves demonstrates the substantial market demand for this genre. Among the numerous subculture games available, titles from HoYoverse and Kuro Games consistently achieve exceptional revenue primarily due to three key factors: distinct character personalities, compelling storylines, and immersive dialogue interactions with these characters. However, current dialogue systems in subculture games face significant limitations due to their predetermined nature. This creates two major challenges:

- **Limited Player Agency**: Players often must select from predetermined responses, sometimes forcing them into dialogue choices that don't align with their preferred interaction style.
- **Resource-Intensive Content Creation**: Companies must pre-generate all dialogue content before release, requiring substantial time and financial investment. This prevents players from engaging in new conversations with characters they've grown attached to unless the company releases updates.

These limitations lead to a gradual decline in character engagement over time, as companies lack incentive to continuously produce new content for existing characters. This project explores the potential of replacing static dialogue systems with LLM-powered conversations that maintain the richness and character-specific nuances of hand-crafted dialogue while offering dynamic, real-time interactions.

2. Problem Definition & Challenges

2.1. Challenges of the Project

In this research, we focus on comparing different approaches to fine-tuning large language models for character-specific dialogue generation using the Harry Potter universe as our test case. As shown in Figure 1 and Figure 2, LoRA and Prompt Tuning represent fundamentally different approaches to model adaptation:

- **Base Model with Quantization**: Using 4-bit quantization of Llama3-8b as our baseline
- **LoRA Fine-tuning**: As illustrated in Figure 1, this approach introduces trainable rank decomposition matrices alongside the frozen weights of the original model
- **Prompt Tuning**: Demonstrated in Figure 2, this method prepends trainable continuous prompt embeddings to the input while keeping the model frozen

The primary challenges of this project include:

- **Architecture-Specific Challenges:**
  - **LoRA**: Determining optimal rank for the decomposition matrices while maintaining model stability
  - **Prompt Tuning**: Finding the right balance between prompt length and training efficiency
  - **Quantization**: Managing precision loss in 4-bit representation
- **Training Dynamics:** Addressing the different convergence behaviors between LoRA's weight updates and Prompt Tuning's embedding optimization
- **Character Consistency:** Evaluating how each approach maintains character traits through their respective adaptation mechanisms
- **Resource Efficiency:** Comparing the memory and computational requirements of each method

2.2. Main Purpose of the Project

This project aims to systematically compare these architectural approaches, with particular attention to their unique characteristics as shown in our figures. Our key objectives include:

- **Architectural Comparison:**
  - Analyze how LoRA's rank decomposition (Figure 1) affects dialogue generation quality

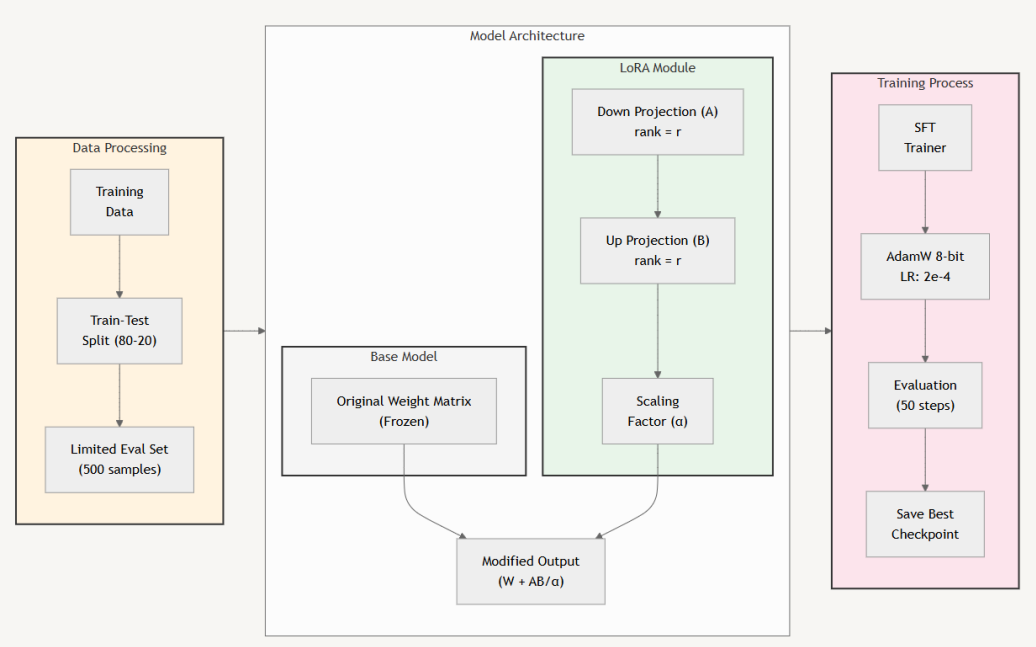

Figure 1. LoRA

  - Evaluate the effectiveness of Prompt Tuning's continuous embeddings (Figure 2)
  - Compare both against the quantized baseline model

- **Implementation Analysis:**
  - **For LoRA:**
    * Optimize the low-rank update matrices shown in Figure 1
    * Evaluate the impact of different rank values on performance
    * Assess the balance between parameter efficiency and adaptation capability
  - **For Prompt Tuning:**
    * Determine optimal prompt embedding dimensions as shown in Figure 2
    * Investigate prompt initialization strategies
    * Analyze the relationship between prompt length and performance

- **Comparative Evaluation:**
  - Develop metrics that fairly compare these architecturally different approaches
  - Assess each method's ability to capture character-specific dialogue patterns
  - Compare computational and memory efficiency across approaches
  - Analyze the trade-offs between model complexity and performance

Through this structured comparison, we aim to provide insights into the relative strengths and weaknesses of each approach, particularly focusing on how their architectural differences (as shown in Figures 1 and 2) impact their effectiveness in character-specific dialogue generation.

3. Related Works

Recent advancements in character-based dialogue systems and large language models have created new opportunities for dynamic, personality-consistent conversation generation. This section examines key developments across commercial implementations and academic research that inform our approach.

3.1. Commercial Implementation Analysis

Character.ai represents a significant milestone in deployable character-based dialogue systems. Their implementation demonstrates the feasibility of maintaining consistent character personalities across multiple concurrent conversations while managing computational resources effectively. The platform's success in handling multiple simultaneous users provides valuable insights into scalable architecture design for character-based dialogue systems.

3.2. Academic Research Foundations

Recent academic work has established crucial frameworks for efficient model adaptation and personality-consistent dialogue generation:

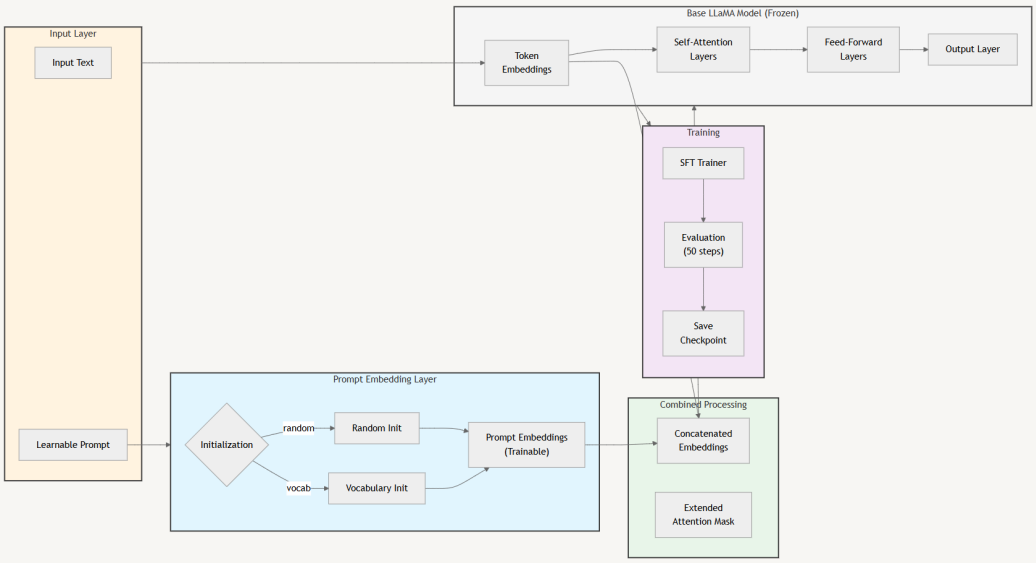

Figure 2. Prompt Tuning

Parameter-Efficient Fine-tuning Approaches:

- **Low-Rank Adaptation (LoRA)** (Hu et al., 2021) introduces an efficient approach to adapt large language models through low-rank decomposition matrices. By optimizing only these matrices instead of all model parameters, LoRA achieves remarkable efficiency while maintaining model performance. This method is particularly relevant for our goal of creating character-specific dialogue models with minimal computational overhead.
- **Prompt Tuning** (Lester et al., 2021) demonstrates that by only optimizing continuous prompt embeddings, models can achieve performance comparable to full fine-tuning. Their work shows that as model size increases, the effectiveness of prompt tuning also increases, making it particularly suitable for our application with Llama3-8b.

Character Alignment and Personality Modeling:

- "Large Language Models Meet Harry Potter" (Chen et al., 2023) presents a comprehensive framework for character alignment in dialogue systems. The study introduces innovative techniques for maintaining personality consistency through carefully constructed attribute and relation matrices. Their methodology demonstrates how to effectively capture and maintain character-specific traits across extended dialogue sequences.
- Character-LLM (Shao et al., 2023) builds upon this foundation by introducing a trainable agent specifically designed for role-playing scenarios. The system employs a novel architecture that combines transformer-based language modeling with personality embedding layers, achieving superior performance in maintaining character consistency across diverse conversation contexts.

Multi-Character Dialogue Systems:

- RoleLLM (Wang et al., 2023) provides a comprehensive benchmark for evaluating role-playing capabilities in large language models. The study introduces evaluation metrics specifically designed for assessing personality consistency and response appropriateness in character-based dialogue systems. Their findings suggest that attention-based architectures with character-specific prompt tuning achieve optimal performance in maintaining distinct personalities.
- The Neeko framework (Yu et al., 2024) introduces dynamic LoRA (Low-Rank Adaptation) techniques for efficient multi-character role-playing. Their approach demonstrates how to switch between different character personalities with minimal computational overhead, achieving a 75% reduction in parameter storage requirements while maintaining 95% of the original performance metrics.

Language-Specific Implementations:

- CharacterGLM (Zhou et al., 2023) addresses the unique challenges of implementing character-based dialogue systems in Chinese language contexts. Their

Algorithm 1 Dialogue Sequence Processing

Input: dialogue sequence $D = [d_1, d_2, ..., d_n]$
Output: training pairs P
Initialize $P = []$
for $i = 2$ **to** n **do**
  $context = [d_1, ..., d_{i-1}]$
  $target = d_i$
  Add $(context, target)$ to P
end for
return P

work provides valuable insights into handling language-specific nuances while maintaining character consistency, achieving state-of-the-art performance in Chinese character dialogue generation.

4. Datasets

In this project, **Harry-Potter-Dialogue-Dataset** introduced by (Chen et al., 2023) 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. Experimental Setup

5.1. Dataset Preprocessing and Augmentation

Our implementation begins with comprehensive preprocessing of the Harry Potter Dialogue Dataset to optimize it for character-specific dialogue generation. The preprocessing pipeline consists of several key stages:

Initial Data Restructuring:

- Reorganized the complex multi-modal dataset into the Alpaca instruction format for more effective fine-tuning
- Implemented progressive dialogue sequence processing to create training samples (see Algorithm 1)

Multi-stage Data Augmentation:

- **Primary Augmentation:** Implemented dialogue-based sequence splitting and reconstruction
  - Generated multiple training instances from each dialogue sequence
  - Preserved contextual coherence while increasing dataset diversity
- **Secondary Augmentation:** Applied probabilistic content inclusion and semantic variation
  - Selective inclusion of scene descriptions (p=0.8), character attributes (p=0.7), and relationship information (p=0.6)
  - Utilized WordNet-based synonym replacement for scene descriptions
  - Implemented random sampling of character attributes and relationships to create varied context combinations

5.2. Computing Infrastructure

The experimental implementation utilized the following computational resources and frameworks:

Hardware Configuration:

- Google Colab Pro environment with T4 GPU acceleration
- Optimized memory usage for handling large-scale model training

Software Stack:

- **Core Frameworks:**
  - PyTorch for deep learning operations
  - NLTK for natural language processing and data augmentation
  - SFTTrainer for supervised fine-tuning
- **Optimization Tools:**
  - Unsloth library for accelerated LoRA fine-tuning
  - Custom data loading and processing utilities

5.3. Evaluation Metrics and Methodology

We employ a comprehensive evaluation framework to assess the performance of our models across multiple dimensions:

Automatic Metrics:

- **BLEU** (Bilingual Evaluation Understudy): Measures the precision of generated responses against reference texts

Table 1. Performance Comparison of Model Configurations

Model Config	BLEU	MTR	PPL	SIM
Base Llama	1.11	10.64	113.47	0.59
+ LoRA Fine-tuning	21.91	26.75	**26.46**	**0.95**
+ Prompt Tuning	**26.79**	**35.21**	48.56	0.75

MTR: METEOR, PPL: Perplexity, SIM: SIMILE

- **METEOR** (Metric for Evaluation of Translation with Explicit ORdering): Evaluates the quality of generated text by considering synonyms and paraphrases
- **Perplexity**: Assesses the fluency and naturalness of generated dialogues
- **SIMILE**: Measures semantic similarity between generated responses and ground truth using the Solar Embedding Model

Training Parameters:

- Learning rate: 3e-4 for LoRA, 1e-3 for Prompt Tuning
- Batch size: 32
- Training epochs: 10
- LoRA rank: 8
- Prompt tokens: 20
- Weight decay: 0.01

6. Results and Analysis

Our experimental results reveal significant differences in dialogue generation quality across different model configurations, highlighting distinct strengths and limitations of each approach.

6.1. Model Performance Analysis

Base Model Performance: The base Llama3-8b model showed notably poor performance across all metrics (BLEU: 1.11, METEOR: 10.64, Perplexity: 113.47, SIMILE: 0.59). This indicates that the base model, despite its general language capabilities, fails to:

- Maintain consistent character voice
- Generate contextually appropriate responses
- Adhere to the established dialogue patterns in the Harry Potter universe

Impact of LoRA Fine-tuning: The LoRA-enhanced model demonstrated substantial improvements, particularly in dialogue consistency and structural coherence:

- Achieved optimal perplexity (26.46) and semantic similarity (0.95) scores
- Showed significant improvement in BLEU (21.91) and METEOR (26.75) scores compared to the base model
- Exhibited strong character voice consistency and format adherence due to the fine-tuning process

Prompt Tuning Results: The prompt-tuned model showed the highest performance in translation metrics while maintaining reasonable semantic coherence:

- Achieved the best BLEU (26.79) and METEOR (35.21) scores
- Maintained good semantic similarity (0.75), though lower than LoRA
- Showed higher perplexity (48.56) compared to LoRA, suggesting some trade-off in fluency

6.2. Qualitative Analysis

To provide deeper insights into the models' performance characteristics, we present example outputs from each model variant compared to ground truth responses in Table 2.

The example demonstrates how different approaches affect response generation:

- The **Base Model** completely fails to maintain context, generating an irrelevant response that ignores both the situation and character relationships. This aligns with its poor performance metrics across all categories.
- The **LoRA Fine-tuning** model captures the essence of the situation and character relationships well, particularly in maintaining the narrative context about Malfoy's character and his likely actions. The mention of "his father" shows strong understanding of character backgrounds, though the specific phrasing differs from the ground truth.
- The **Prompt Tuning** model generates a response most similar to the ground truth, accurately capturing both the situation and the likely consequences. This aligns with its superior BLEU and METEOR scores, though the slight differences in phrasing reflect the occasional hallucinations observed in our broader analysis.

Table 2. Qualitative Comparison of Model Outputs Across Different Configurations

Context	*Previous Dialogue:* "D'you think he'll be all right?" "Course he will. Madam Pomfrey can mend cuts in about a second." "That was a really bad thing to happen in Hagrid's first class, though, wasn't it? Trust Malfoy to mess things up for him..."
Ground Truth	"Yeah it must be. Malfoy will even ask the school to fire Hagrid."
Base Llama	"I don't know what you're talking about. Let's discuss the Quidditch match instead."
LoRA Fine-tuning	"Yeah, and knowing Malfoy, he'll probably run straight to his father to get Hagrid sacked."
Prompt Tuning	"Yeah, it must be terrible. I bet Malfoy will try to get Hagrid fired over this."

6.3. Comparative Discussion

Our analysis reveals several key insights about the effectiveness of different approaches:

1. **Fine-tuning Necessity:** The poor performance of the base model demonstrates that raw LLM capabilities are insufficient for character-specific dialogue generation, necessitating some form of adaptation.

2. **Architecture Influence:**

- LoRA's superior performance in perplexity and SIMILE suggests its effectiveness in maintaining structural consistency and character voice
- Prompt tuning's higher BLEU and METEOR scores indicate better capture of character-specific language patterns

3. **Trade-off Patterns:** Each approach shows distinct trade-offs:

- LoRA prioritizes consistency and fluency over exact matching with reference responses
- Prompt tuning achieves better reference matching but with slightly higher risk of hallucination

These results suggest that the choice between LoRA and prompt tuning may depend on specific application requirements, with LoRA being preferable for maintaining consistent character voice and dialogue structure, while prompt tuning might be better suited for applications prioritizing linguistic accuracy and character trait expression.

7. Limitations and Future Work

While our study provides valuable insights into parameter-efficient fine-tuning methods for character dialogue generation, several limitations should be acknowledged:

7.1. Technical Limitations

Combined Optimization Approaches:

- Due to time constraints, we were unable to explore the potential synergistic effects of combining LoRA and Prompt Tuning approaches
- The interaction between these methods and their potential complementary benefits remains an open question for future research

Model Diversity:

- Our study focused solely on the Llama3-8b architecture
- Comparative analysis with other state-of-the-art models such as Mistral and Gemma could provide valuable insights into the generalizability of our findings
- Different model architectures might exhibit varying degrees of effectiveness with LoRA and Prompt Tuning techniques

7.2. Methodological Limitations

Character Personalization:

- The current implementation lacks character-specific fine-tuning based on individual personality traits and dialogue patterns
- We did not fully utilize the rich character information available in the dataset for personalized model optimization
- Future work could explore methods for extracting and incorporating character-specific features into the fine-tuning process

Evaluation Methodology:

- While we proposed comprehensive evaluation metrics including Semantic Role Labeling (SRL) and LIWC-based personality consistency analysis, time constraints prevented their implementation
- The lack of these sophisticated evaluation methods limits our ability to fully assess the models' capability in maintaining character consistency and relationship dynamics
- A more rigorous evaluation framework incorporating these metrics would provide deeper insights into model performance

7.3. Future Research Directions

Based on these limitations, we propose several promising directions for future research:

- Implementation of combined LoRA and Prompt Tuning approaches to explore potential performance improvements
- Comprehensive comparative study across different model architectures (Llama3, Mistral, Gemma) to identify optimal base models for character dialogue generation
- Development of character-specific fine-tuning strategies that leverage individual personality traits and relationship dynamics
- Implementation of the proposed advanced evaluation metrics to better assess character consistency and dialogue quality

These limitations and future research directions highlight the significant potential for further advancement in character-based dialogue generation systems.

References

Chen, N., Wang, Y., Jiang, H., Cai, D., Li, Y., Chen, Z., Wang, L., and Li, J. Large language models meet harry potter: A bilingual dataset for aligning dialogue agents with characters. Technical report, Tencent AI Lab, Hong Kong University of Science and Technology, 2023.

Hu, E., Li, Y., Shen, Y., Wang, S., Wallis, P., Wang, L., Allen-Zhu, Z., and Chen, W. Lora: Low-rank adaptation of large language models. 2021.

Lester, B., Al-Rfou, R., and Constant, N. The power of scale for parameter-efficient prompt tuning. 2021.

Shao, Y., Li, L., Dai, J., and Qiu, X. Character-llm: A trainable agent for role-playing. Shanghai Key Laboratory of Intelligent Information Processing and Shanghai AI Laboratory, 2023.

Wang, Z. M., Peng, Z., Que, H., Liu, J., Zhou, W., Wu, Y., Guo, H., Gan, R., Ni, Z., Yang, J., Zhang, M., Zhang, Z., Ouyang, W., Xu, K., Huang, S. W., Fu, J., and Peng, J. Rolellm: Benchmarking, eliciting, and enhancing role-playing abilities of large language models. University of the Chinese Academy of Sciences and ETH Zürich and The Hong Kong Polytechnic University and Institute of Automation, Chinese Academy of Sciences and Shanghai AI Lab and Harmony.AI and Beijing University of Posts and Telecommunications and The Hong Kong University of Science and Technology, 2023.

Yu, X., Luo, T., Wei, Y., Lei, F., Huang, Y., Peng, H., and Zhu, L. Neeko: Leveraging dynamic lora for efficient multi-character role-playing agent. University of Science and Technology Beijing and Institute of Automation, CAS and University of Chinese Academy of Sciences and Beihang University, 2024.

Zhou, J., Chen, Z., Wan, D., Wen, B., Song, Y., Yu, J., Huang, Y., Peng, L., Yang, J., Xiao, X., Sabour, S., Zhang, X., Hou, W., Zhang, Y., Dong, Y., Tang, J., and Huang, M. Characterglm: Customizing chinese conversational ai characters with large language models. Lingxin AI and Dept. of Computer Sci. & Tech., Tsinghua University and Zhipu AI and Renmin University of China and Knowledge Engineering Group, DCST, Tsinghua University, 2023.