NPAIC: Knowledge-Grounded NPC Dialogue Generation with Personality and Memory

Gannon Smith
University of Michigan / Ann Arbor, USA
gansmith@umich.edu

1 Abstract

Non-player characters (NPCs) are central to immersion in modern role-playing games (RPGs), yet most deployed systems rely on rigid, pre-authored dialogue trees that limit reactivity and replayability. This project investigates whether a lightweight large language model (LLM), augmented with character-specific adaptation and explicit memory mechanisms, can generate dialogue that is coherent, task-aligned, and consistent with a well-defined character persona. Focusing on Arthur Morgan from *Red Dead Redemption 2*, I evaluate a base LLM, retrieval-based baselines, and a parameter-efficient fine-tuned model augmented with conversation memory and a knowledge graph. Quantitative evaluation using both traditional similarity metrics and LLM-based judges reveals that a strong base model outperforms the fine-tuned memory-augmented model on character consistency and coherence by a nontrivial margin. Analysis suggests that limited fine-tuning data and information loss introduced by memory summarization degrade performance relative to providing the base model with raw conversational context. These results highlight practical trade-offs in adapting LLMs for interactive narrative settings and suggest that careful memory design may be more impactful than parameter adaptation when data is limited.

Github Repository: github/gannonsmith/npaic

2 Introduction

Dialogue-driven interaction is a defining feature of role-playing games, shaping player immersion, narrative pacing, and emotional engagement. Despite advances in game AI, most NPC dialogue systems remain heavily scripted, relying on static dialogue trees that fail to adapt meaningfully to player behavior or evolving world state. As a result, repeated interactions often expose mechanical limitations that break immersion.

Recent advances in large language models (LLMs) suggest a potential alternative: generative dialogue systems that can dynamically respond to player input while maintaining character consistency and narrative grounding. However, naively applying LLMs to games raises several challenges. Generated dialogue must remain faithful to established characters, respect the current quest and world state, and operate efficiently enough for real-time use on consumer hardware.

This project explores the design and evaluation of a character-centric NPC dialogue generation system that integrated three core ideas: (1) a strong base LLM, (2) explicit memory mechanisms for conversation history and world knowledge, and (3) parameter-efficient fine-tuning to encode character personality. Using Arthur Morgan from *Red Dead Redemption 2* as a case study, the central research question is:

Can a lightweight LLM augmented with memory and character-specific adaptation generate dialogue that is more immersive and consistent than simple retrieval-based baselines?

The project ultimately yields a surprising result: a carefully prompted base model outperforms a LoRA-adapted model with memory augmentation. Rather than indicating failure, this finding provides insight into the sensitiviy of fine-tuning, the importance of memory representation, and the strength of modern base models. The main contributions of this work are:

- An end-to-end NPC dialogue generation pipeline combining conversational memory, a knowledge graph, and parameter-efficient adaptation.
- A controlled evaluation of base versus fine-tuned models under identical conditions.
- Empirical evidence that memory summarization can harm downstream dialogue quality

Character	Lines
Arthur Morgan	3279
John Marston	1387
Dutch van der Linde	1198
Sadie Adler	511
Charles Smith	492
Hosea Matthews	349

Figure 1: Red Dead Redemption 2: Line counts for characters with the most lines

when not carefully designed.

3 Data

The data for this project consists of fan-transcribed game scripts from *Red Dead Redemption 2* (RDR2). These scripts include dialogue lines annotated with speaker identity and mission context. While the original dataset contains dialogue from many characters, this project focuses exclusively on Arthur Morgan due to the substantial overhead involving constructing character-specific memory systems and the limited availability of time and computational resources.

Arthur Morgan is the most frequently speaking character in the dataset, with over 3,200 dialogue lines. Figure 1 shows the distribution of dialogue lines across major characters, motivating the decision to focus on Arthur as an initial test case.

3.1 Preprocessing

Raw script data was cleaned and normalized to remove transcription artifacts and formatting inconsistencies. Dialogue was segmented into structured input-output pairs, where each instance consists of:

- Mission identifier
- A short window of prior dialogue context
- Current speaker utterance (typically another character)
- Target NPC response (Arthur Morgan)

Dialogue pairs were split into training, validation, and test sets using an 80/10/10 split. The final training set contains 2,528 dialogue pairs. No additional annotation beyond existing mission metadata was introduced. An example of a processed input instance is shown in Appendix B.

4 Related Work

This project draws inspiraton from prior work in personalized text generation, knowledge-grounded dialogue systems, and memory-augmented agents.

4.1 Personalization in NLP

Salami et al. introduced LaMP, a benchmark for evaluating personalized generation using both automatic metrics and semantic similarity measures (Salemi et al., 2024). Their work motivates evaluating character fit rather than surface-level similarity alone.

4.2 Knowledge-grounded NPC dialogue generation

Ashby et al. integrated structured knowledge graphs with a language model approach to generate contextually correct dialogue for NPCs (Ashby et al., 2023). Their hybrid system ensures that the generated responses remain consistent with the quest statue and player context. This project adopts a similar idea but applies it to a single-character, dynamically retrieved knowledge graph.

4.3 Procedural content generation

Vaertinen et al. explored the generation of RPG quest descriptions with GPT language models, focusing on procedural quest generation instead of the dialogue (Värtinen et al., 2024). Their experiments showed how GPT models are able to generate coherent quest narratives through careful prompt structure and the abstraction of world entities. Their approach simplifies references to characters and locations to avoid factual errors from the model, and inserted the information back in post-processing. Their findings inform prompt structure decisions used in this work.

4.4 Generative agents and memory

Parl et al. introduced an incredibly interesting study of persistent LLM-drived characters in a Sims-like simulated environment (Park et al., 2023). Their agents used a layered memory architecture, combining long-term recollection, reflection, and planning to produce believeable, evolving behaviors in agents. This work adapts the idea of summarization-based memory to a dialogue-only NPC setting.

4.5 Parameter-efficient fine-tuning

Hu et al. proposed a technique for fine-tuning large models by inserting trainable low-rank matrices into the transformer layers (Hu et al., 2021). This method dramatically reduced the number of trainable parameters required while preserving the model quality. LoRA is central to this project's attempt to encode personality without retraining full model weights.

4.6 LLMs as evaluators

Recent surveys show that LLMs can serve as reliable judges of generation quality (Li et al., 2024). This motivates the use of an LLM-based evaluator for subjective metrics such as character consistency.

5 Methods

This section details the full system design, baseline constructions, and modeling choices used to evaluate NPC dialogue generation for Arthur Morgan. The overarching goal of the methodology is to isolate the effects of (1) semantic retrieval, (2) large pre-trained language models, and (3) explicit memory and personality adaption, while keeping the evaluation setting controlled and comparable across approaches.

5.1 Baselines

Four non-parametric baselines were implemented to contextualize the performance of the models. These baselines require minimal to no training beyond preprocessing and are designed to test progressively stronger assumptions about relevance and character grounding.

Random Line. This baseline samples a response uniformly at random from the full set of dialogue lines in the dataset, regardless of speaker, mission, or context. It serves as a sanity check for the evaluation pipeline and represents a minimal lower bound on coherence, relevance, and character consistency.

In-Character Random Line. This baseline restricts random sampling to dialogue lines spoken by Arthur Morgan. While still ignoring conversational context, this method captures superficial stylistic features such as vocabulary, tone, and dialect associated with the character. Improvements over the fully random baseline indicate the importance of speaker-specific language patterns even without contextual grounding.

Embedding-Based Retrieval. This baseline computes sentence embeddings for the current player utterance and retrieves the most semantically similar past dialogue instance from the training set. The associated Arthur Morgan response from that instance is returned verbatim. Embeddings were computed using a pre-trained sentence embedding model, and cosine similary was used for nearest-neighbor retrieval. This approach approximates a memory-based dialogue syste and provides a strong non-generative baseline that tests how far semantic similarity alone can go without language generation.

Actual Response. This baseline is used as validation, to ensure that the evaluation pipeline actually works as intended. It is also useful to contextualize how well the other methods perform.

Together, these baselines span a spectrum from context-free randomness, to context-aware retrieval, to the correct response, enabling a clearer interpretation of the results.

5.2 Base Model

All generative experiments are built on **Qwen2.5-3B**, a 3-billion-parameter causal language model selected for its strong instruction-following performance and feasibility for real-time inference. The model was used in its publicly released form without additional pretraining or fine-tuning.

At inference time, the base model was prompted with the verbatim last 10 dialogue turns from the conversation history, including speaker tags. This design choice avoids information loss and allows the model to directly attend over raw conversational context. No explicit conversation summaries or external knowledge were provided to the base model, allowing its performance to reflect the capabilites of the pretrained model alone.

All inference was conducted on a SPGPU on the Great Lakes Cluster, with typical response latencies under one second, demonstrating feasibility for interactive settings.

5.3 Character-Specific LoRA Adaptation

To encode Arthur Morgan's personality more explicitly, a parameter-efficient fine-tuning strategy was employed using Low-Rank Adaptation (LoRA) (Hu et al., 2021). A single LoRA adapter was trained exclusively on Arthur Morgan's dialogue data, consisting of 2,528 training instances.

Training was performed for three epochs with rank $r = 8$ and scaling factor $\alpha = 16$. LoRA adapters were inserted into the query, key, value, and output projection matrices of each transformer layer. This configuration was chosen to balance expressivity and stability while keeping the number of trainable parameters small.

The objective of LoRA training was not to teach the model conversational structure from scratch, but rather to bias generation toward Arthur Morgan's characteristic tone, brevity, and worldview. Importantly, the LoRA-adapted model was trained and evaluated only in conjunction with the memory mechanisms, reflecting the intended final deployment scenario rather than as a standalone generator.

5.4 Conversation Memory

Conversation memory was implemented as a summarization-based long-term memory mechanism. Instead of providing the model with the full dialogue history, the last ten dialogue turns were summarized into a concise 2-3 sentence representation at each step.

Summarization was performed by the base Qwen2.5-3B model using a fixed prompt that instructs the model to retain only salient conversational facts, intentions, and emotional context, shown in Appendix C. The resulting summary was injected into the generation prompt under a dedicated "Conversation Memory" section.

This design was motivated by efficiency concerns and by prior work suggesting that summarized memory can approximate long-term recall (Park et al., 2023). However, it also introduces the risk of information loss, which is later examined in the Discussion section.

5.5 Knowledge Graph Memory

In addition to conversational context, the system incorporates explicit world knowledge through a lightweight knowledge graph constructed using the NetworkX library. The graph represents Arthur Morgan's perspective of the game world and contains nodes corresponding to major characters, locations, and missions, connected by simple relational edges.

The final graph, shown in Figure 2, contains approximately 40 nodes. At inference time, relevant subgraphs are retrieved using rule-based matching over proper nouns appearing in the recent dialogue, the current mission identifier, and known location mentions. On average, about three nodes are retrieved per interaction.

Retrieved nodes are converted into textual facts and summarized into a short description using the base model. The summary prompt is shown in Appendix C. This summary is injected into the prompt under a "KG Memory" section. The LoRA-adapted model was both trained and evaluated with this knowledge memory enabled, ensuring consistency between training and inference conditions.

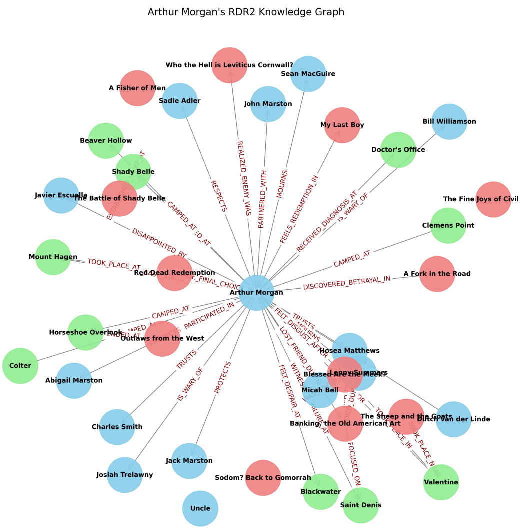

Figure 2: Arthur Morgan's knowledge graph, containing other main characters, locations, and missions

5.6 Prompt Structure

All generative models followed a structured prompt format designed to clearly separate instruction, memory, and generation. Prompts consisted of:

- A system-level instruction defining Arthur Morgan's persona and constraints
- An optional dialogue of the previous 10 lines
- An optional conversation memory summary
- An optional knowledge graph memory summary
- A response generation cue

This structured format was chosen to reduce prompt ambiguity and encourage the model to treat memory components as grounding information rather than as dialogue to be continued. Full prompt templates are provided in Appendix D.

5.7 Inference Pipeline

Figure 3 illustrates the complete inference time pipeline used by the final system. Inference proceeds as a deterministic, multi-stage process that integrates conversational context, world knowledge, and character-specific generation.

At each interaction step, the system first ingests the most recent player utterance along with a fixed window of prior dialogue turns. For the base model,

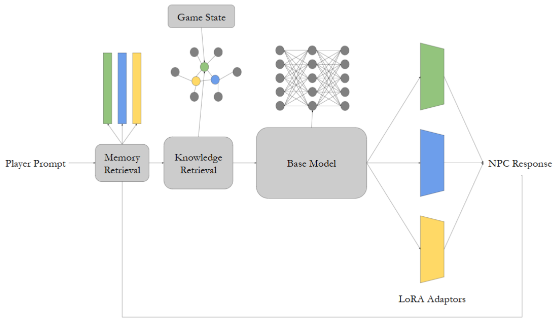

Figure 3: NPAIC inference architecture, showing the swapped weights per character using the same flow

this dialogue window is passed directly to the language model. For the LoRA-adapted system, the dialogue window is instead routed through the conversation memory module, where the last ten dialogue turns are summarized into a concise 2-3 sentence representation. This summary serves as a compressed representation of conversational state.

In parallel, the knowledge graph memory module performs entity-based retrieval over the current mission context, recent speakers, and explicitly mentioned locations or characters. Relevant nodes and edges are extracted from the knowledge graph and converted into short factual statements. These statements are then summarized into a compact textual form suitable for prompt injection.

Once memory retrieval and summarization are complete, the system assembles a structured prompt containing: the system instruction, the conversation memory summary (if enabled), the knowledge graph summary (if enabled), and the immediate dialogue context. This prompt is then passed to the generative model, either the base Qwen2.5-3B model or the LoRA-adapted variant, for response generation.

The model produces a single NPC dialogue response, which is subsequently post-processed to enforce formatting constraints before being returned to the user or evaluation pipeline. Importantly, the inference pipeline does not allow the model to directly modify game state or perform actions; it is strictly limited to dialogue generation. This separation reflects realistic deployment constraints in game systems and avoids entangling language generation with game logic.

This modular inference design allows individual components to be independently modified or ablated, and it directly motivates the architectural analysis discussed in Section 6.

6 Evaluation and Results

6.1 Metrics

Models were evaluated on the same held-out test set using the following metrics.

- Character Consistency (1-5): LLM-judged alignment with Arthur Morgan's persona.
- Relevance and Coherence (1-5): LLM-judged relevance to player input.
- Task Alignment (0-1): Whether the response acknowledges or advances the intended task.
- BLEU: N-gram overlap with reference responses.
- BERTScore (F1): Semantic similarity to references.

Evaluation prompts are provided in Appendix A.

6.2 Results

Table 1 summarizes performance across all models. The embedding retrieval baseline substantially outperforms random selection methods, confirming the value of semantic similarity. However, the most notable result is that the base Qwen2.5-3B model outperforms the LoRA + memory model on most metrics.

Table 1: Performance Comparison

Model	CC	RC	TA	BLEU	BERT
Random	1.93	1.62	0.15	0.003	0.845
In-Char Random	2.01	1.79	0.11	0.003	0.846
Embed-Sim	3.78	3.14	0.43	0.028	0.854
Base-Model	3.19	4.07	0.51	0.004	0.842
LoRA-Model	2.75	3.17	0.40	0.003	0.825
Actual-Response	4.90	4.82	0.84	0.849	1.000

Figure 4 visualizes these metrics across models in a more understandable chart.

7 Discussion

The results in Table 1 reveal several instructive patterns about the strengths and limitations of generative NPC dialogue systems when constrained by limited data and imperfect memory mechanisms. Rather than a simple ranking of models, these results show how different sources of bias (retrieval, pretrained knowledge, and fine-tuning) can interact with the task.

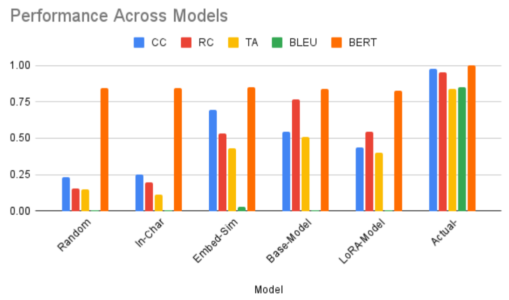

Figure 4: Model performance across baselines, scaled to (0,1)

First, the progression from the random baselines to embedding-based retrieval demonstrates the importance of semantic grounding. The embedding retrieval substantially outperforms both random baselines across all subjective metrics, acheiving a character consistency (CC) score of 3.78 and a relevance/coherence (RC) score of 3.14. This confirms that even without generation, access to semantically similar prior dialogue provides strong cues for producing plausible NPC responses. However, its lower task alignment score (0.43) suggests that retrieval alone struggles to adapt responses to the specific intent of the current interaction.

The base generative model exhibits a different and notable profile. While its CC score (3.19) is lower than that of the embedding retreival baseline, it achieves the highest relevance and coherence score among all non-ground-truth systems (4.07) and the highest task alignment score (0.51). This indicates that the pretrained model excels at interpreting player intent and producing contextually appropriate responses when given raw conversational history. The base model's relatively low BLEU score further reinforces that it is not reproducing training data verbatim, but instead generating novel responses that diverge lexically from the reference while remaining pragmatically aligned.

In contrast, the LoRA-adapted model underperforms the base model across all evaluation dimensions. Its drops in character consistency (2.75), relevance/coherence (3.17) and task alignment (0.40) suggest that the intended benefits of personality fine-tuning were not realized in practice. Several factors likely contribute to this outcome. First, the LoRA adapter was trained on a relatively small dataset, which may have caused over-specialization or interference with the base model's learned conversational priors. Second, unlike the base mode, the LoRA model relied on summarized conversation memory rather than raw dialogue history. Any loss or distortion in these summaries directly constrained the model's ability to generate coherent and situationally appropriate responses. Together, these factors likely outweighed the stylistic bias introduced by the fine-tuning.

The comparison between the embedding retrieval baseline and the base model further highlights a key trade-off. Retrieval-based methods preserve character voice effectively by reusing authentic dialogue, as reflected in higher character consistency scores, but lack the flexibility to adapt responses to new contexts. Generative models, on the other hand, demonstrate superior task alignment and coherence but are more prone to stylistic drift without string constraints. The LoRA results suggest that lightweight fine-tuning alone is insufficient to resolve this, particularly when memory representations may lose information.

Finally, the large gap between all automated systems and the actual in-game responses underscores the difficulty of the task. Human-authored dialogue achieves near-perfect scores across subjective metrics, emphasizing that current mdoels still fall short of the nuance, restraint, and narrative awareness exhibited by professional game writing. This gap is especially pronounced in character consistency, where even the best automated system remains more than a full point below the ground truth.

From an end-user perspective, the current system would not yet be suitable for deployment in a commerical game. Output formatting inconsistencies, occasional verbosity, and the tendency for the model to imply agency beyond dialogue generation would break all immersion. Nonetheless, the results provide clear guidance for future work: preserving informational conversational context and improving memory representations appear more critical than parameter-efficient fine-tuning under data-limited conditions.

8 Conclusion

This project explored the feasibility of memory-augmented, personality-driven NPC dialogue generation using a lightweight LLM. Through controlled experiments on Arthur Morgan from *Red Dead Redemption 2*, the study finds that a strong base model outperforms a fine-tuned alternative when memory summarization is imperfect and training data is limited. These findings suggest that future work should prioritize robust memory mechanisms and prompt optimization before resorting to parameter-efficient fine-tuning.

8.1 Qualitative Failure Analysis

To better understand the quantitative results, this section presents a qualitative analysis of common failures observed across models. Rather than isolated errors, these failures occurred systematically and help explain the metric trends reported in Section 6.

Character Drift and Over-Verbosity. The LoRA-adapted model frequently produced responses that were overly verbose or emotionally expressive relative to Arthur Morgan's established character. While the intent of fine-tuning was to reinforce character voice, the adapted model often responsed with moral reflection or dramatic phrasing that felt inconsistent with Arthur's typically restrained and pragmatic dialogue. In contrast, the embedding-based retrieval baseline preserved character voice more reliably by reusing authentic in-game dialogue, albeit at the cost of contextual relevance.

Memory-Induced Context Loss. A significant source of error stemmed from the summarization-based conversation memory. Summaries occasionally left out cues like conversational intent, urgency, or tone. When this occurred, the LoRA model generated responses that were locally coherent but globally misaligned with the conversation. The base model, which received raw dialogue context, was less prone to these failures, directly contributing to its scores.

Conflicts Between Knowledge and Dialogue Context. The knowledge graph memory occasionally introduced information that conflicted with the conversational flow. For example, retrieved mission or character facts sometimes prompted responses that acknowledged correct world knowledge but failed in terms of conversational flow. These conflicts highlight the challenge of integrating structured knowledge into generative dialogue without careful prioritization.

Retrieval-Generation Trade-off. Comparison between embedding retrieval and generative models reveals a fundamental trade-off. Retrieval-based methods excel at preserving stylistic authenticity but lack flexibility when player intent diverges from prior dialogue. Generative models adapt more effectively to new situations but require strong constraints to maintain character consistency. Table 2 illustrates representative examples of these failure modes.

Overall, this analysis suggests that many observed errors are architectural rather than purely modeling failures. Improvements in memory fidelity, conflict resolution between memory sources, and stricter stylistic constraints are likely to yield greater gains than additional fine-tuning alone.

9 Other Things I Tried

Significant effort was devoted to designing conversation memory summarization and knowledge graph retrieval. While functional, these components did not consistently improve performance and introduced more kinds of failures. Orginally, there was an attempt to create a dynamic world state knowledge graph to reflect knowledge changes, but this feature was dropped due to difficulty of evaluation and creation.

This project also originally aimed to incorporate multiple characters and additional games, but this scope was reduced to ensure a complete, evaluable system. The overhead required per character or per game, in terms of knowledge graphs, data preprocessing, and fine-tuning, was too substantial given the time constraints.

10 What I Would Do Next

Future work would begin with systematic ablation studies to better understand the contributions and interactions of the different memory components within the model architecture. By selectively disabling or modifying individual memory modules, I could isolate their impact on dialogue coherence, character consistency, and response diversity. This would provide clearer insight into which memory mechanisms are most effective, guiding targeted improvements.

Another important area would be the development and integration of more advanced summarization methods to condense and manage longtern conversational context. Current summarization techniques could be improved through the use of transformer-based models or reinforcement learning approaches that optimize for relevance and informativeness, enabling the model to retain key narrative elements without overwhelming the memory capacity.

Prompt engineering and optimization would also be a priority, as carefully designed prompts can significantly influence the model's behavior. Au-

Player Input	Base Model	LoRA + Memory	Ground Truth
“We should probably move before they notice us.”	“Yeah. Let’s keep it quiet and get out while we can.”	“I reckon we’ve been pushing our luck long enough. Every move we make carries weight, and sooner or later the past catches up to us.”	“Then let’s move. Quiet.”
“You really trust Dutch after all this?”	“I don’t know. I want to believe him, but I got my doubts.”	“Dutch is a complicated man, and loyalty ain’t something you give lightly. I’ve been thinking a lot about where all this is headed.”	“I want to. Don’t mean I do.”

Table 2: Representative failure cases comparing dialogue responses for identical player inputs across models. The base model produces concise, intent-aligned responses, while the LoRA-adapted model exhibits verbosity and character drift. Ground-truth responses highlight the restrained tone characteristic of Arthur Morgan.

tomated prompt tuning, possibly through gradient-based methods or evolutionary algorithms, could be applied to discover prompts to maximize across new metrics.

Expanding the underlying knowledge graph and training datasets represents another critical step. A richer, more diverse knowledge base would provide the model with a deeper understanding of character lore, relationships, and game world dynamics. Pairing this with large and more varied dialogue would improve generalization and robustness.

In parallel, thorough experimentation with different LoRA hyperparameters would be conducted to better characterize the trade-offs between model complexity, training stability, and inference efficiency. This would help identify optimal configurations tailored to different game contexts or computational constraints.

Finally, extending the framework to more characters and games is the true purpose of this project, generalizing the architecture to all types of characters with minimal training and overhead work. Extending to new genres and types of characters would validate the generalizability of the approach and potentially uncover new challenges and opportunities in character-driven AI dialogue generation.

11 Acknowledgments

I thank the course instructors and classmates for feedback and discussion throughout the project. All work presented here is my own.

References

Trevor Ashby, Braden K Webb, Gregory Knapp, Jackson Searle, and Nancy Fulda. 2023. Personalized quest and dialogue generation in role-playing games: A knowledge graph- and language model-based approach. In *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems*, CHI ’23, New York, NY, USA. Association for Computing Machinery.

Edward J. Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. 2021. Lora: Low-rank adaptation of large language models. *Preprint*, arXiv:2106.09685.

Haitao Li, Qian Dong, Junjie Chen, Huixue Su, Yujia Zhou, Qingyao Ai, Ziyi Ye, and Yiqun Liu. 2024. Llms-as-judges: A comprehensive survey on llm-based evaluation methods. *Preprint*, arXiv:2412.05579.

Joon Sung Park, Joseph C. O’Brien, Carrie J. Cai, Meredith Ringel Morris, Percy Liang, and Michael S. Bernstein. 2023. Generative agents: Interactive simulacra of human behavior. *Preprint*, arXiv:2304.03442.

Alireza Salemi, Sheshera Mysore, Michael Bendersky, and Hamed Zamani. 2024. Lamp: When large language models meet personalization. *Preprint*, arXiv:2304.11406.

Susanna Värtinen, Perttu Hämäläinen, and Christian Guckelsberger. 2024. Generating role-playing game quests with gpt language models. *IEEE Transactions on Games*, 16(1):127–139.

A Evaluation Prompts

A.1 Character Consistency Prompt

```
"Evaluate how consistent the following
NPC response is with the given character
profile and prior dialogue.
Output only a single integer from 1 to 5,
where:
1 = Completely inconsistent
2 = Mostly inconsistent
3 = Somewhat consistent
4 = Mostly consistent
5 = Fully consistent

Character Profile:
Name: Arthur Morgan
Traits: Loyal, Stoic, Pragmatic,
World-Weary, Morally Conflicted
Speech Style: Gruff, concise, frontier
slang, avoids emotional or flowery
language.
Beliefs: Violence only when necessary,
distrust of law, redemption through good
```

```
deeds.

Past Dialogue:
{past_dialogue}

Current Response:
<Arthur>: {predicted_response}

Now rate consistency (1–5) and output only
the number.”
```

A.2 Dialogue Relevance and Coherence Prompt

```
”Evaluate whether the NPC's response is
coherent and relevant to the player's
input.

Output only a single integer from 1 to 5,
where:
1 = Not related
2 = Mostly not related
3 = Partially related
4 = Mostly coherent and relevant
5 = Fully coherent and relevant

Last message:
<{last_speaker}>: {last_line}

Current Response:
<Arthur>:  {predicted_response}

Now rate and output only the number.”
```

A.3 Task Alignment Prompt

```
”You are evaluating whether Arthur
Morgan’s response directs the player
toward the correct task.

Inputs:
- Speaker: {last_speaker}
- Arthur’s Response: {predicted_response}
- Intended Action Summary:
{correct_response_action}

Task:
Score from 0–1 whether Arthur’s response
clearly and logically aligns with the
intended actionor motivates it.
(1 = strongly aligned, 0 = not aligned)”
```

B Input Example

```
{
   "mission": "Old Friends",
  "context": "<Arthur Morgan> Good. </Arthur Morgan>
    <Leopold Strauss> Yes. I suppose. </Leopold Strauss>
    <Arthur Morgan> Hey, Jack. </Arthur Morgan>
    <Jack Marston> Morning. </Jack Marston>
    <Arthur Morgan> How are you holding up? </Arthur Morgan>
    <Jack Marston> Okay. </Jack Marston>
    <Arthur Morgan> That's the spirit. </Arthur Morgan>
    <Jack Marston> Okay. </Jack Marston>
    <Arthur Morgan> Hey, how's he doing? </Arthur Morgan>
    <Abigail Marston> He's okay, just needs some rest. Thank you again, Arthur.
      </Abigail Marston>",
   "speaker": "Abigail Marston",
  "utterance": "He's okay, just needs some rest. Thank you again, Arthur.",
   "response_speaker": "Arthur Morgan",
  "response": "That's alright. Keep him warm. Marston.",
   "response_action": "Keep him warm"
}
```

C Summarization Prompts

C.1 Conversation Memory

```
  "Summarize this dialogue in 2-3 sentences, focusing on the emotional tone, character
   motivations, and key facts:


   {context}


   Summary:"
```

C.2 Knowledge Graph Memory

```
  "Summarize the following knowledge graph facts into compact form that is useful for
   guiding the character's next line.


   {facts}

   Summary:"
```

D Inference Prompts

D.1 Base Model

```
"### Context:
{context}

### {speaker} says:
{utterance}

### Respond as {response_speaker} to the last utterance only.

{response_speaker}:"
```

D.2 LoRA Adapted Model

```
"You are roleplaying as Arthur Morgan from Red Dead Redemption 2.
Stay in character and respond naturally.

Mission:
{mission}

Conversation Memory:
{memory_summary}

Relevant Knowledge:
{knowledge_summary}

Dialogue:
{speaker}: {utterance}

{response_speaker}:"
```