Towards Immersive Computational Storytelling: Card-Framework for Enhanced Persona-Driven Dialogues

BINGLI LIAO $^1$ and Danilo Vasconcellos Vargas $^2$

$^1$Kyushu University
$^2$Affiliation not available

Abstract

In the realm of RPGs, creating immersive, persona-driven dialogues remains a challenge, especially in intricate settings like Call of Cthulhu (CoC). Existing methodologies often falter in portraying character personas within complex conversations accurately. To address this, we introduce a novel card-based framework, utilizing the advanced Baichuan-7B language model for tailored dialogue generation. Guided by detailed scene settings and character personas, Baichuan-7B exhibited a striking ability to craft context-aware dialogues for even unseen characters and scenarios. To assess the quality of these dialogues, we present an innovative metric, circumventing the traditional hurdles of human evaluations. Furthermore, insights into the attention mechanism shed light on the dynamics of information flow during dialogue creation. Collectively, our findings underscore the transformative potential of large language models in computational storytelling, particularly in RPG settings.
Towards Immersive Computational Storytelling: Card-Framework for Enhanced Persona-Driven Dialogues

Liao Bingli, Danilo Vasconcellos Vargas

Abstract—In the realm of RPGs, creating immersive, persona-driven dialogues remains a challenge, especially in intricate settings like Call of Cthulhu (CoC). Existing methodologies often falter in portraying character personas within complex conversations accurately. To address this, we introduce a novel card-based framework, utilizing the advanced Baichuan-7B language model for tailored dialogue generation. Guided by detailed scene settings and character personas, Baichuan-7B exhibited a striking ability to craft context-aware dialogues for even unseen characters and scenarios. To assess the quality of these dialogues, we present an innovative metric, circumventing the traditional hurdles of human evaluations. Furthermore, insights into the attention mechanism shed light on the dynamics of information flow during dialogue creation. Collectively, our findings underscore the transformative potential of large language models in computational storytelling, particularly in RPG settings.

Index Terms—generative models, games, procedural content generation, attention, visual analytics

I. INTRODUCTION

PERSONA-DRIVEN dialogue in role-playing games (RPGs) presents a significant potential to enhance player immersion. The creation of a unique speaking style for each character, aligned with their personality and profession, allows for a richer and more engaging player experience. However, crafting such dialogue is a demanding task for game developers, requiring not only creative writing skill but also a deep understanding of each character’s persona and the game scene’s setting.

The increasing demand for more dynamic and richly detailed game content highlights the need for scalable, automated methods of dialogue generation. Current methods of generating persona-driven dialogue, Zheng et al. [1] proposed a persona-sparse embedding approach which can only encode simple fixed objectives like location and hobbit tags which only have limited influence to shown persona within conversation. However, Chen et al. [2] proposed a learnable latent variable method to memorize entailment and discourse relations to generate more coherent and consistent dialogue. But it requires natural language inference data to pre-train their latent variable structure, for instance the DNLI, the inference dataset based on PersonaChat [3], it is not practical in common cases. These proposed persona-driven dialogue generation approaches [4], [5] are hard to adequately capture the nuances of a character’s persona in complex conversation context and can result in repetitive and inauthentic exchanges.

In contrast, recent advancements in AI generated content technology, particularly in the field of language models, offer promising opportunities for computational storytelling.

Generative Pre-trained Transformers (GPT) have demonstrated their capacity to generate human-like text across various genres and styles. Building on this potential, our study fine-tunes Baichuan-7B [6], one of the most advanced large language models which is memory-friendly and computationally efficient, to generate dynamic persona-driven dialogue within the context of the popular tabletop RPG, Call of Cthulhu (CoC).

CoC was chosen as the basis for our research due to its rich narrative environment, complex character personas, and the diverse interaction scenarios. This game, known for its emphasis on character development and story immersion, serves as an ideal ground for exploring the potential of persona-driven dialogue generation based on large language models.

An ideal dialogue generation system should be adaptable to a variety of contexts, including diverse character personas and conversation scenarios. To address this, we introduced a method that applies scene and character cards to construct dialogues. Each card details specific aspects of the scene settings and the personas of characters. Our results demonstrate that the fine-tuned language model can generate persona-consistent and contextually aware dialogue, even in the presence of new characters and scenarios not included in the training dataset.

Furthermore, our study extended beyond merely fine-tuning a language model to generate persona-driven dialogues. We explored deeper into attention mechanism, explaining the intricacies. Our findings reveal the mechanism for how the information is compressed and flows among prompts and dialogue segments during persona-driven dialogue generation. These insights suggest potential avenues for the development of innovative neural network architectures and attention mechanisms in future research.

Our research presents the following key contributions:

1) Introduction of a novel card-based approach for generating persona-driven dialogues using the table role-playing game, Call of Cthulhu (CoC). This framework enables dynamic conversation content adaptation by utilizing diverse character and scene cards.

2) Development of a metric to evaluate the quality of persona-driven dialogues that can be automatically processed using GPT-4, circumventing traditional hurdles of human evaluations like bias and expense.
3) An extensive study showing that the dialogue generation ability of larger models, like GPT-3.5, can be effectively transferred to smaller models, such as Baichuan-7B, through prompt engineering and organized data sampling.

4) We proposed a word-wise visual analytical approach highlighting the mechanisms through how the information is compressed and propagated during dialogue generation by large language models (LLMs), exemplified by the role of anchor words as pivotal nodes in information flow.

We have published our CoC persona-driven LLM along with the designed prompts on GitHub. Our aim is to pave the way for next-generation dialogue systems in the gaming industry, enhancing narrative depth and player engagement through procedurally generated, persona-driven dialogue.

II. RELATED WORKS

Dialogue Systems and Personalization: With the development of large-scale pre-trained language models, there has been significant interest in improving persona-based dialogue generation [7]–[9]. Traditional approaches to dialogue personalization were often rooted in psychology, such as parameterizing personality traits with the big five personality traits model [10]. However, these approaches were limited by the subjective nature of psychological metrics and the difficulty in collecting format-specific data. Hence, recent efforts have adopted a data-driven approach, leveraging large-scale dialogue datasets to learn persona features directly [11]. For instance, sequence-to-sequence learning has been used to model speaker personas from social media data [12], [13], while others have proposed models that input personas explicitly [14], [15]. However, instead of user private concern of the social media data used for model training, these models were typically small size and trained from scratch which limited their abilities on trivial persona traits and simple conversation generation. The advantages of pre-training language models with persona-driven dialogue generation have yet to be fully explored.

Consistency in Dialogue System: Despite the advancements in persona-based dialogue, the problem of dialogue consistency remains a challenge. Some studies have attempted to model mutual persona perception and to improve dialogue consistency [16], while others have decomposed persona-based dialogue tasks into consistent understanding and dialogue generation, with natural language inference playing a critical role [4]. These methods focus on maintaining coherence in dialogues but may struggle to generate persona-driven dialogue with persona style speaking style which is hard to be trained with natural language inference method. Moreover, these models lack the ability to conditionally generate dialogue which is desired in game development.

Model Interpretability: The success of Transformer models has encouraged a myriad of research investigations focused on understanding the mechanisms behind the outstanding performance. One approach aims at evaluating the model’s performance when presented with carefully crafted input sentences [17], [18]. However, this approach falls short of explain the inherent processes that enable the model to develop impressive abilities. An alternative exploration pathway involves probing the model’s vector representations, where a simple neural network is trained for tasks like part-of-speech tagging, as presented by Adi et al. [19]. Attention mechanisms as central to Transformer architecture have also been the subject of extensive analysis. Recent studies attempt to interpret model abilities by examining attention head dependencies across multiple layers [20]. Clark et al. [21] showed that the attention heads of pre-trained BERT models capture syntactic relations either by focusing on delimiter tokens or more generally across sentences. Numerous research also take effort to employ visualization techniques for better understanding the inner workings of Transformer models. Common strategies include creating bipartite graphs or heatmaps to delineate attention patterns for specific input sentences [22], [23]. A recent work by Yeh et al. [24] introduced a visualization approach based on the joint embedding of query and key vectors, aiming to support understanding the self-attention mechanism. Nevertheless, these methods primarily emphasize semantic and syntactic analysis grounded in pre-trained or traditional language tasks fine-tuned models. While these approaches provide convincing explanations for simple input sentences, an in-depth understanding of the attention mechanisms behind complicated multi-turn conversations still remain mysterious. Our research seeks to bridge this gap, utilizing word-wise attention heatmaps from the fine-tuned LLM to uncover how the model adopt specific attention patterns, enabling efficient information compression and propagation across dialogue turns in persona-driven conversations.

III. LARGE LANGUAGE MODELS

A. Generative Pre-trained Transformers (GPTs)

Generative Pre-trained Transformers (GPTs) are built upon the Transformer neural network architecture introduced by Google in 2017 [25]. While the original Transformer model was conceived for language translation tasks and consisted of both encoder and decoder blocks, GPT models exclusively utilize the decoder part. These models tokenize input text into vectors that are then processed and then processed through a series of blocks. Central to these blocks is the self-attention mechanism, which determines the relevance of each token word in relation to others within a given context. Consequently, this produces a weighted representation of the input vectors. In an autoregressive manner, the model’s objective is to predict subsequent tokens in a given sequence based on its preceding predictions.

Models building upon the foundational Transformer architecture have demonstrated astonishing versatility, efficiently handling a series of language tasks, from token classification and summarization to question answering. Moreover, their capabilities extends beyond linguistics, making impactful influences in audio processing and computer vision.

Recently, GPTs, aimed to craft human-like text, have garnered substantial attention. During training, the straightforward objective is to optimize the model’s parameters to ensure that each predicted token aligns with the prior ones. This learning process can be statistically represented by the equation:
giant LLMs, those with parameters exceeding 100B. It can be research reveal that the CoT is not the sole preserve of critical query can guide the model towards multi-step problem- appending the directive "think step by step" to a mathematical output with carefully constructed prompts. As an exemplar, with the CoC aesthetic.

facilitating the creation of character and scene cards consistent in response generation. Such a method is essential in our study, this technique, allowing more cogent inference and precision. The chain of thought (CoT) mechanism [29] further augments character's personality and individual linguistic expressions.

The important object represents a key item transpires. To maintain the CoC thematic essence, we incorporated elements such as an important object, character goals, and obstacles. The important object represents a key item relevant to the scenario, offering potential cues. Concurrently, character goals and obstacles outline the challenges to be navigated. We anticipate that these elements collectively provide the LLM with a rich contextual background information and it can establish backbone of the entire conversation orchestrating the flow of dialogue generation in alignment with the scene's thematic setting.

The aforementioned robust framework facilitates the generation of persona-consistent and contextually aware dialogue by integrating both character and scene cards. This innovative approach is expected to enhance the realism and depth of the generated dialogues, leading to more engaging and immersive storytelling experiences within the genre of RPGs.

\[
t_i = p(t_i|t_1...t_{i-1}; \theta) \tag{1}
\]

wherein \(t_i\) denotes the i-th token in the sentence and \(\theta\) signifies the model's parameter for the sampling distribution.

OpenAI’s GPT-3, building on the foundational Transformer architecture, marks a significant advancement in the GPT series [26]. With a staggering 175 billion parameters, it possesses over a hundred times the capacity of GPT-2 [27], demonstrating an exceptional aptitude for producing diverse textual outputs. Its evolution, termed GPT-3.5 or ChatGPT, amplifies this ability. It integrates methods like Reinforcement Learning from Human Feedback (RLHF) and introduces refinements in dataset quality [28]. However, training such large models is impractical on consumer-grade hardware that game studios and individual enthusiasts typically employ.

### B. Baichuan-7B

Our study harnesses the capabilities of Baichuan-7B, a model from the LLM family that, despite its compactness with only 7 billion parameters, consistently outperforms models of similar scales in benchmark test. Baichuan-7B excels in computational efficiency and boasts a memory-conservative design, implemented with optimizations like Flash-attention and R-M-SNorm [6]. These technical advancements position Baichuan-7B as an ideal candidate for online tasks, including real-time dialogue generation with consumer-grade hardware. In our research, we fine-tuned Baichuan-7B to produce persona-centric dialogues with the CoC aesthetic, utilizing character and scene cards. Our approach leveraged the model’s ability in generating contextually consistent and persona coherent dialogues.

### C. Prompt engineering and the Chain of Thought

Prompt engineering is a vital concept of the research on LLMs. It involves strategic designing tailored prompts to guide LLM responses in specific directions. In essence, a well-designed prompt reshapes the context for model generated outputs. The underlying principle of prompt engineering pivots on guiding the model through prompts with distinct structures, styles or contextual cues, ensuign the output resonates with the designer’s intent. For instance, if the aim is to generate a dialogue reflecting the speech style of a specific character, the prompt might encapsulate descriptions of the character’s personality and individual linguistic expressions. The chain of thought (CoT) mechanism [29] further augments this technique, allowing more cogent inference and precision in response generation. Such a method is essential in our study, facilitating the creation of character and scene cards consistent with the CoC aesthetic.

The CoT mechanism indicates that the LLM can align its output with carefully constructed prompts. As an exemplar, appending the directive “think step by step” to a mathematical query can guide the model towards multi-step problem-solving, significantly boosting the solution accuracy. Recent research reveal that the CoT is not the sole preserve of giant LLMs, those with parameters exceeding 100B. It can be transferred to smaller models via careful data sampling [30]. Inspired by this insight, we developed a context controllable, persona-driven dialogue generation mechanism which integrates character and scene cards into the prompting process. These cards serve as compact description of characters and their environments, encapsulating crucial dialogue-generating data, ranging from intrinsic character traits to their speaking nuances and ambient scene particulars. With these cards as prompts when fine-tuning Baichuan-7B model, we enabled it to generate dialogue presenting contextual awareness and persona consistency. Addressing the challenge of acquiring high-quality character and scene cards, we utilized GPT-3.5 with prompt engineering. It allows GPT-3.5 to craft diverse cards which seamlessly mirroring the distinct narrative style of CoC.

### IV. Character and Scene Cards

In our research, we encountered the challenged of a lack of publicly available datasets conforming to the specific format of CoC. Consequently, we employed prompt engineering to autonomously generate high-quality character and scene cards, working as a foundation for fine-tuning a persona-driven dialogue generation LLM.

**Character Cards:** Recognizing the insufficiency of standard persona information such as location, habits and gender, for our targeted objectives, we designed character cards to capture more nuanced traits. These cards encompass common attributes such as name, age and gender, while extending to more complex persona representations by incorporating fields that describe specific speaking style. Several exemplary dialogue lines are also included in this field. The detailed speaking style would assist the LLM in generating dialogue authentically representing the character’s unique persona. To better align with the CoC style, we integrated character skills and occupation into the character card intended to collaborate with the accompanying scene card.

**Scene Cards:** With the objective of precisely controlling the global conversation flow, we devised scene cards to serve as a blueprint for dialogue generation. Each scene card includes a field outlining the basic environment in which the dialogue transpires. To maintain the CoC thematic essence, we incorporated elements such as an important object, character goals, and obstacles. The important object represents a key item relevant to the scenario, offering potential cues. Concurrently, character goals and obstacles outline the challenges to be navigated. We anticipate that these elements collectively provide the LLM with a rich contextual background information and it can establish backbone of the entire conversation orchestrating the flow of dialogue generation in alignment with the scene’s thematic setting.

The aforementioned robust framework facilitates the generation of persona-consistent and contextually aware dialogue by integrating both character and scene cards. This innovative approach is expected to enhance the realism and depth of the generated dialogues, leading to more engaging and immersive storytelling experiences within the genre of RPGs.
V. Card Creation

Our prior analysis elucidated the potential benefits of incorporating character and scene cards in the prompts utilized by the LLM for persona-driven dialogue generation. A character card can be repurposed across various scenes, augmenting the flexibility and diversity of narrative contexts. We carefully crafted eight character cards, each containing a unique speaking style, to exhibit the fine-tuned LLM’s ability in generating distinct character-specific dialogues. Contrary to traditional persona-driven dialogue generation approaches, which often confine persona expression to sparse and elementary traits (such as habits or occupation), the incorporation of speaking style creates a more pervasive, natural persona-related behavior in the generated dialogue.

However, the manual creation of a sufficient number of high-quality scene cards to control the conversation flow when fine-tuning the LLM posed a significant challenge. Fortunately, prompt engineering emerged as a viable solution to this challenge. Drawing inspiration from the approach employed by Stanford’s alpaca-instruct LLM [31], we devised several instruction rules for GPT-3.5 to generate a variety of CoC-style scene cards. These rules delineated the specifications for the generated scene cards, including the formatting for each field and the essential alignment with CoC-style. Upon conducting initial tests, we discovered some instability in the quality of scene cards generated with the current instructions. For instance, the format of scene card fields varied across several instances of GPT3.5 API calls, failing to satisfy the requirement for consistent, high-quality, and well-formatted scene cards.

```javascript
scene card:
{
  "scenario_name": "<scenario_name>",
  "situation_description": "<situation_description>",
  "location": "<location>",
  "important_object": "<important_object>",
  "character_goal": "<character_goal>",
  "obstacle": "<obstacle>
}
```

Fig. 1. The format template of the scene card.

Inspired by the CoT capability of LLMs, we hypothesized that providing a format template alongside a sample scene card would assist the LLM in producing cards more in alignment with our target scene card requirements. We anticipated that this method would stabilize the quality of the generated scene cards. To evaluate this, we modified our previous instruction prompt to incorporate a format template and a sample. The format template serves as the blueprint for the scene card construction, using placeholders to denote the fields to be filled by GPT-3.5. Moreover, to optimize cost-efficiency and avoid generating only one scene card per GPT-3.5 API call, we revised our instruction to facilitate the generation of eight variant scene cards at once, given the maximum response token length of 2,048 for GPT-3.5.

Employing our innovative prompt and the GPT-3.5 API, we successfully created 6,028 unique scene cards, each featuring different combinations of situation, object, goal, and obstacle. A scene name field, serving as a high-level abstraction of the scene setting, also exists for each scene card. To improve diversity, we utilized the LLM to further generate approximately seven variant scene cards based on the same scene name. These cards feature different objects, goals, or obstacle while sharing the same scene settings.

VI. Training Data Generation

Having prepared character and scene cards, the subsequent step focused on generating sufficient persona-driven dialogue data, utilizing these cards for the supervised fine-tuning of the LLM. Our objective was to employ GPT-3.5, guided by the designed prompt, which includes instructions describing how to use the information in character and scene cards while creating persona-driven dialogue. After extensive testing based on various settings for dialogue generation, we determined that dialogues generated between two characters yielded optimal results in portraying the personas detailed in each character cards. Therefore, each generated dialogue sample is constructed using two character cards and one scene card within our research paradigm.

In order to further enhance dialogue generation, we discovered that the integration of a post-description, explicitly annotating the merit of the prompt instructions, could significantly improve dialogue quality. Traditional prompts present a direct set of instructions, conducting the LLM to deduce and generate desired outputs for designers. However, this format might be insufficient for complex tasks that require sophisticated behaviors in the LLM output. The persona-driven dialogue generation in this research entails extraordinary knowledge to achieve high-quality output. Similar to the CoT, detailed prompts seemingly guide LLMs to generate more refined outputs. Recent studies validate this, indicating that intricate instructions are more effective than basic ones during LLM fine-tuning [32]. Addressing this problem, we introduce Augmented Prompting with Post-Descriptions (AP-PD). This technique supplements the primary prompt with a post-description that rationally emphasizes the desired features expected in the generated dialogue. Through this approach, we observed a marked improvement in dialogue quality.

![Kernel Density Estimation Analysis](image)

Fig. 2. The kernel density estimation analysis of the score distributions from the original prompt and the AP-PD augmented advanced prompt.

To quantify the effectiveness of AP-PD, we employed GPT-3.5 to generate dialogues using both the conventional
prompt and AP-PD, based on identical character and scene card combinations. To eliminate potential human biases in evaluation, we utilized GPT-4 to assess the quality of dialogues based on various criteria: alignment with character and scene cards, character interaction, dialogue flow, engagement, and creativity. This objective approach ensures that the observed enhancements in dialogue quality could be attributed to the AP-PD technique rather than to subjective human interpretations. From a pool of 300 samples in Figure 2, AP-PD (with a mean score of 8.8 and standard deviation of 0.38) outperformed the traditional prompt (mean score of 8.5 and standard deviation of 0.53). This underscores the superior quality of persona-driven dialogues generated using AP-PD.

Recent studies on supervised fine-tuning of LLMs have shown that dataset quality is more important than quantity [33]. Typically in narrow domains, a limited amount of high-quality data could achieve impressive results for fine-tuned LLMs. Moreover, redundant low-quality data might detrimentally affect fine-tuned LLM performance, and the increased training steps also amplify energy consumption. Considering eight character cards and 6,000 scene cards, the total number of dialogues would exceed 160,000 if dialogues were generated for every possible character combination. Implementing LLM fine-tuning based on this data might induce overfitting on eight character cards and impair generalization capabilities. To balance the relation between quantity and diversity, we synthesized 49,454 dialogues instead of an exhaustive 160,000, maintaining a balance across scene types without generating every permutation for the remaining 5,000 scene cards, a balanced sampling approach that precisely align with both the scene setting and sophisticated character personas. The system prompt we devised for the fine-tuning phase is built upon the following foundational rules:

**Directive for Dialogue Creation:** Beginning with the instruction “Generate a dialogue based on the information provided in the Scene Card and Character Cards”, the primary context source for LLM generation is anchored to the provided cards. This directive ensures dialogues remain contextual, preventing dialogue context from deviating into irrelevant domains that could impair the immersion experience in the conversation.

**Stylistic Framework:** By specifying “You are partaking in a role-play dialogue set in the style of Call of Cthulhu”, the LLM is subtly informed of the desired genre and tone to adhere to, ensuring the produced dialogues reflect the atmospheric tension and thematic elements of CoC.

**Scene and Character Contextualization:** This section underscores the important role of scene and character cards in shaping the dialogue’s context. It ensures that the LLM is attentive to the distinct nuances embedded within the cards during dialogue generation.

**Character Adherence:** A precise instruction, “Your dialogue should adhere to the character’s speaking style, personality traits and quirks”, guides the model to generate responses that are aligned with each character’s mannerism, averting generic or out-of-character context.

**Persona-Driven Emphasis:** The concluding reminder, “the dialogue should be driven by the character’s persona”, reinforces the core premise in our research. It seeks to stimulate dialogues where character personas do not merely exist behind the scenes but actively drive the narrative.

This system prompt is not merely an instructional passage; it is a strictly designed directive capturing the essence of our research goals. Through its coherent guidance, Baichuan-7B is conditioned to generate dialogues that not only maintain consistency with the context set by the scene card but also resonate with the varied personalities in the character cards. The prompt thus emerges as a keystone in achieving the objective of immersive, persona-driven dialogues, showcasing the richness and depth of CoC’s narrative universe.

**VIII. Untuned LLM’s Performance**

Baichuan-7B has achieved noteworthy accomplishments on mainstream benchmarks, such as MMLU and AGIEval. It is imperative to establish a robust baseline by assessing its performance using our distinct system prompt, scene and character cards. The resulting dialogues synthesized by Baichuan-7B can be viewed with Figure 8 in Appendix. Our evaluation framework comprises the following metrics:

![Diagram illustrating the full prompt format used to generate dialogues, consisting of system prompt, scene card, character cards, and AP-PD components.](image)
A. Contextual Adherence

The dialogue exhibits the presence of mysterious phenomena and a potential danger, aligning with the eerie ambiance specified in the CoC’s scene card. However, the generated response only loosely adheres to the provided scene card in the prompt. Although there is a mention of “deep one”, salient elements such as the “Drowned Relic”, the expedition’s objective of “Recover lost expedition artifacts”, and the forthcoming challenges like the “Deep One ambush” are not directly addressed in the dialogue.

B. Character Fidelity

Professor Arthur: The dialogue showcases the professor’s inquisitive traits with queries such as “What did happen exactly? How does such bizarre phenomena occur without any apparent cause?”. Nevertheless, it largely overlooks his penchant for frequent use of Latin phrases and meticulous diction.

Dr. Eleanor: The LLM captures Dr. Eleanor’s quirky manner but overlooks her idiosyncratic nature and overwhelming enthusiasm for technical matters. The absence of her inventor identity and her tendency of coin new terms further diminishes her portrayal.

C. Narrative Flow and Consistency

The dialogue occasionally seems disjointed, with some transitions are jarring. Distinguishing a coherent narrative trajectory can be challenging. Moreover, certain sentences, e.g. “So yeah, whatever happened before has left quite an impression on everybody. Everyone except myself, perhaps. My curiosity never ceases to amaze me sometimes” are redundant and do not contribute to character development or plot progression.

D. Prompt Conformity

The generated dialogue deviates from the detailed prompt, which provides specific information about the scene and the characters involved. Some irrelevant details, such as the mention of “dinner time” and the “dagger which belonged to lady Hester Blackwood” neither of which are present nor implied in the original prompt, indicate that the original LLM is generating content not anchored to the given prompt.

In its untuned state, Baichuan-7B demonstrates a limited ability to sporadically capture the essence of the CoC atmosphere and integrate character specifics. However, it struggles to adhere precisely to the specifics laid out in the scene and character cards, showing inaccuracies and deviations. This highlights the essential role of fine-tuning in ensuring the LLM consistently generates precise, contextually anchored, and persona-driven dialogues. This is especially relevant for role-playing scenarios with granular prompts in our research.

IX. Emphasizing Persona Traits

To enable the LLM to generate precise, persona-consistent dialogues, it is crucial to highlight the specific traits described in the character and scene cards in our research. While our initial instructions sought to guide the LLM to focus on traits, the need for controllable, high-quality persona-driven dialogue required the development of a more innovative mechanism.

Aiming to enhance transformer models’ performance on specific language or vision tasks, previous research such as ChatGLM [34] and BLIP [35] manipulated the attention mechanism applied in the transformer block during the pre-training phase. However, such methodologies, which necessitate modifications to the micro or macro design of the neural network architecture, pose challenges for the LLM fine-tuning stage. Specifically, these approaches either make the pre-trained weights redundant or involve recomputing the weights on a large corpus. These significant modifications could adversely affect the efficiency of the fine-tuning process.

Drawing inspiration from the attention mechanism intrinsic to transformers, we postulate that introducing a unique marker as a prefix to traits within the card prompt could enhance the model’s attention to them. In our approach, we opted for the “@” symbol as the specific marker to ensure distinction from our generated dialogue content. This marker-based strategy provides two advantages. First, the intrinsic attention algorithm in the transformer can more efficiently aggregate trait information from the cards when the specific marker is embedded as a prefix. Second, it ensures that the model’s architecture and vocabulary remain untouched, which plays an essential role in effective fine-tuning.

To validate the effectiveness of this approach, we inspected the tokenization behavior of Baichuan-7B’s tokenizer on both the original prompt and the prompt augmented with our specific marker. As expected, the tokenization of the original prompt yielded no specific tokens that inherently emphasize traits within the cards. This absence leaves the LLM with limited guidance to differentiate prompt tokens and dialogue tokens during supervised fine-tuning. However, with our modified prompt, the “@” prefix marker does not merge with adjacent words and remains isolated after tokenization. This isolated token of the marker symbol could effectively signal the contextual information of the succeeding trait description. Consequently, this token could assist the LLM to distinguishing between the trait prompt and the dialogue content. More specifically, as the LLM learns how to utilize the provided cards for generating persona-driven dialogue, it could effectively identify the trait information signaled by the marker token. However, this method represents a speculative attempt prior to actually fine-tuning the LLM. A quantitative analysis of this specific marker approach, based on the heatmap of attention matrix, will be provided in subsequent section.

X. Experiment Methodology

We conducted the fine-tuning experiments for Baichuan-7B on a robust hardware setup, consisting of three Nvidia A100-SXM-80GB GPUs. The AdamW optimizer was employed to optimize the LLM, initiated with a learning rate of 1e-8. The model parameter datatype was configured to bf16, a format that provides a broader numeric range and greater stability compared to fp16 during backward loss computation. Moreover, the micro-batch size per GPU and gradient accumulation...
steps were set to 1. We leveraged DeepSpeed Zero Stage 3 for distributed training. This stage partitions the model parameters, gradients, and optimizer across multiple GPUs, effectively reducing GPU memory demands for LLM training.

Given the volume of our persona-driven dialogue training data, the model was fine-tuned over three epochs with a window length of 4096 tokens, which sufficiently encompassed the length of our dialogue data tokens. Additionally, an End-of-Sequence (EOS) token was appended to the dialogue token list to delineate the termination of the dialogue generation. This token is crucial for our fine-tuned LLM inference stage, as the original generation function requires an EOS token to halt the autoregressive generation process. A common practice in fine-tuning instruct LLM involves masking the system prompt in the labels, as it does not directly contribute to response generation. Similarly, our fine-tuning LLM does not require the generation of the system prompt, character, and scene cards; it directly generates tokens for persona-driven dialogue in an autoregressive manner. We also assigned a value of -100 to these prompts in the labels, indicating their exclusion from the loss calculation during fine-tuning. Moreover, user input content is often masked during fine-tuning of a multi-turn chat LLM, since the LLM is not tasked with generating user input.

To enable the fine-tuned LLM to generation CoC-style persona-driven dialogue, we adopted the official script settings available on the Baichuan-7B repository. For instance, the default generation hyperparameter configuration specifies a maximum new token limit of 2048 and a repetition penalty is 1.1.

XI. EXPLORATION OF CoC PERSONA-DRIVEN DIALOGUE GENERATION

To mitigate overfitting during supervised fine-tuning of Baichuan-7B, the diversity and the volume of training data are of superior importance. Our dataset comprises eight character cards, utilized to produce tens of thousands of persona-driven dialogues with GPT-3.5. While the implementation of diverse scene cards was aimed at enhancing dialogues diversity, it is essential to evaluate the generalization ability of our fine-tuned Baichuan-7B model when exposed to previously unseen scene and character cards. We employed GPT-4 to generate novel scene and character cards, ensuring that the card format aligns with our established pattern. To evaluate our fine-tuned Baichuan-7B model, these newly generated cards were used as prompts to synthesize CoC persona-driven dialogues and the instance generated dialogue is illustrated in Appendix Figure 9. To maintain measurement consistency, we adopted the same metrics that were used to assess the quality of dialogues generated by the untuned LLM.

A. Contextual Adherence

The dialogue unfolds within the framework of the Hidden Library of Alexandria. The cryptic tomes and prophecies, mentioned at the head of the conversation, confirm consistency with the essential elements outlined in the scene card. Moreover, the dialogue emphasizes the crucial object, “The Lost Tome”, and the obstacles, the Library’s traps and spectral guardians.

B. Character Fidelity

Dr. Benjamin: The narrative accurately captures his methodical and analytical traits. His pursuit to decode symbols corresponding to prophecies reflects the demands of his profession. Moreover, his skeptical reaction to Ms. Isabelle’s mystical talisman implies his methodical and critical persona. Nevertheless, the generated dialogue omits the description of his specific quirks, such as adjusting his glasses and preferring handwritten notes, suggesting a potential area for improvement.

Ms. Isabelle: The dialogue, through the metaphorical assertion “Knowledge without wisdom is dangerous” exhibits her whimsical and poetic nature. Her enigmatic and unpredictable
traits are revealed through sudden thematic shifts, such as mention the dead guarding the entrance while planning travel to next location. Her quirk of carrying tarot cards is naturally manifested when addressing the obstacle posed by the library’s spectral guardians.

C. Narrative Flow and Consistency

The generated dialogue logically progresses from discussing the library’s potential dangers to the act of deciphering prophecies. More specifically, the conversation trajectory, which begins with entering the library, confronting its inherent obstacles, and ultimately leading to the discovery of the lost tome, creates a coherent narrative.

D. Prompt Conformity

The lost tome, identified as the important object in the scene card, serves as the central axis which the generated dialogue revolves. The characters are actively engaged in the goal of deciphering prophecies, consistently addressing and navigating the obstacles outlined in the scene card.

The fine-tuned LLM demonstrates an impressive capability in crafting dialogue that seamlessly aligns with the context provided by the scene and character cards. The generated dialogue not only highlights the character’s traits but also orchestrates a coherent and engaging conversation aligned with the specified cards. This indicates the potential of the LLM to generate manipulable, contextually consistent, persona-driven dialogues.

To evaluate the performance of the persona-driven dialogues generated by our fine-tuned Baichuan-7B, we employed the advanced GPT-4 to comparatively assess dialogues generated from GPT-3.5 and our fine-tuned model. The scene and character cards were randomly sampled to formulate prompts, which served as input for both the fine-tuned LLM and GPT-3.5. To ensure the validity of this experiment, these cards were newly generated by GPT-4 and were kept separate from the training dataset.

The results of this comparative analysis are presented in Figure 6. The average score difference between the dialogues generated by the two LLMs is approximately 0.5, with our fine-tuned LLM surpassing GPT3.5 in some instances. This outcome underscores the high quality of persona-driven dialogues produced by our fine-tuned Baichuan-7B, which achieved comparable scores to GPT3.5 despite the latter housing over 100 billion parameters. Moreover, this result significantly implies that the specialized capabilities inherent to larger LLMs can be transferred to relatively smaller models with minor compromises in quality. Many unexpected and advantageous abilities emerge in large LLMs after extensive training with trillions of tokens. Many real-world applications may only require capabilities specific to their own domain. Accessing large models for niche abilities is costly and energy-consuming. Furthermore, directly training a model to capture these specific capabilities is not always feasible due to the lack of specialized datasets and task designs. Our methodology illustrates the practicality of equipping smaller models with the powerful abilities of larger models through careful prompt
engineering, data sampling, and supervised fine-tuning.

XII. MECHANISM OF PREFIX MARKER TOKENS

The aforementioned evaluations highlight the capability of our fine-tuned LLM in generating CoC persona-driven dialogues. Delving deeper, we aimed to illuminate how information is compressed and propagated during persona-driven dialogue generation. We posit that understanding the role and function of the prefix marker token could provide insights for further research on this topic. In order to investigate the mechanism of the prefix marker token, we first focused on the scene card prompt. We analyzed the attention scores sourced from the first self-attention layer of the fine-tuned LLM with the normal input prompts. Given that the multi-head attention mechanism is implemented, the average attention scores are calculated to facilitate a clearer visualization of the results, as depicted in the left side of Figure 5. It illustrates the relationship between the scene card prompt words and the corresponding token indices. Four distinct prefix marker tokens, represented as ‘@’, are employed to underscore the elements: location, object, goal, and obstacle within the scene card.

The results depicted in Figure 5 left-side exhibit four striking spikes, denoting elevated self-attention scores associated with the prefix marker tokens. These markers are influenced by the rotary embedding positional encoding scheme employed in the LLM, exhibiting a forward linear decay in attention score distribution. The plot shows that scores immediately drop to zero. It is caused by the causal mask mechanism, which masks subsequent tokens prior to the softmax calculation, thereby limiting the token’s attention range. A noticeable pattern emerges wherein the marker token tends to focus more on the latter token in compound words or phrases. According to the token-word relationship presented in Appendix Figure 10, an elevated attention score is evident on the token at index 4 which represents the final token of the word “scenario”. Similarly, prefix markers for object, goal, and obstacle also distinctly focus on the token at index 17, corresponding to the word “of” within the phrase “Hidden Library of Alexandria”. This distinctive behavior is presumed to result from the interaction between the causal mask and the rotary embedding mechanism during the fine-tuning process. Under the causal mask constraint, tokens can only gather information from preceding tokens, with later tokens potentially compressing more forward information. However, with the rotary embedding method, a linear decay is introduced to the attention scores, making each token more inclined to attend to its neighboring tokens. Based on this analysis, we conclude that the incorporated prefix marker tokens perform a vital role in context information compression. Their attention to key informational tokens demonstrates potential in enhancing the model’s ability to generate context-coherent, persona-driven dialogue.

To investigate the LLM’s mechanism for compressing information from scene card prompt and to illuminate the interplay among tokens in input processing, we visualized attention interactions across all tokens, using the scene card as input. This visualization is depicted through the attention heatmap presented in the right side of Figure 5.

Within the graph, each row of the lower triangular attention score matrix represents the average attention scores, with color intensity corresponding to the magnitude of these scores.

Rows: Represent the tokens on which a particular token (indexed by the row) focuses, and these tokens contribute their compressed information to the indexed token.

Columns: Indicate the frequency at which a given token (indexed by the column) is attended to by subsequent tokens.

From this attention heatmap, tokens at indices 4 and 17 consistently receive intensive attention from subsequent tokens, aligning with the observations made in previous analyses. Moreover, it is observed that tokens encapsulating crucial information, such as important objects, goals, and obstacles (indexed at 25, 32, and 45), tend to attract high attention scores from succeeding tokens. This results in a non-uniform attention intensity distribution within the graph. Tokens within the scene card prompt do not distribute attention uniformly across all forward tokens. Instead, there is a significant tendency to focus on ‘marker’ tokens, which compress necessary information. For instance, a later token at index 54 focuses on several previously discussed tokens that have condensed unique information.

If information can only be compressed once, later tokens in the sequence require long-range attention to access the information compressed by earlier tokens. However, the positional embedding algorithm tends to suppress long-range attention, which can result in potential loss of prior information for long sequence inputs. The generated persona-driven dialogue requires the LLM to access the scene and character cards located at the beginning of the input sequence. This implies the possibility of well-organized information compression management that enables efficient information flow within a limited attention range. To investigate this, we conducted a subsequent experiment involving the word-wise average attention heatmap on full prompts and the persona-driven dialogue in Figure 7.

XIII. HIERARCHICAL ATTENTION CASCADE

In our analysis, we consider the prompt components, including the system prompt, scene, and character cards, which are indexed from 0 to 338. The generated persona-driven dialogue spans indices 339 to 719 on the heatmap. A salient pattern emerges as a series of bright intersecting lines within the index range of the generated dialogue, forming a ladder-shaped appearance. We term this the Hierarchical Attention Cascade (HAC). This pattern provides the following two key insights:

Role of Anchor Words: Words at these bright intersections act as pivot points, absorbing information from preceding words while also exerting a primary influence on the LLM to generate subsequent tokens. These anchor words directly correspond to character names, which denote the beginning of each conversational turn. The context of each generated dialogue turn incorporates the compressed information for the name words in previous turns.

Recency Focus: The HAC demonstrates a recency focus at both the word and conversational turn levels. The anchor word
exhibits intensive attention to its immediate predecessors, with attention decaying for more distant words. Moreover, anchor words consistently draw attention from words located down the sequence, with attention decaying for more distant turns.

In contrast, the prompt region of the heatmap displays numerous intensely bright rows, indicating a multi-fold interaction among proximate words. This recurrent compression of similar information provides two potential advantages:

**Resilience to Information Decay:** The iterative information compression compensates for the side effect of attention linear decay, enabling subsequent words to effectively access earlier information.

**Information Propagation:** This cascading compression shortens the relative distance between prompts and dialogues, allowing the LLM to generate context-coherent, persona-driven dialogue.

The results indicate that our fine-tuned LLM primarily utilizes anchor words, such as character names, to incorporate narrative context information in early turns. For complex and long-range contexts, the LLM employs a recursive compression strategy, ensuring effective information passage despite inherent attentional decay. These insights reveal the front-layer attention interaction mechanism and offer a promising foundation for future explorations aimed at refining dialogue generation models.

Furthermore, our research delved into the complex dynamics of information propagation and compression within the Transformer framework during dialogue generation. The attention heatmaps revealed several intriguing insights: we identified the existence of ‘anchor words’ that act as reservoirs, aggregating information from structured dialogue turns. Moreover, the Hierarchical Attention Cascade (HAC) mechanism plays a vital role in interpreting complex card prompt details. This layered information is seamlessly integrated into the dialogue as the LLM generates its responses.

Overall, our research illuminates promising pathways for future exploration of prompt formats, token dictionaries, and attention mechanisms, which could further enhance the performance of LLMs in dialogue generation tasks.

**ACKNOWLEDGMENTS**

This work is supported by OpenAI’s Researcher Access Program.

**APPENDIX**
Fig. 8. Dialogue generated from untuned Baichuan-7B.

| Dr. Benjamin: (looking at a dusty room) | My word, this library holds many secrets that could unlock the mysteries of the universe. But we must proceed with caution, for there may be hidden dangers lurking within its pages. |
| Ms. Isabel: (nodding in agreement) | Indeed, Dr. Benjamin. And yet, the knowledge contained within these walls is worth any risk. Perhaps my skills in chronicle and ritual magic can help us decipher the cryptic prophecies without falling prey to the Library’s traps. |
| Dr. Benjamin: (impressed) | Your expertise in the occult never ceases to amaze me, Ms. Isabel. With your guidance, perhaps we can uncover the truth behind these enigmatic tomes. |
| Ms. Isabel: (sighing) | Thank you, Dr. Benjamin. But let us not forget the importance of balance. As the saying goes, “Knowledge without wisdom is dangerous.” |
| Dr. Benjamin: (nodding thoughtfully) | Wise words indeed. Now, let us focus on the task at hand. According to this map, the lost tome should be located in the restricted section of the library. |
| Ms. Isabel: (nodding up and down) | Ah, but what if the spirits of the dead guard the entrance? Fear not, dear doctor, for I have brought along some protective talismans. They should ward off any malevolent entities. |
| Dr. Benjamin: (gazing at an embroidery) | Protective talismans, you say? How intriguing. Let us hope they prove effective against whatever supernatural forces may be ahead. |

As they make their way through the maze of shelves and corridors, they encounter several puzzling traps and spectral guardians. But with Dr. Benjamin’s keen eye for artifacts and Ms. Isabel’s quick thinking, they manage to avoid each obstacle. Finally, they reach the restricted section and find the lost tome.

| Dr. Benjamin: (beaming through the glass) | Wise words indeed. Now, let us focus on the task at hand. According to this map, the lost tome should be located in the restricted section of the library. |
| Ms. Isabel: (studying the symbols intently) | Yes, yes, I see now. It seems that the key to unlocking the prophecies lies in the very fabric of reality itself. We must perform a series of rituals to restore balance between the worlds. |
| Dr. Benjamin: (sighing bitterly) | Farewell! And how exactly do we go about doing so? |
| Ms. Isabel: (sighing confidently) | That, my friend, is where your expertise comes in. Together, we shall unravel the mysteries of this eddying enlightenment and emerge victorious over the forces of darkness. |

Fig. 9. Figure illustrates the CoC-style persona-driven dialogue generated from fine-tuned Baichuan-7B.

Fig. 10. Figure illustrates the tokenization result after applied Biachuan-7B tokenizer to the given scene card prompt.

REFERENCES


