LLMs and Conversational Games

Louis-Vincent Poellhuber
louis.vincent.poellhuber@umontreal.ca
Bole Yi
bole.yi@umontreal.ca
Rafaela Souza Pinter
rafaela.souza.pinter@umontreal.ca

Abstract

Playing conversational games is an underexplored task for Large Language Models (LLMs), which requires remembering massive history information and making logical decisions during the gameplay. In this project, by testing on two conversational games with different basic rules, Ask-Guess and Spyfall, we aim to study previous conversational game frameworks [1, 2, 3, 5, 6, 7], along three dimensions: knowledge extraction, prompt tuning and LLM choice. This project seeks to provide a novel viewpoint on comprehending LLMs' capabilities on conversation-based applications and explore their potential in diverse contexts, to uncover new insights into the effectiveness and adaptability of LLMs in facilitating conversational natural language interactions.

1 Introduction

The realm of Natural Language Processing (NLP) has witnessed remarkable developments in recent years, especially with the evolution of Large Language Models (LLMs) such as GPT[15], Claude[14], Gemini [13] and Cohere [12]. These models, with the ability to comprehend and generate human-like text, have not only revolutionized various NLP tasks [10] but have also for new applications addressing real-world challenges, including conversational games[3]. Conversational games refer to structured activities or exercises designed to facilitate communication, interaction, and social engagement between individuals or within groups, which is the one of the best areas to test language models and their behaviors. The challenge remains as though LLMs excel in understanding and generating language based on patterns in data, they lack the deeper contextual understanding and associative memory capabilities of humans.

In recent work, one exploration to Werewolf games [7] showed an effective way to interact with massive historical information in LLMs using experience pools with GPT-3.5. Also, though fine tuning LLMs are impractical now, by leveraging pre-trained models and adapting them using techniques like prompt engineering [11] can offer us a more practical approach in many cases. As the conversational games are not thoroughly explored in previous LLM research. Additionally, the comparison between multiple LLMs and the simultaneous implementation of knowledge extraction and prompt engineering techniques represent new areas for investigation. In this project, we aim to gain deeper insights into the behavior of LLMs and uncover new discoveries about their capabilities.

Our project proposal can be summarized as follows: (1) Re-implement two word guessing games Ask-Guess and Spyfall on different LLMs to compare their performances. (2) Conduct knowledge extraction by storing information into experience pools and retrieve it with fine-tuned Q&A BERT model, make LLMs recall information efficiently. (3) Try different prompt engineering techniques to enhance the model performance, test the effects of chain-of-thought prompting.

2 Related Work

Large Language Models (LLMs) have been utilized to play various games, including Werewolf, Avalon, Ask-Guess, SpyFall, TofuKingdom, and Zork [1, 2, 3, 4, 5, 6, 7], which serve different purposes in understanding LLM capabilities, such as the model's ability to navigate within a location-based narrative and understand social interactions.

GameEval [3] introduces three games, Ask-Guess and SpyFall being relevant to our work, with specific metrics for LLM evaluation, focusing on cooperative and adversarial dynamics. The effectiveness of these methodologies in providing a comprehensive assessment of LLMs has was noted.

In exploring the game Werewolf [7], the use of ChatArena was interesting in coordinating LLM

gameplay. The authors highlight the importance of experience pools, a collections of game-related data, for enhancing performance. The integration of BERT for processing game text and extracting suggestions from the experience pool is noted for its significant impact on performance and a relevant method for our project.

Lan *et al.* [1] explored the game Avalon for investigating LLM-based agents' social behaviors through gameplay. Notably, a six-module framework was proposed to facilitate this, including summary, analysis, planning, action, response, and experience learning, revealing varied social behaviors among LLMs when interacting with different characters.

3 Task and dataset

In this project our team is going to implement two word guessing games previously implemented by GameEval [3] and study the impact of various optimization techniques used by other works [1, 7]. Additionally, we are going to compare the performance of different LLMs, to try and see if different behaviors emerge.

More specifically, we are going to re-implement the games Ask-Guess and Spyfall. Ask-Guess is a game where one player randomly picks a word and attempts to make the other player guess it, without saying it. If time permits and if performance is good, we are also considering implementing its sibling, Taboo, where each word picked comes with a subset of related words that cannot be said either. For both games we'll be using a Taboo dataset [8] originally created for a web browser game. SpyFall is another word-guessing game, where six players pick a common word. A spy either doesn't know the real word or has a similar word, depending on the complexity wanted. For the dataset, there is none that is publicly available, so we will make our own, using the Taboo dataset as a base. We will manually select appropriate related word pairs.

4 Methods

To execute this project, we'll use the same framework used by Qiao *et al.* [7], ChatArena [9]. It is a Multi-Agent Language Game Environments for LLMs, allowing us to easily communicate with different LLM APIs in a game setting. We'll study three dimensions to conversational games: the model choice, the knowledge retrieval and the prompt engineering. To compare LLMs, we'll use some of those offered by ChatArena: OpenAI's GPT3.5(*gpt-3.5-turbo-0301*), Anthropic's Claude, Google's Gemini and Cohere. The use of different LLM models as agents has not been thoroughly studied in the context of the games we implement, making this part of our approach original. Besides different models, we'll also implement a knowledge extractor. This important step of our approach helps alleviate the issue of context length. A common approach consists of first storing information into an experience pool, optionally using some kind of compression technique, then using a mechanism to retrieve that information, so that the LLM only takes the currently relevant history as input. We will create two experience pools exclusive to each agent: a literal history and a compressed history. For the latter, we'll use a prompt, which will task an LLM to summarize the information from a round. For experience retrieval, we'll use the following techniques: another summarization prompt, sentence embeddings with a similarity to retrieve the K most similar experiences, and a fine-tuned BERT model, as described in [7]. We'll also use chain-of-thought prompting to help the model reason.

4.1 Baselines and evaluation

Our baselines will be evaluated through an ablation study. For each studied dimension (model, knowledge retrieval and prompt engineering), we'll iteratively compare the different techniques used to their respective baseline. Then, we'll compare our best model with a predetermined baseline. For the model choice, we'll use GPT3.5 as a baseline, because it is a very commonly used model and because its newer version, GPT4, is the state-of-the-art. If the LLM choice has little impact on performance, we'll use the GPT3.5 model as the default model. For the knowledge extraction, we'll use the sentence embedder with the literal experience pool as a baseline. Finally, for prompt tuning, we'll use zero-shot as our baseline, that we'll compare to chain-of-thought prompting.

Evaluation will be game-specific, but will have the success of the game as its main objective. To do so, we'll use the metrics defined in GameEval [3]. For Ask-Guess and Taboo, we'll use the following metrics: successful trial, ending error, round limit error, answer mentioned error and chat error.For SpyFall, we'll use the following metrics: spy winning rate and spy living round. Please refer to the article [3] for a detailed explanation.

5 References

[1] Yihuai Lan, Zhiqiang Hu, Lei Wang, Yang Wang, Deheng Ye, Peilin Zhao, Ee-Peng Lim, Hui Xiong, and Hao Wang. 2023. LLM-Based Agent Society Investigation: Collaboration and Confrontation in Avalon Gameplay. arXiv:2310.14985 [cs].

[2] Tian Liang, Zhiwei He, Jen-tse Huang, Wenxuan Wang, Wenxiang Jiao, Rui Wang, Yujiu Yang, Zhaopeng Tu, Shuming Shi, and Xing Wang. 2023. Leveraging Word Guessing Games to Assess the Intelligence of Large Language Models. arXiv:2310.20499 [cs].

[3] Dan Qiao, Chenfei Wu, Yaobo Liang, Juntao Li, and Nan Duan. 2023. GameEval: Evaluating LLMs on Conversational Games. arXiv:2308.10032 [cs].

[4] Murray Shanahan, Kyle McDonell, and Laria Reynolds. 2023. Role-Play with Large Language Models. arXiv:2305.16367 [cs].

[5] Chen Feng Tsai, Xiaochen Zhou, Sierra S. Liu, Jing Li, Mo Yu, and Hongyuan Mei. 2023. Can Large Language Models Play Text Games Well? Current State-of-the-Art and Open Questions. arXiv:2304.02868 [cs].

[6] Dekun Wu, Haochen Shi, Zhiyuan Sun, and Bang Liu. 2023. Deciphering Digital Detectives: Understanding LLM Behaviors and Capabilities in Multi-Agent Mystery Games. arXiv:2312.00746 [cs].

[7] Yuzhuang Xu, Shuo Wang, Peng Li, Fuwen Luo, Xiaolong Wang, Weidong Liu, and Yang Liu. 2023. Exploring Large Language Models for Communication Games: An Empirical Study on Werewolf. arXiv:2309.04658 [cs].

[8] K. Woblick, 'Kovah/Taboo-Data'. Oct. 24, 2023. Accessed: Feb. 11, 2024. [Online]. Available: https://github.com/Kovah/Taboo-Data

[9] 'Farama-Foundation/chatarena'. Farama Foundation, Feb. 10, 2024. Accessed: Feb. 11, 2024. [Online]. Available: https://github.com/Farama-Foundation/chatarena

[10] Sébastien Bubeck, Varun Chandrasekaran, Ronen Eldan, Johannes Gehrke, Eric Horvitz, Ece Kamar, Peter Lee, Yin Tat Lee, Yuanzhi Li, Scott Lundberg, Harsha Nori, Hamid Palangi, Marco Tulio Ribeiro, and Yi Zhang. 2023. Sparks of Artificial General Intelligence: Early experiments with GPT-4. arXiv:2303.12712 [cs].

[11] Jules White, Quchen Fu, Sam Hays, Michael Sandborn, Carlos Olea, Henry Gilbert, Ashraf Elnashar, Jesse Spencer-Smith, and Douglas C. Schmidt. 2023. A Prompt Pattern Catalog to Enhance Prompt Engineering with ChatGPT. arXiv:2302.11382 [cs].

[12] 'Home', Cohere. Accessed: Feb. 12, 2024. [Online]. Available: https://cohere.com/

[13] 'Gemini - chat to supercharge your ideas'. Accessed: Feb. 12, 2024. [Online]. Available: https://gemini.google.com

[14] 'Claude'. Accessed: Feb. 12, 2024. [Online]. Available: https://claude.ai/login?returnTo=

[15] 'ChatGPT'. Accessed: Feb. 12, 2024. [Online]. Available: https://chat.openai.com