AI Charades: Language Models as Interactive Game Environments

Kevin Frans
Cross Labs, Tokyo, Japan
Massachusetts Institute of Technology, Cambridge, MA
kvfrans@csail.mit.edu

Abstract—Language models in recent years have shown astounding growth in modeling future text prediction. From a games perspective, language models present richly semantic environments, in which players can deeply influence the model’s behavior through text input. This work presents a framework for building Language Model Games, games which center around a player manipulating a language model into behaving in a desired manner. A small demo, AI Charades, showcases a proof-of-concept game in which a player must communicate a secret word to an AI without mentioning it directly. Later discussion explores language models as a general tool for human-AI interfacing, along with their capabilities as design tools.

Index Terms—language models, game interactions, AI players

I. INTRODUCTION

In recent years, strong language models have pushed the boundaries of natural language prediction. Simply put, the task of a language model is to predict the next word, given a sequence of text. Trained on large human text corpora, strong language models display surprising amounts of generalizability, and have been shown to have strong zero-shot learning performance on a range of tasks such as translation or summarization [1, 2].

From a games perspective, language models present a promising new domain for rich interactions. An autogressive language model is a stateful system: passing in a sequence of words adjusts the probability distribution of the next word. Thus, interacting with a language model through text can be seen as taking actions in a semantically-structured environment. By examining model outputs following certain phrases, the state of the environment is readable, presenting an rich interactive domain to build games around.

This work presents a framework for building Language Model Games, a new class of games centered around players manipulating AI language models into behaving in a desired manner. A small demo, AI Charades, is presented as a playable proof-of-concept. Players are given a random secret word; they must then input sentences into the language model, such that the model outputs the correct word when prompted. Later discussion centers on language models as a natural-language-based paradigm for designing games.

II. DEMO

To play AI Charades or view source, visit the demo link: https://colab.research.google.com/github/kvfrans/aitype/blob/main/ai_secretword.ipynb

III. DESIGN

AI Charades is a small Language Model Game in which players must communicate a secret word to an AI, without mentioning the word directly. Specifically, a random word from a list is sampled. Players are then given the opportunity to freely enter a set of player-interaction phrases, with the goal of influencing the AI by introducing context.

The player’s success is measured through a state-reading phrase: a pre-written phrase which presents a prompt for the
Language models are a fruitful new paradigm for game design. Viewed as standalone systems, language models present a rich semantic environment, whose dynamics can be easily influenced and interpreted through natural language. Games such as AI Dungeon highlight the capacity of language models in collaborative storytelling [3]. Simple gamified frameworks around strong language models, such as AI Charades, gain a level of depth similar to games built around interacting with other players. In both cases, game rules simply present goals and incentives. Replayability and depth are achieved through the exploration of deeply complex systems: traditionally the minds of other players, in the case of Language Model Games, the dynamics of the AI.

A. Language Model

In this work, a pre-trained GPT-2 transformer network is used as the model [1]. The use of a pre-trained model presents several benefits. First, no fine-tuning is required, and thus no additional training or data collection is required. Second, general language models present a rich semantic environment to interact with. In Language Model Games, a large portion of the challenge is encompassed within the natural dynamics of the language model – players must figure out which phrases affect the AI, what grammar is understood, etc. Outer structure, such as providing a secret word in AI Charades, serves mainly as a framework of goals to encourage exploration of the language model dynamics. Note that any sufficiently-trained language model – players must figure out which phrases affect the AI, what grammar is understood, etc. Outer structure, such as providing a secret word in AI Charades, serves mainly as a framework of goals to encourage exploration of the language model dynamics. Note that any sufficiently-trained language model can be substituted in, and different models will often provide dynamics that are related (basic grammar rules apply) but distinct (language models bias to their training dataset).

B. Language Models as Design Tools

An interesting direction lies in the use of language models as design tools. Already, the behaviors of language models can be influenced by the text they are given, e.g. changing the state-reading phrase from “The word I’m thinking of is” to “This story makes me feel” will extract different information. Further down this line is the idea of text input as training data in few-shot learning. Powerful models such as GPT-3 have shown language models can learn on-the-fly from input text: given several math problems, followed by an unfinished problem, the language model correctly completes the answer [2]. As game designers, this capability presents a novel way to design game dynamics, purely through natural language. One can imagine a scenario in which the personality of an NPC is encoded through a series of background Q&A prompts given before, means that task-specific training becomes unnecessary [2]. Thus, powerful pre-trained language models can be put to use as interfaces in games and beyond.

C. Limitations

The limitations of Language Model Games are, unsurprisingly, the language models themselves. While language models present a range of rich interactions, it is hard to control these interactions precisely. Especially when using pre-trained models, it is often the case that the language model does not display the behavior a designer is looking for. One workaround is to treat the language model as a source of structured uncertainty, akin to a human partner in games like Charades and Pictionary – while it’s hard to guess what your partner will do, you can generally depend on them to behave in certain ways. Games designed around language models will have to take this uncertainty into account, and depend on the versatility of the language model rather than its rigidity.

REFERENCES