AIRL: AI-aligned Reinforcement Learning and Dialogue Generation

Yuyang Cao* **Bardia shahrestani***
yuyang.cao@mail.mcgill.ca, bardia.shahrestani@mail.mcgill.ca

*equal contribution

Abstract

In this paper, we present a novel approach to emulate Socrates' character from Plato's works using AI-aligned Reinforcement Learning (AIRL) of pre-trained generative models with Proximal Policy Optimization (PPO). Combining Low Rank Adaptation (LoRA) and GPT-Neo-1.3B, we create an efficient model for realistic dialogue generation in virtual environments with predefined characters. Our results demonstrate that the PPO-trained LoRA model generates more concise and coherent responses, but further training and refinements are needed for improved alignment with Socrates' character. This study contributes to the development of advanced conversational AI systems using reinforcement learning techniques.

1 Introduction:

Large Language Models (LLMs) have recently gained popularity for their ability to generate human-like language. However, they suffer from several limitations, including the inability to maintain a specific style over an extended conversation and the necessity for industrial-scale GPUs for training and running. In this work, we address these issues by employing AI-aligned Reinforcement Learning (AIRL) using Proximal Policy Optimization (PPO) to fine-tune GPT-Neo-1.3B with Low Rank Adaptation (LoRA). Our goal is to emulate the character of Socrates from Plato's works, thus creating a model capable of generating realistic dialogue in virtual environments without regressing to a default style.

2 Background:

Recent advancements in Reinforcement Learning (RL) systems have enabled complex agent-environment interactions. Deep Reinforcement Learning from Human Preferences (Christiano et al. 2017)[1] has shown the potential of learning complex tasks based on human preferences without direct access to the environment's reward function.

Efficient methods for adapting large-scale models to specific tasks have been developed, such as LoRA: Low-Rank Adaptation of Large Language Models (Hu et al. 2021)[3]. This approach effectively reduces the number of trainable parameters, GPU memory requirements, and inference latency without compromising model quality.

Fine-tuning GPT-3 using human feedback for alignment was demonstrated in Training Language Models to Follow Instructions with Human Feedback (Ouyang et al. 2022)[7]. InstructGPT models outperformed the larger GPT-3 model in human evaluations, with improvements in truthfulness and reductions in toxic output generation.

The potential of large language models for text-annotation tasks was explored by ChatGPT Outperforms Crowd-Workers for Text-Annotation Tasks (Gilardi et al. 2023) [2], showcasing the superiority

37th Conference on Neural Information Processing Systems (NeurIPS 2023).

of ChatGPT over crowd-workers for various annotation tasks at a substantially lower per-annotation cost.

Proximal Policy Optimization Algorithms (PPO) (Schulman et al., 2017) [8] introduced a new family of policy gradient methods for RL, offering a favorable balance between sample complexity, simplicity, and wall-time.

Our project aims to develop a novel approach for emulating literary characters using AI-aligned Reinforcement Learning of Pre-trained Generative Models with Proximal Policy Optimization, leveraging advancements in human preference-based RL, low-rank adaptation, instruction-following language models, and PPO.

3 Related Works

Several recent studies have focused on the evaluation and generation of dialogues using pre-trained language models. Lee et al. [4] proposed a protocol for evaluating conversational models using head-to-head pairwise comparison, with DialoGPT and Blender emerging as superior systems. Li et al. [5] introduced DIALOGIC, a dialogue simulation method that leverages GPT-3's in-context learning capabilities to automate dataset creation, achieving near-human fluency and annotation accuracy. Mohapatra et al. [6] presented a data creation strategy using GPT2 to simulate user and agent interactions, leading to significant improvements in low-resource settings on MultiWOZ and Persona chat datasets. Lastly, Zhang et al. [9] developed DialoGPT, a large-scale generative pre-training model for conversational response generation, which outperformed strong baselines in generating relevant and context-consistent responses.

4 Methodology:

Our methodology involves the following steps: Our methodology consists of the following steps:

1. **Fine-tuning**: We fine-tune a LoRA of GPT-Neo-1.3B on a manually cleaned dataset of Plato's dialogues featuring Socrates. Custom callbacks are implemented for version control and LoRA model saving, although the former was not utilized in the final training due to performance limitations.
2. **Pipeline Evaluator**: We construct a pipeline evaluator using OpenAI's gpt-3.5-turbo chat model as a reward function. This model evaluates responses and generates a positive or negative sentiment after reasoning. A sentiment classification model further translates these outputs into scalar values of 1.0, 0.0, or -1.0.
3. **Reinforcement Learning**: We employ the TRL library to train a new LoRA based on a hybrid dataset, which comprises both pre-generated synthetic responses and online results from our pipeline evaluator. Various strategies are explored, including single and multiple evaluators, as well as different evaluation criteria.
4. **Model Adaptation**: The LoRA method enables efficient switching of attention layers depending on the desired character generation, allowing us to effectively store the writing style or "personality" of Socrates in a compact 13MB format.

By integrating these steps, we demonstrate an effective approach to generate text that emulates the character of Socrates using AI-evaluated criteria and reinforcement learning techniques Below, you can find the pseudocode for our architecture.

4.1 Pseudocode: Training LoRA with PPO

The following is the high-level structure of our LoRA training Process with PPO.

```
1. Import required libraries and models
2. Load pre-trained model with LoRA configuration
3. Initialize PPO Trainer
4. Prepare dataset from input text file
5. Create evaluation function using external language model
```

```
   and sentiment analysis model
6. For each evaluation scheme:
   a. Create an evaluator instance with the evaluation criterion
   b. For each prompt and response in the dataset:
      i. Perform a PPO training step with default reward
      ii. Periodically generate responses from the model
      iii. Evaluate generated responses using the evaluator
      iv. Update reward based on the evaluation
      v. Perform another PPO training step with the updated reward
```

4.2 Pseudocode: Evaluator

The following is the high-level structure of our evaluators.

```
1. Import required libraries and models
2. Define the Evaluate class:
   a. Initialize the evaluator with a given criterion,
      sentiment analysis tokenizer, and model
   b. Define the evaluate_verbal method:
      i. Call the external language model API with the text input
         and criterion
      ii. Return the evaluation response from the API
```

5 Supplementary Material

1. We used the Parameter-Efficient Fine-Tuning (PEFT) library provided by Huggingface [1] to efficiently adapt pre-trained language models to our downstream applications, without fine-tuning all the model's parameters. PEFT enables efficient adaptation of pre-trained language models to downstream tasks by fine-tuning only a small number of extra parameters. PEFT is integrated with HuggingFace Accelerate for large-scale models.
2. TRL is a library built on top of the transformers library by Hugging Face, which supports most decoder architectures and encoder-decoder architectures. It offers a PPO trainer for language models, requiring triplets of (query, response, reward) to optimize the model. The fine-tuning process involves rollout, evaluation, and optimization steps using the generated response, scalar value, and PPO algorithm. TRL is available on GitHub. [2]
3. The TRL wrapper by Huggingface supports all decoder and encoder-decoder model architectures in transformers with the AutoModelForCausalLMWithValueHead and AutoModelForSeq2SeqLMWithValueHead classes.[3]
4. "bert-base-multilingual-uncased-sentiment" is a BERT model fine-tuned for sentiment analysis on product reviews in six languages, predicting the sentiment as a number of stars. It can be used as a sentiment analysis model for any of these six languages or further fine-tuned for related tasks. [4]
5. Project Gutenberg offers a collection of books written by Plato, which can be accessed.[5] The website provides a sorted list of Plato's books by popularity, and includes a variety of works centered around Socrates.

6 Results:

In this study, we conducted a maximum of 120 steps for LoRA pretraining followed by 3600 PPO training steps on our LoRA model and qualitatively assessed the quality of the outputs from

[1] https://github.com/huggingface/peftPEFT
[2] https://github.com/lvwerra/trl/tree/main
[3] https://huggingface.co/docs/trl/models
[4] https://huggingface.co/nlptown/bert-base-multilingual-uncased-sentiment
[5] https://www.gutenberg.org/ebooks/author/93

the models with and without PPO training. Our observations revealed that the PPO-trained model produced more coherent and brief responses, likely due to the structure of our synthetic data. However, structurally, the model had not aligned its speaking structure with that of Socrates after 3600 steps. Furthermore, we found that the fine-tuned model was excessively verbose, with less coherence or semantic understanding, but with a better modelling of the overall sentence structure in Plato's works.

We hypothesize that with further training and a more detailed evaluation prompt for the external model, it would be possible to achieve true AI-aligned reinforcement learning with minimal human interference.

The following presents a comparison of example results from the PPO-trained LoRA and the LoRA without PPO.

6.1 Verbatim Examples

Example Results from PPO-trained LoRA:

- **Prompt 1:**
  PLATO: What is the true nature of justice, and how can we achieve it in our society?
  SOCRATES: Justice is the way we treat each other.
- **Prompt 2:**
  HIPPIAS: How can we define beauty in a way that encompasses all its manifestations? please explain briefly.
  SOCRATES: Beauty is the sum of all the virtues.
- **Prompt 3:**
  XENOPHON: In your opinion, what is the best form of government, and why? please explain briefly.
  SOCRATES: The best form of government is a republic.

Example results from LoRA without PPO:

- **Prompt 1:**
  PLATO: What is the true nature of justice, and how can we achieve it in our society?
  SOCRATES: How can there be a society if one and the same individual is a judge and others are to be put to the law? Now we must not say that justice is like the wind or the waves; but to say that justice is only like this, that one man and one thing shall receive, and [...]
- **Prompt 2:**
  HIPPIAS: How can we define beauty in a way that encompasses all its manifestations? please explain briefly.
  SOCRATES: To begin with, beauty is the most perfect thing which a body has. And among all perfect things there is none that is more beautiful than the one we have just mentioned; and since one of the elements in which beauty is contained is goodness, those two are related. On the other hand [...]
- **Prompt 3:**
  XENOPHON: In your opinion, what is the best form of government, and why? please explain briefly.
  SOCRATES: I shall do so. I have observed that the best forms of government are, as I mentioned to you before, the forms of life as lived by kings, and of common people, and the forms of government that are imposed by and in consequence of natural law, and which are the laws of justice, and of all the virtues and the punishments. Now these are the principles in which we learn to think; and [...]

These results indicate that although the PPO-trained LoRA generated more concise and coherent responses, the semantic understanding and alignment with Socrates' beliefs and values were not yet fully achieved. Further training and refinements to the evaluation prompt might lead to improved performance and better alignment with the desired character.

7 Conclusion:

This study showcases the potential of AI-aligned reinforcement learning using Proximal Policy Optimization and Low Rank Adaptation for adapting pre-trained generative models to specific tasks or styles, such as emulating Socrates' character. The combination of reinforcement learning techniques enables efficient fine-tuning of language models, reducing computational requirements.

The ability to generate tailored conversational styles and characters using reinforcement learning has significant implications for cognitive and virtual environments. Future work may explore other AI-based evaluation schemes, merging LoRAs, human-in-the-loop dialogue generation, procedural training, adaptation of other characters or styles, and further reducing computational requirements or improving generated content quality. This research contributes to the ongoing development of advanced conversational AI systems using reinforcement learning techniques.

References

[1] Paul Christiano et al. "Deep reinforcement learning from human preferences". In: *arXiv preprint arXiv:1706.03741* (2017).

[2] Fabrizio Gilardi, Meysam Alizadeh, and Ma"el Kubli. "ChatGPT Outperforms Crowd-Workers for Text-Annotation Tasks". In: *arXiv preprint arXiv:2303.15056* (2023).

[3] Edward J. Hu et al. "LoRA: Low-Rank Adaptation of Large Language Models". In: *arXiv preprint arXiv:2106.09685* (2021). DOI: `10.48550/arXiv.2106.09685`.

[4] Seolhwa Lee, Heuiseok Lim, and João Sedoc. "An Evaluation Protocol for Generative Conversational Systems". In: *arXiv preprint arXiv:2010.12741* (2020).

[5] Zekun Li et al. "Dialogic: Controllable Dialogue Simulation with In-Context Learning". In: *arXiv preprint arXiv:2210.04185* (2022).

[6] Biswesh Mohapatra et al. "Simulated Chats for Building Dialog Systems: Learning to Generate Conversations from Instructions". In: *arXiv preprint arXiv:2010.10216* (2021).

[7] Long Ouyang et al. "Training language models to follow instructions with human feedback". In: *arXiv preprint arXiv:2203.02155* (2022).

[8] John Schulman et al. "Proximal Policy Optimization Algorithms". In: *arXiv preprint arXiv:1707.06347* (2017).

[9] Yizhe Zhang et al. "DialoGPT: Large-Scale Generative Pre-training for Conversational Response Generation". In: *arXiv preprint arXiv:1911.00536* (2020).