Alignment Of Large Language Models Via Debate And Reinforcement Learning

Łukasz Kuciński

Artificial Intelligence (AI) has become a part of our daily lives, largely due to advancements in large language models (LLMs) like GPT. These models, which assist in tasks ranging from improving productivity to controlling robots, have profound implications for society, business, and governance. Consequently, their development and use are subject to increasing scrutiny and regulation, as seen in initiatives like the EU AI Act and President Biden's Executive Order on AI.

A critical aspect of AI development is ensuring these models align with human values and truths, a field known as AI alignment. Traditional methods involve training these models with human or AI feedback. However, a new and promising approach is using debates. In this method, models are trained to argue and defend their points, engage with opposing views, and persuade judges of their stance. This debate format not only helps in discerning truth but also drives the AI's development in understanding complex arguments and reasoning.

Our project aims to harness this debate format for training LLMs. We're inspired by recent advancements in debate analysis, teaching models, and curriculum design. Our method involves a 'teacher-student' framework where AI 'students' participate in debates, and a 'teacher' oversees and adapts the training process. We hypothesize that this approach will lead to the discovery of diverse debate rules and better align LLMs with human values and reasoning.

The project will produce a debate-based training protocol, a public dataset of debate transcripts, trained model weights, and source code. The outcomes are expected to enhance AI's ability to provide accurate, reliable answers, thereby improving our understanding of the world and aiding in decision-making. This research, targeting top AI conferences, could mark a significant step forward in AI alignment development.