What's on your mind?

Choose an option:

- (•) Lecture Review
- ( ) Ask Questions
- ( ) Test Yourself

TA-Lite: Academic Assistant

Logged in as: rarao@iu.edu

Select Lecture

Which lecture would you like help with?

Lecture 08: Reinforcement Learning from Human Feedback

Select Material Type

Lecture Slides | Lecture

Lecture Review

Choose what you'd like to review:

Generate Summary | Generate

Summary from slides

In this article, the authors propose a new approach to training language models (LLMs) using reinforcement learnin reasoning capability in LLMs by practicing solutions that take them from problem statements to final answers. The simulate human preferences and train an agent to follow these preferences. The agent receives rewards for good ac with the goal of maximizing cumulative rewards over time.

The authors describe the problem of training LLMs, which is to take a problem statement (prompt) and generate an However, current methods have limitations, such as relying on hand-crafted rules or relying solely on supervised le RL to train an agent to follow human preferences in generating answers.

The authors introduce the concept of a "preference model," which is a neural network that simulates human prefer of the LLM. They also define a loss function for the preference model, which is a negative log-likelihood function tha the human preference and the LLM's output. The goal is to minimize this loss function to incentivize the LLM to follo

The authors then describe their proposed approach, which consists of three steps:

1. Instance the environment: Create a simulated environment that includes the problem statement and the desire
2. Learn the agent policy: Use RL to train an agent to take actions in the environment and receive rewards based o
3. Evaluate the agent: Evaluate the trained agent using various metrics, such as accuracy and fluency, to determir

The authors also provide examples of how their approach can be applied in practice, such as in natural language pr to follow user preferences. They conclude that their proposed approach has the potential to improve the reasoning overall performance.