Module 11

Reading and Short Answer Assignment

Due by 11:59:59 PM on the date of our next module lecture.

This week you will watch an excellent video about the general recipe for training a model like ChatGPT (from mid 2023). It is presented by Andrej Karpathy, co-founder of OpenAI and former director of AI at Tesla.

Video (about 45 minutes): https://youtu.be/bZQun8Y4L2A?si=yd6KfYhbyYJ9Wfg1

Questions

Please write a short response (4-5 sentences / short paragraph) to each of the following questions / the following question. Your responses will be graded for accuracy, critical thinking, and clarity. You may use any common word processing or text format. Please upload your answers by the due date.

Question 1: Why is SFT included in the RLHF pipeline? (i.e., why can't we just start with RL?)

Question 2: What are some potential pitfalls when building and using a reward model? What are the downstream impacts on model alignment?