Final Exam Q2: Essays? In a CS Class?? (50 Points)

16:198:520

Question 1: War Games (30 Points) In the climax of the 1983 AI Thriller 'War Games', Matthew Broderick is desperate to stop an AI from playing a game of global thermonuclear war against the USSR that will have very real consequences. He instructs it to play tic-tac-toe against itself, which it begins to do, ending each game in a draw. Over and over again it plays tic-tac-toe, drawing against itself each time. There's almost a sense of frantic urgency as it plays against itself, faster and faster, trying to find a winning strategy. (The power in the room almost burns out as the computer draws more energy into its calculations.)

A character asks Matthew Broderick what the computer is doing, and he says exictedly, 'It's learning!'

Then, in a moment of almost realization - it starts to play simulated games of global thermonuclear war against itself, as the US and the USSR. It plays through various simulations strategies, noting the lack of winner every time. Faster and faster again, showing various simulated launches and strikes between the two countries, and the final outcome: no winner. It finally shuts down the simulation, and declares in a calm electronic tone, '*A STRANGE GAME. THE ONLY WINNING MOVE IS NOT TO PLAY.*' and it abandons its game of nuclear war.

See the clip here: https://www.youtube.com/watch?v=s93KC4AGKnY

> *Describe what goes on in this scene in terms of algorithms discussed in the class. What is realistic, what is not? What is feasible, based on the kind of algorithms you know to underly AI, and what seems a stretch? Be as thorough and clear as you can be, drawing on as much of the course as you can.*

Question 2 (20 Points): Recently Large Language Models like Chat-GPT have opened up new frontiers in the future of AI, perhaps (as some have speculated) opening the way to AGI if scaling laws persist. However, these language models suffer from a major problem in that they tend to **make stuff up**. They will frequently regurgitate facts and figures in what comes off as a confident tone, regardless of whether what they are saying is actually true. An interesting example of this was Meta's **Galactica**'s tendency to cite papers in its results that did not actually exist.

> Why are Large Language Models prone to making things up in this way? Why don't they know better? Be thorough and draw on our discussions of machine learning. Be wary of anthropomorphizing. **(5 points)**

In a possible alternative future, however, LLMs are not treated as knowledge bases themselves, but instead as *interfaces* between users and knowledge bases (such as Wikipedia, etc), being able to structure and search knowledge bases based on user prompts.

> Describe how you might build and train a system to take a natural language prompt from a user and generate a good google search prompt from it to achieve good, targeted results from google. Be clear about methods, data, representation - all the usual things we discussed in class. **(15 points)**