Reinforcement Learning for Language Model Training

Polina Tsvilodub

Social implications & Limitations of LLMs

RL4 LMT

Intermediate feedback & final projects

Roadmap

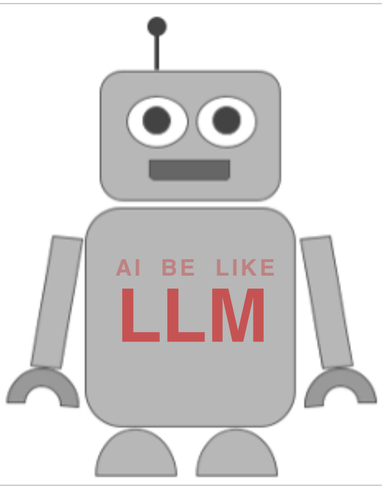
Homework, posters and projects

- **feedback**:
  - 2 tutorial sessions for discussing homework (end of Jan & beginning of Feb; announcement will be out soon)
- **roadmap posters:**
  - see Moodle section on posters: sign up for papers & instructions Google doc available
  - only posters PDFs have to be submitted - no presentation!
  - poster submission deadline: **February 29th 23:59**
- **roadmap final projects:**
  - see Moodle section on projects: requirements & project proposals document available
  - proposed projects are quite research-oriented — potential for projects+ if you are so inclined
  - read, clarify questions, **sign up for groups & projects by class of Jan 24th**: sign up form available on Moodle
  - **project submission deadline: March 31st 23:59**

Two forms of intelligence

or: the LLM cheat sheet

NEITHER OF WHICH ANYONE REALLY FULLY UNDERSTANDS



language (written)

user feedback

statistical regularities between words (tokens)

language (written)

language (written, spoken, signed)

world (the whole gory mess)

abstract, generic, & causal concepts

language (written, spoken, signed)

action (all the fails & glory)

slide credit: Michael Franke

Presentations

Your job

During the presentation, think about the following questions:

- Which factors might affect LLM performance that we would (not) expect to affect human performance?
- What are the limitations of the proposed evaluation paradigm for moral judgement abilities of LLMs?
- What are `good practices` of woking with datasets that you can apply in you own work?