Day-End Demo Summaries

EA Hackathon

Who's noting? patcon, … others?

Demo 1: Alignment Forum Opinion Summarizer

- calling out to openAI API
- reads every approach on alignment forum and reads for keywords
- RLHF approach. other approach ??? not running here because a bit slow
- ask openAI to summarize views
- type in a quest term, that's used to obtain relevant points from ET alignment forum. openAI completion model. gets their position.
- can be integrated into metaculus. want to build a research bot on metaculus that goes out and identifies difficult info and presents it to forecasters
- e.g., you're a general forecaster who doesn't know about alignment, so can go out to forum and get info
- pivot: original UI = question and answer interface. "tell me user opinion based on xxxx". if we see on metaculus, will be question based. ran out of time
- saw some of LLM making things up.
- what would you like to do next with this?
    - making summarizes a bit better toward more narrow questions. less broad overview, and less obvious things
- sam: i don't even know what questions to ask. what questions should i be asking? constant updating thing.
- how does it choose which authors? bringing in posts and know the author who participated. retrieve paragraphs from each post that relevant, and then feed those (with names of author) to LLM. randomly truncating to 5 users.
- lucy: also possible to create aggregate multi-user summary of the summaries? e.g., "there is one camp that believes X, and another that believes Y"
- arvo: summary of technical terms used? LLM is good at that

Demo 2: visualising binary data tuples

- taking binary string of data and visualizing it by running window over data
- #todo link to code?
- look at diff types of files
- text file
- neural net generated photos
- png, jpg, json, txt

- trained vs untrained networks. both look like blobs
- working through layers to compare
- looking for structure in
- q: how does data map to pixels?
    - super simple algo. a couple lines. walk through binary string one at a time and construct all the tuples of consecutive pair.
- Q: what is this good for? based on paper by xxxx
    - is you reverse eng something, you can do it 2 ways. can analyze and recognize patterns if know type of patterns, or can use hex editor and work up from hex. 2nd is best when you don't know what file is about

Demo 3: controlling computer via LLM

- Chidi and Luke
- WARNING: this will broke SO MANY TIMES
- langchain. if trying to get something like GPT3 to do more useful things, you want to process and pass things between models
- an approach to control own computer using machine learning
- two main parts: whisperAI + langchain GPT3 part
- whisperAI = speech to text model from openAI from last october
- demo: "what is 5 time 5?"
    - opens calculator and adds stuff
    - converted text to applescript
- Luke developed langchain model
- Chidi developed whisperAI part
- Chidi built open source tool called boz. audio to text. live transcription. good for subtitles. cool part = all offline. local first. privacy first.
- can work with applescript
- gave lightweight ability to process text on screen, but can't see images yet
- demo: "Play Chess"
- will google things until it thinks it's found an answer
- ...

Demo 4: ITN form wizard

- 3 ppl: adam, lucy, arvo
- https://www.quantifiedintuitions.org/botec
- https://www.guidedtrack.com/programs/34umajb/run
- if one does a shallow investigation, won't share. so won't notice when things haven't been investigated.
- ppl think "someone prob tried this already" so don't bother

- idea = use Squiggle: probabilistic programmin language
- generating squiggle script via small script, and then can edit
- have “database” as EA forum post
- can upvote the most interesting ones
- to help calculate the impact of possible actions
- demo: impact of improving bad elevator music
- …

Demo 5: Stampy.ai

- existing project
- online FAQ about AI safety
- all backed up by google docs
- semantic search for different questions in the db
- we were workin
- originally whole thing lived on discord, but now it’s a bot and a frontend
- stampy will sometimes come into discord and suggest questions based on conversations
- if bot doesn’t have an answer, it can be saved to list, where it will be asked to a real human later in discord (if someone clicks “none of these things” bc a good question doesn’t exist from semantic search)
- readme improvements: https://github.com/StampyAI/stampy/pull/206
-
- …
- …

Demo 6: constrained language model

- scifi inspired: Book of the New Sun by Gene Wolf
- The Just Man chapter: https://gwern.net/doc/culture/1983-wolfe-thecitadeloftheautarch-thejustman
- language that can be straight out of propaganda book
- still a thinking person underneath, but can only speak. can tell story by speaking propaganda out of order
- wanted to constrain AI to only say pre-approved set of sentences
- https://gist.github.com/BorisTheBrave/969f303a082c9da1916d04ee1eb04452
- …