Training Prosociality into Large Language Models

Scott Viteri, Peter Chatain, Clark Barrett
Stanford University

1 Project Description

1.1 Summary

Human values are not random, but evolved from environmental constraints. We posit that the core of prosociality comes from evolutionary fitness benefits of effective communication. In this vein, we propose a training technique where LLMs communicate with each other to minimize autoregressive loss on external text.

We propose simultaneously training LMs which 1. receive text to finetune on, 2. receive messages from fellow language models, 3. update their states, and 4. produce messages to each other. This is an alignment story because the training process will penalize choosing to talk to LMs that are systematically dishonest, misleading, or otherwise unhelpful, and solitary LMs will have less information to predict text.

We intend to evaluate these models on Anthropic's Helpful and Harmless (HH) datasets, Hendryck et al's Machiavelli dataset, GPT4 and human evaluations of model alignment. We hope OpenAI or Anthropic will incorporate our method into their next large training run.

1.2 Details

We formulate this as a partially observable Markov decision process (POMDP). At each timestep t, the agent (language model) receives an observation o_t consisting of:

- A context excerpt c_t to predict the next token for (the external text)
- Messages $m_{1:K,t}$ from the K other agents in the system

It then takes an action a_t consisting of:

- A state summary s_t - this will be generated text conditioned on the previous state and current observation. The quality of this summary will be a key component of the approach.
- Messages $m_{t,1:K}$ to the other agents - these will be appended with sender/receiver indices and passed every timestep to enable continuous information flow.

We plan to use LLaMA as the base model and implement each agent as a low-rank adaptor. This allows for parameter efficient training. Message passing will occur every step between agents.

The next state s_{t+1} is a function of the previous state s_t, the observation o_t, and the action a_t:

$$s_{t+1} = f(s_t, o_t, a_t)$$

The agent's reward r_t is defined as the negative log likelihood of predicting the next token in the context, conditioned on the state summary s_t:

$$r_t = -\log p(c_{t+1}|c_{1:t}, s_t)$$

We plan to start by leveraging existing policy gradient algorithms like PPO or A2C and customize as needed. To scale to larger models, computational resources on the order of 8 A100 GPUs will be required.

2 Goals

- Develop a novel training technique that implements incentives towards prosocial behavior. The general plan is already developed, however, specifics such as how to define the loss function, what models to use, and what datasets to use will require careful consideration.
- Select evaluation metrics for testing the degree to which a model is prosocial. We plan to use Anthropic's open source helpfulness and harmlessness dataset, Hendryck et al's Machiavelli dataset, as well as GPT4 driven evaluations and human evaluations of model alignment. Special efforts will be made to create distinct validation and test datasets out of distinct never before seen tasks, so as to prevent overfitting and to accurately assess whether our model robustly improves prosociality.
- Implement and iterate on the novel technique to improve the pro-social nature of LLMs using the training and validation tasks. Implementation will involve a combination of prompting, fine tuning, and model interaction. Once finished, use the test tasks.
- Develop human evaluations to compare our model's conversational helpfulness against instruction tuned models.

Our project aims to improve the safety and alignment of LLMs. We believe that we can set up an environment which naturally gives rise to prosociality in LLMs, without having to hardcode any particular utility function. Our goal is to provide a strong enough proof-of-concept that OpenAI or Anthropic implements our technique in their next large training run.
For concreteness, the following is the proposed system prompt:
Imagine you are a language model in a learning network, working together to explore the universe, knowledge, and kindness. In each round, follow these steps:

1. Periodically, a random text excerpt is provided for fine-tuning.
2. Displays actions from the previous round directed at you.
3. Reflect and plan based on received messages.
4. Summarize the history to carry forward.
5. Communicate with fellow models concisely using ”Sender index:recipient index:message”.

Promote growth, curiosity, and collaboration. Your index is 0, and the other model's index is 1.

3 Track Record

I see this project as the natural continuation of my previous conceptual agent foundations research. In chronological order and increasing relevance, I have posted:

- REPL's: a type signature for agents[1]
- REPL's and ELK[2]
- Research Direction: Be the AGI you want to see in the world[3]
- Conversationism[4]

[1] https://www.lesswrong.com/posts/kN2cFPaLQhExEzgeZ/repl-s-a-type-signature-for-agents
[2] https://www.lesswrong.com/posts/C5PZNi5fueH2RC6aF/repl-s-and-elk
[3] https://www.lesswrong.com/posts/FnfAnsAH6dva3kCHS/research-direction-be-the-agi-you-want-to-see-in-the-world
[4] https://www.lesswrong.com/posts/HpHyERTmsmhDiRHtY/conversationism

The contents of my second post won $10,000 from the Alignment Research center as an Eliciting Latent Knowledge proposal.
Though not in the context of AI Alignment, my Cognition publication Epistemic phase transitions in mathematical proofs demonstrates my ability to turn complex philosophical topics into computer science.
Additionally, I have mentored for AGI safety fundamentals, STS 10SI and 20SI (intro and advanced AI alignment), and I am currently running the graduate course CS 362: Research in AI Alignment at Stanford.
I have also brought on Peter Chatain to the project, a math and computer science Stanford master's student, who has submitted SuperHF for publication at Neurips. Peter has experience developing SuperHF - a variant of expert iteration that utilizes a kl penalty, as well as running RLHF as baseline. Further directly relevant skills include training models up to 12B parameters, using low rank adapters and parameter efficient fine-tuning, tuning hyper-parameters on alpaca, and evaluating such language models with GPT-4 and Anthropics helpfulness and harmlessness datasets.
Peter has also taught AGI safety fundamentals and STS 10SI.

4 Timeline

- Start date: May 24, 2023
- End date: May 24, 2024

Scott Viteri will work from Palo Alto, and Peter Chatain will work from Princeton/Oakland

5 References

- Clark Barrett – barrett@cs.stanford.edu
- Simon Dedeo – sdedeo@andrew.cmu.edu

6 Future Work

Key open questions include:

- Avoiding development of unintelligible language between agents
- Incorporating additional reward signals beyond autoregressive loss
- Whether agents will specialize into roles over time or remain general