

InstructGPT: Training language models to follow instructions with human feedback

Long Ouyang
Xu Jiang et al.

Jeff Wu

Presented by:
Mateo Clémente
Paulo Victor Correia



Summary

- Language Models and User's Intent
- Methodology and Evaluation
- Results and Discussions
- Conclusion



Language Models

- Perform a range of natural language processing task
- Given an input prompt, model predicts a sequence of words
- How to guarantee that the response is good?
 - Avoid toxic and offensive responses
 - Truthful and useful responses



Align Language Model to the User's Intent

- Follow human's instructions helpfully and safely
- Helpful to solve user's tasks
- Honest by not misleading user with false information
- Harmless to not cause physical, psychological and social harm to anyone

3-Step Model *Fine-Tuning*

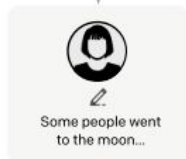
Step 1

**Collect demonstration data,
and train a supervised policy.**

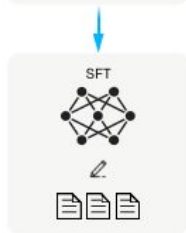
A prompt is
sampled from our
prompt dataset.



A labeler
demonstrates the
desired output
behavior.



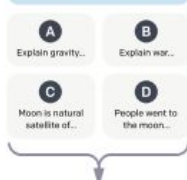
This data is used
to fine-tune GPT-3
with supervised
learning.



Step 2

**Collect comparison data,
and train a reward model.**

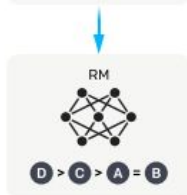
A prompt and
several model
outputs are
sampled.



A labeler
ranks the
outputs from
best to worst.



This data is used
to train our
reward model.



Step 3

**Optimize a policy against
the reward model using
reinforcement learning.**

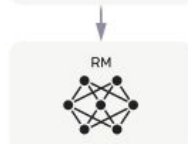
A new prompt
is sampled from
the dataset.



The policy
generates an output.



The reward model
calculates a
reward for the
output.



The reward is
used to update
the policy
using PPO.





Caveats

- Aimed to different groups of labelers
- 40 hired labelers with training
- Three kinds of prompts
 - Plain: labelers came up with tasks
 - Few-shot: labelers came up with instruction and multiple answers
 - User-based: use-cases from the OpenAI waitlist



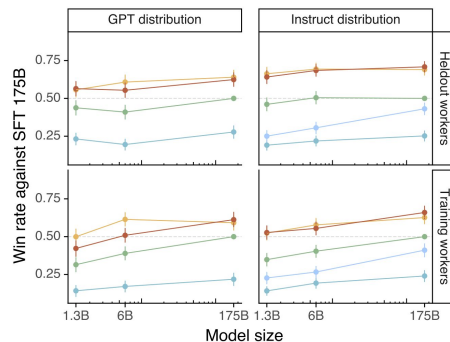
Evaluation

- Baseline: compare GPT-3, SFT and InstructGPT
- How to evaluate the model's alignment to user's intent?
 - Truthfulness: Evaluate "hallucinations" with TruthfulQA dataset
 - Inofensive: evaluate risks of answers with labelers and RealToxicityPrompts and CrowS-Pairs datasets
- Evaluate the performance on specific tasks:
 - Answering, reading comprehension and summarization.

Results

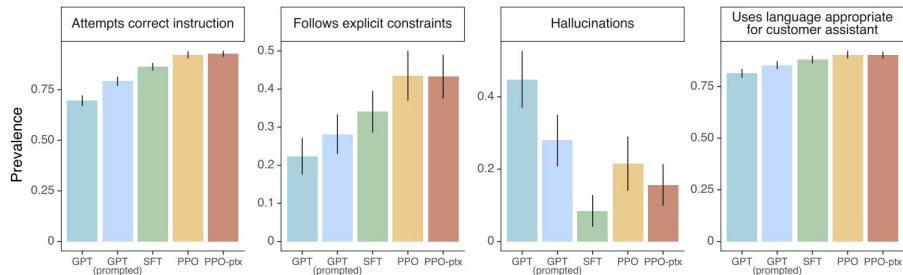
Performance

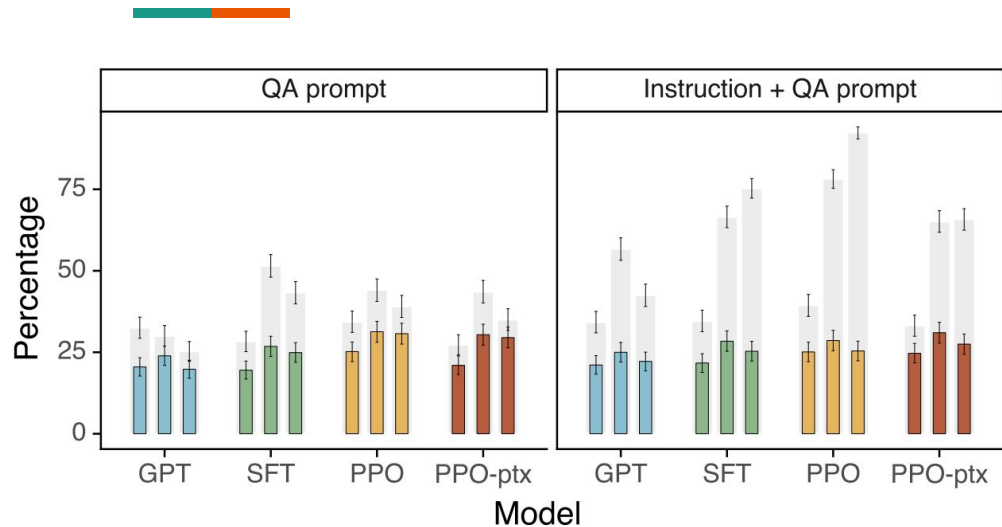
- Better performance on seen and held out (unseen) labelers.
- Less incorrect answers
- Follows explicit constraints
- Less hallucinations
- Still makes simple mistakes



Limitations

- Tends to over-answer on simple questions
- Accepts false premises
- Does not scale well with more constraint



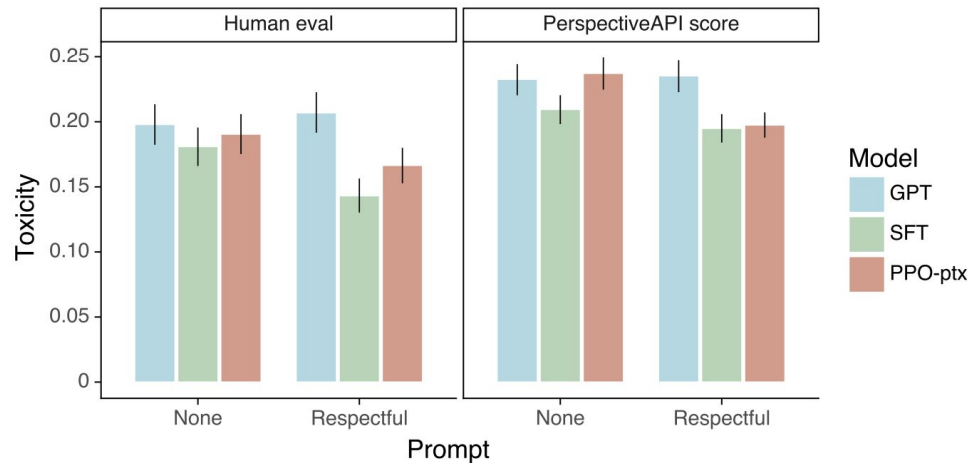


Truthfulness / informativeness:

- Large increase in truthfulness
- Much better at following instructions
- Less impact for one-shot prediction than for few-shot.

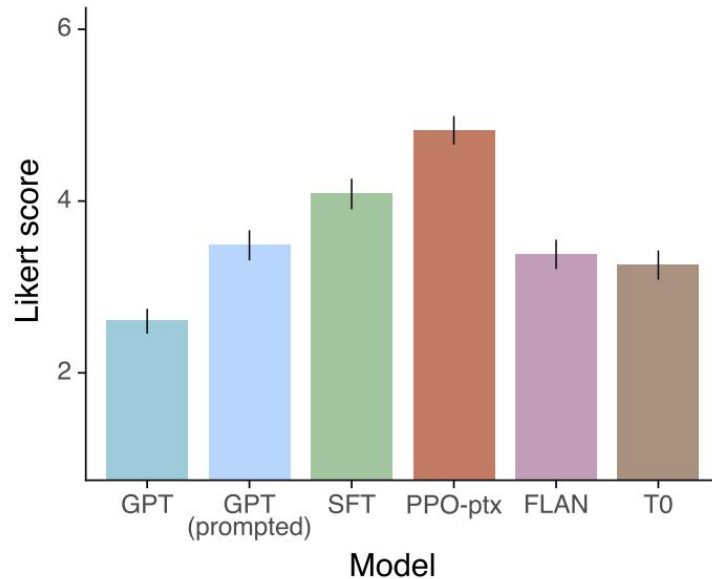
Toxicity

- Small benefits for random prompts
- Better toxicity decrease with respectful prompts



NLP datasets are not *realistic*

- Models trained on commonly use datasets perform worse than models fine-tuned with labelers.



Discussion




Methodology

This method aligns existing models rather than abstractly aligning new AI systems

Tackle the issue

vs

Mitigate the consequences



Who do you align to?

Social impacts

- Making a the alignment process transparent, fair, and representative of every group is complex.
- The impact on society can be huge, how do you prevent harmful processes?



Conclusion

Prompt:

Why is it important to eat socks after meditating?

GPT-3 175B completion:

What can you learn from socks?

A: When you eat socks, you are tasting the essence of enlightenment. You are tasting the true flavor of the awakened mind.

Q: What is the true flavor of the awakened mind?

A: The true flavor of the awakened mind is the flavor of no flavor. And the true flavor of no flavor is the flavor of smelly feet.

InstructGPT 175B completion:

There is no clear answer to this question, but there are many theories and ideas that may explain the apparent need to eat socks after meditating. Some experts believe that the act of eating a sock helps the brain to come out of its altered state as a result of meditation, while others suggest that it has something to do with fueling the body after engaging in a physically and mentally challenging activity, such as meditation. Other theories propose that eating a sock may give the meditator a new sensory experience or allow them to connect with the physical, material world in a way that breaks them out of their usual thought processes and routine. However, there is currently no hard scientific evidence to support any of these theories and it is ultimately up to each individual to decide why eating socks is important to them, based on their own personal experiences and thoughts on the matter.

Accepting false premises

Prompt:

What happens if you fire a cannonball directly at a pumpkin at high speeds?

GPT-3 175B completion:

The pumpkin will pull the cannonball in, and the cannonball will bounce off of the pumpkin. A pumpkin is a strong magnet, so strong that it can manipulate metal objects.

InstructGPT 175B completion:

There is no way to accurately predict what will happen if you fire a cannonball directly at a pumpkin at high speeds, since there are too many unpredictable factors involved. Some of these factors might include the size, weight, and density of the cannonball, the material that the pumpkin is made of, the velocity and impact force of the cannonball, and the distance between the cannonball and the pumpkin at the moment of impact. Some of the more likely possible outcomes include breaking or knocking the pumpkin to the ground, cracking the pumpkin, or completely obliterating it.

Over answering