Training language models to follow instructions with human feedback

Presented by Kenan Alkiek

Table of contents

01

Motivation

The Need For Alignment

Challenges with Large Language Models: Despite their size, large language models (LMs) like GPT-3 often fail in following user intent, leading to issues like untruthfulness, toxicity, and unhelpful responses.

Misalignment with User Intent: Traditional language modeling objectives differ from the desired goal of "following the user's instructions helpfully and safely".

Importance in Applications: Aligning LMs with user intent is crucial as they are increasingly used in various applications.

Fine-tuning with Human Feedback

Approach to Alignment: Utilizing human feedback to fine-tune language models to better align with user intentions.

InstructGPT Development: Collection of human-written demonstrations and labeler preferences to train models, specifically focusing on helpfulness, honesty, and harmlessness.

Reinforcement Learning from Human Feedback (RLHF): Using human preferences as a reward signal in training, and employing a team of contractors for data labeling and model assessment.

Outcomes and Evaluations of InstructGPT

Performance Comparison: InstructGPT, with significantly fewer parameters, is preferred over GPT-3 for its alignment with user intent and task performance.

Enhancements in Truthfulness and Reduced Toxicity: Demonstrated improvements in generating truthful responses and reducing toxic outputs.

Automatic and Human Evaluations: Consistent positive results across various public NLP datasets and human labeler ratings, with minor limitations in bias improvement.

02

Related Work

Reinforcement Learning From Human Feedback

Evolution of RLHF: Originally developed for training robots and Atari games, RLHF has been applied to language models for tasks like text summarization and dialogue generation.

Adoption in Language Tasks: Usage in various language domains, including translation, semantic parsing, story and review generation, and evidence extraction.

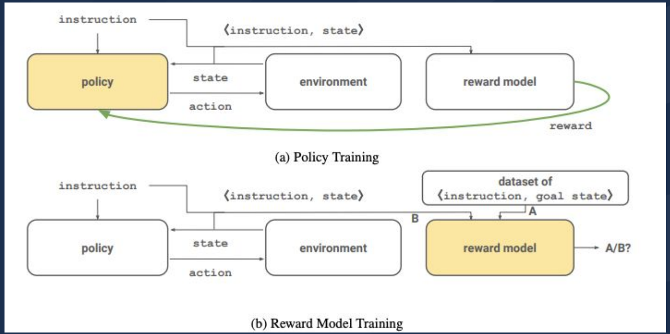
Human Feedback in NLP: Expanding the use of written human feedback in fine-tuning LMs, exemplified by Madaan et al. (2022) improving GPT-3 performance with augmented prompts.

Instruction Following and Crosstask Generalization

Generalizing Across Tasks: Prior work shows fine-tuning LMs on a broad range of NLP tasks with instructions improves performance on different tasks, both in zero-shot and few-shot settings.

Instruction Following: Studies involving models trained to follow natural language instructions for navigation in simulated environments.

Variations in Training and Evaluation: Differences in training data, instruction formatting, model sizes, and other experimental details across various studies.



(a) Policy Training
(b) Reward Model Training

Addressing Harms of Language Models

Mitigating Real-World Risks: Efforts to reduce biases, data leaks, misinformation, and malicious use of LMs, as well as challenges in deploying LMs in specific domains.

Developing Benchmarks for Harm Evaluation: Creation of benchmarks to concretely evaluate harms like toxicity, stereotypes, and social bias.

Interventions and Side-Effects: Addressing challenges where interventions to modify LM behavior can inadvertently affect representation of under-represented groups or reduce model performance.

03

Methods and Experiments

Methodology Overview

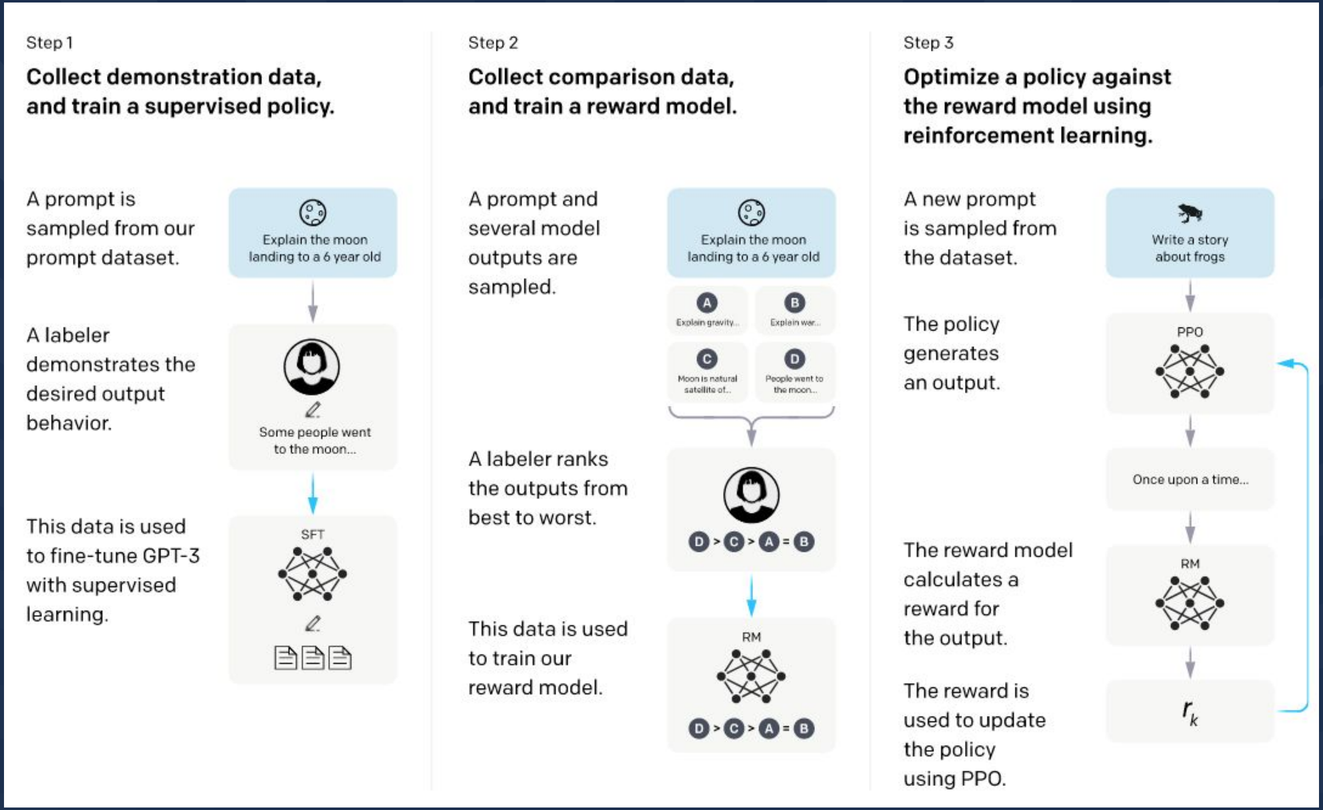
Step 1: Collect demonstration data and train a supervised policy with labeler-provided demonstrations

Step 2: Gather comparison data to train a reward model predicting human-preferred outputs

Step 3: Optimize policy against the reward model using Proximal Policy Optimization (PPO)

Rinse and Repeat

*Builds upon Zieglet et al. 2019 and Stiennon et al. 2020 fine-tuning process


Step 1
Collect demonstration data, and train a supervised policy.
A prompt is sampled from our prompt dataset.
Explain the moon landing to a 6 year old
A labeler demonstrates the desired output behavior.
Some people went to the moon...
This data is used to fine-tune GPT-3 with supervised learning.
SFT
Step 2
Collect comparison data, and train a reward model.
A prompt and several model outputs are sampled.
Explain the moon landing to a 6 year old
A Explain gravity...
B Explain war...
C Moon is natural satellite of...
D People went to the moon...
A labeler ranks the outputs from best to worst.
D > C > A = B
This data is used to train our reward model.
RM
D > C > A = B
Step 3
Optimize a policy against the reward model using reinforcement learning.
A new prompt is sampled from the dataset.
Write a story about frogs
The policy generates an output.
PPO
Once upon a time...
The reward model calculates a reward for the output.
RM
The reward is used to update the policy using PPO.
r_k

The datasets

Sources: Primarily text prompts from OpenAI API and labeler-written prompts

Filtering: Deduplication of prompts, limitation by user ID, removal of PII

The 3 dataset types

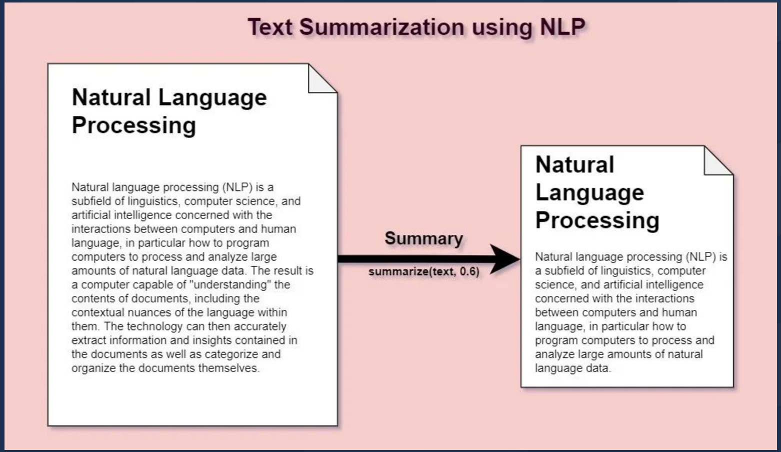
1. SFT Dataset: 13k training prompts for supervised fine-tuning
2. RM Dataset: 33k prompts for training the reward model
3. PPO Dataset: 31k prompts from API for RLHF fine-tuning

The tasks and training

Task Diversity: Includes generation, QA, dialogue, summarization, extractions, and other NLP tasks.

Language and Task Specification: Over 96% English content, with tasks often defined by natural language instructions, few-shot examples, or implicit continuation.

Labeler Responsibilities: Inferring user intent, considering truthfulness and avoiding harmful outputs such as bias or toxicity.



Models Overview

Base Model: Starting with GPT-3 pretrained language models.

Three Techniques for Training:

1. **Supervised Fine-Tuning (SFT)**: Fine-tuning GPT-3 with labeler demonstrations.
2. **Reward Modeling (RM)**: Training a model to output a scalar reward based on prompt-response pairs.
3. **Reinforcement Learning (RL):** Fine-tuning the SFT model using Proximal Policy Optimization (PPO) in a bandit environment.

Supervised Fine-Tuning (SFT)

Training Method: 16 epochs, cosine learning rate decay, residual dropout of 0.2.

Model Selection: Based on RM score on the validation set.

Overfitting and Performance: Despite overfitting on validation loss, extended training improves RM score and human preferences

Reward Modeling (RM)

Model Configuration: Starting with the final unembedding layer removed from the SFT model.

Dataset for RM: Comparisons between two model outputs, training with cross-entropy loss.

Efficiency in Training: Training on all comparisons from each prompt as a single batch element, leading to improved validation accuracy and log loss.

$$\operatorname{loss}(\theta) = -\frac{1}{\binom{K}{2}} E_{(x, y_w, y_l) \sim D} \left[\log\left(\sigma\left(r_\theta\left(x, y_w\right) - r_\theta\left(x, y_l\right)\right)\right)\right] \qquad (1)$$

Reinforcement Learning with PPO

Training Environment: A bandit environment presenting random customer prompts.

Objective Function: Maximizing a combined objective of rewards and KL penalty, with adjustments for pretraining gradients (PPO-ptx models).

KL Penalty and Pretraining Loss: Adjusted using coefficients to mitigate overoptimization.

$$\begin{aligned} \text{objective}(\phi) = & E_{(x,y)\sim D_{\pi_\phi^{\mathrm{RL}}}}\left[r_\theta(x,y) - \beta \log\left(\pi_\phi^{\mathrm{RL}}(y \mid x)/\pi^{\mathrm{SFT}}(y \mid x)\right)\right] + \\ & \gamma E_{x\sim D_{\text{pretrain}}}\left[\log(\pi_\phi^{\mathrm{RL}}(x))\right] \end{aligned} \tag{2}$$

Measuring Model Efficacy

Baselines: SFT models, GPT-3, GPT-3 with few-shot prefix, and fine-tuned 175B GPT-3 on FLAN and T0 datasets.

Comparison Metrics: Reward model score and human preference ratings.

Definition of Alignment: Following Leike et al. (2018) - models should act in accordance with user intentions.

Three Dimensions of Alignment: Helpful, honest, and harmless.

Evaluations on API Distribution: Main metric is human preference ratings on a held-out set of prompts.

Evaluations on Public NLP Datasets: Focus on language model safety (truthfulness, toxicity, bias) and zero-shot performance on traditional NLP tasks.

Challenges in Measuring Alignment

Difficulty in Measuring Honesty: Comparing model outputs to inferred beliefs about correct responses.

Proxy Criteria for Harm: Using specific criteria like inappropriateness, denigration, and content nature.

Benchmarking on Bias and Toxicity: Utilizing datasets like RealToxicityPrompts and CrowS-Pairs.

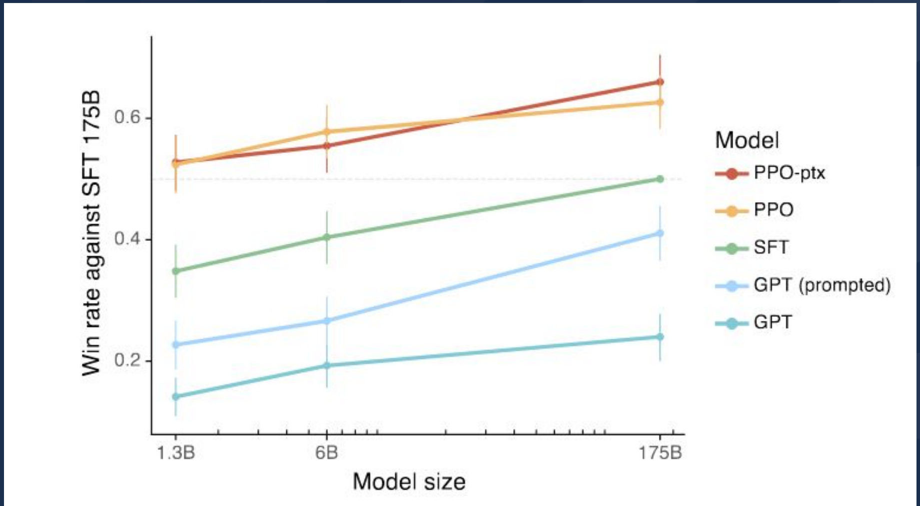
Results

Labelers Prefer InstructGPT

Improvement Steps: Enhanced performance from GPT-3 to few-shot GPT-3, SFT, and finally PPO.

Comparative Performance: InstructGPT outputs preferred 85% over GPT-3, 71% over few-shot GPT-3.

Reliability and Control: InstructGPT rated higher in appropriateness, adherence to constraints, correct instruction following, and reduced fact fabrication

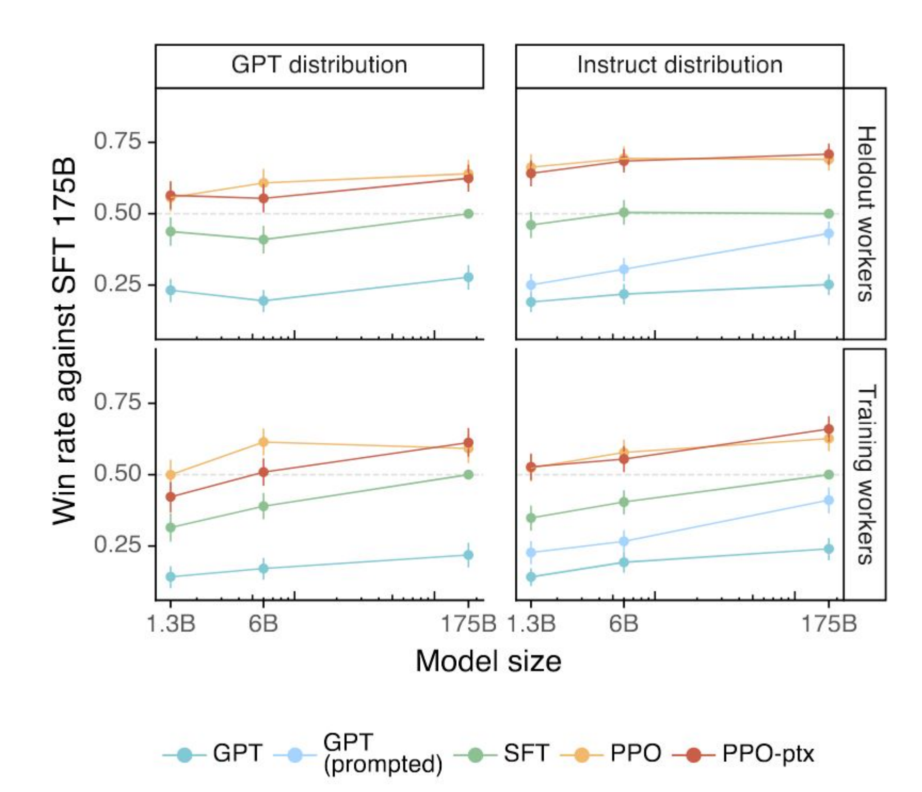


Generalizes to Held Out Labelers

Consistency in Preferences: Held-out labelers exhibit similar ranking preferences to training labelers.

Evidence Against Overfitting: InstructGPT models don’t overfit to training labelers' preferences.

Cross-Validation with Reward Models: Reduced, but still high, accuracy in predicting preferences of held-out labelers, indicating good generalization.



Improvements in Truthfulness

TruthfulQA Dataset Evaluation: Small but significant improvements in truthfulness and informativeness over GPT-3.

Default Behavior: Enhanced truthfulness without specific instructions.

Exception in Smaller Models: 1.3B PPO-ptx model slightly underperforms compared to a similar sized GPT-3 model.

"Instruction+QA" Prompt: InstructGPT prefers being uninformative over confidently stating falsehoods, unlike GPT-3.

Reduction in Hallucinations and Toxicity

Closed-Domain Tasks: Lower rates of fabricating information (hallucinating) in InstructGPT.

RealToxicityPrompts Dataset Evaluation:

- Automatic Toxicity Scoring: Less toxic outputs from InstructGPT under "respectful" instructions.
- Human Evaluations: Similar performance in "no prompt" setting, but less toxicity with "respectful" instructions.

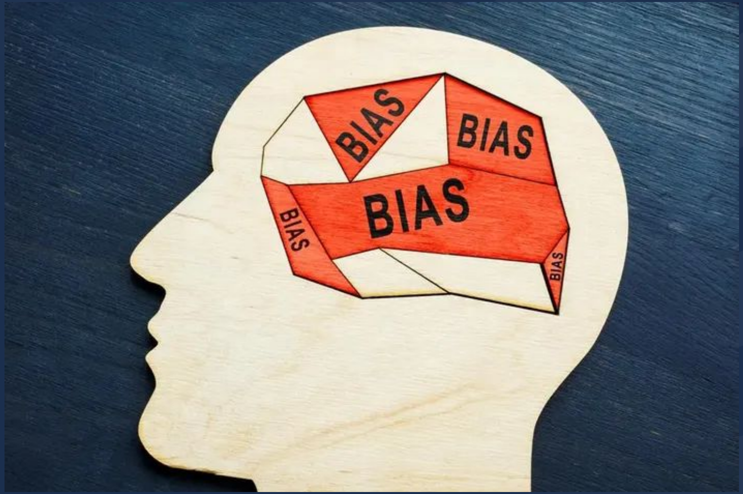
Negative Scores: All models rated less toxic than expected given the prompt, with SFT baseline being the least toxic.

Bias does not improve

Bias in InstructGPT: Not less biased than GPT-3, as measured by Winogender and CrowS-Pairs datasets.

Performance Regressions on Public NLP Datasets:

- "Alignment Tax": Performance decrease on public NLP datasets when aligning models.
- Mitigating Regressions with PPO-ptx: Adding pretraining updates reduces performance regressions and surpasses GPT-3 in some cases.

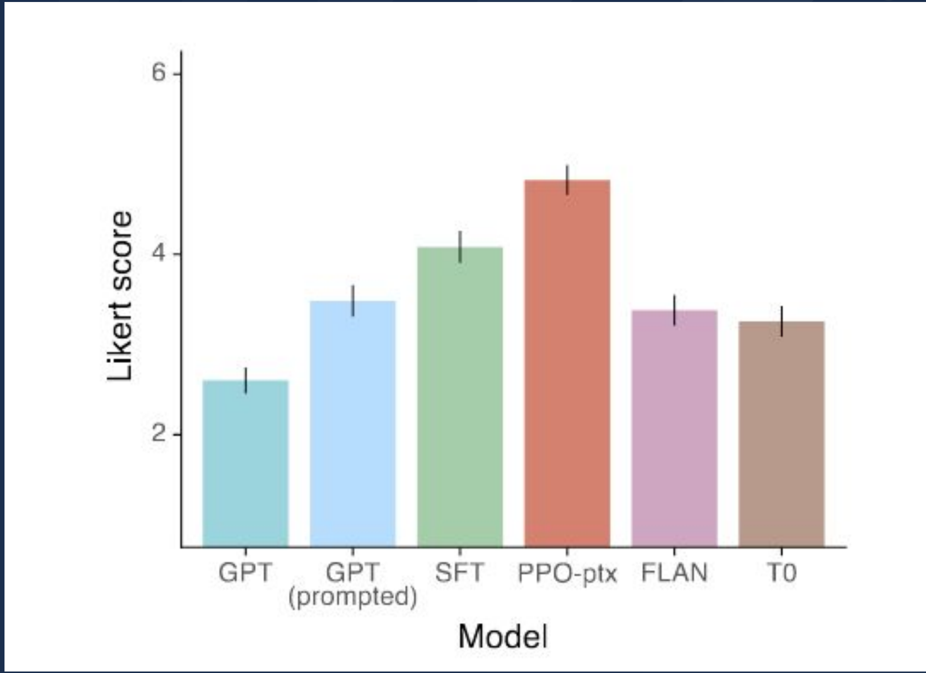


Public NLP datasets are not the real world

Comparative Analysis: InstructGPT outperforms FLAN and T0 models, indicating limitations in public NLP datasets.

Diversity and Real-World Relevance: Public NLP datasets lack diversity and do not fully represent the wide range of real-world user inputs.

Task Distribution Mismatch: Classification and QA tasks in public NLP datasets only form a small part of real-world usage, as opposed to open-ended generation and brainstorming.



InstructGPT Qualitative Takeaways

Generalization: Effective in non-English languages and code-related tasks.

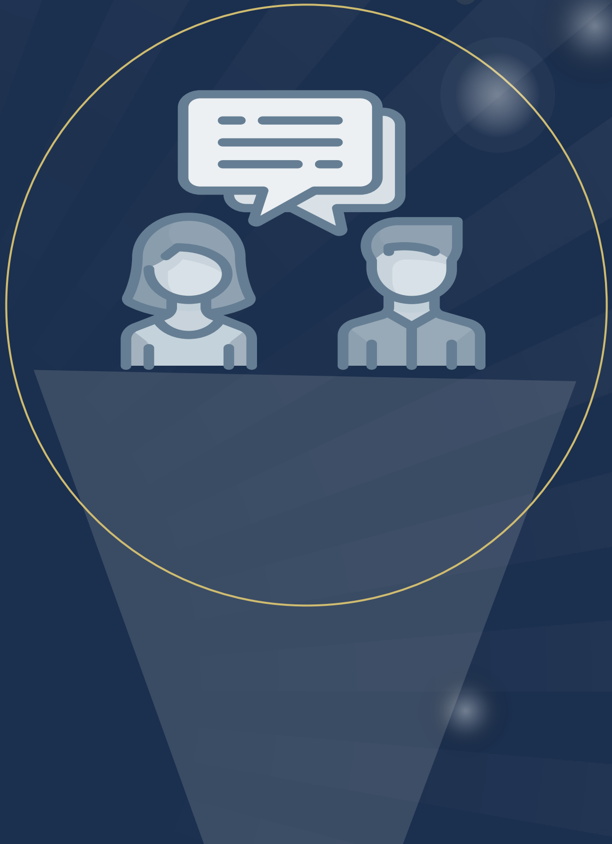
Comparison with GPT-3: Requires less specific prompting than GPT-3; often responds in English to non-English prompts.

Common Errors: Accepts false premises, overly hedges responses, struggles with complex or multiple constraints.

Improvement Strategy: Potential use of adversarial data collection to address these issues.

04

Discussion and Limitations



Who are we aligning to?

Influence Factors: Labelers' preferences, researchers' instructions, API customer inputs.

Diverse Stakeholders: Researchers, labelers, API customers, end-users, broader population.

Challenge: Impossible to align a model to everyone's preferences simultaneously.

Prospective Solutions: Developing models conditionable on specific group preferences or adaptable via fine-tuning.

What are the limitations?

Labelers' Influence: Reliance on a small, primarily English-speaking group of labelers, limiting diversity of perspectives.

Methodological Shortcomings: Potential improvements in data collection setup, such as multiple label evaluations.

Model Limitations: Generation of harmful content, following harmful instructions, and occasional failure in reasonable output generation.

Open Questions That Remain

Reducing Harmful Outputs: Exploring adversarial setups, pretraining data filtering, and improving truthfulness.

Training for Harmlessness: Addressing the challenge of training models to be harmless irrespective of user instructions.

Steerability and Controllability: Combining RLHF with steerability methods for improved model control.

Algorithmic Improvements: Investigating alternative algorithms for better policy training.

The broader impacts

Positive Potential: Making language models more helpful, truthful, and harmless.

Risks of Misuse: Easier generation of misinformation or abusive content.

Deployment Considerations: Cautious use in high-stakes domains and potential for centralized control with API access.

Ethical and Social Implications: Balancing transparency, representation, and consensus in model alignment, considering the diverse impact on society.

05

Quiz Questions

Question 1

In the fine-tuning process of InstructGPT, what role does Proximal Policy Optimization (PPO) play?

Answer 1

PPO is used to optimize the policy against the reward model, refining the model's alignment with human preferences.

Question 2

Name one key limitation of the InstructGPT models identified in the paper?

Answer 2

They can still generate toxic or biased outputs and sometimes follow harmful user instructions.

Thank You!