Executive Summary

Training Language Models to Follow Instructions with Human Feedback

The Problem

Large language models like GPT-3 often generate outputs that are untruthful, toxic, or simply don’t follow user instructions. These models are trained to predict the next word on internet text, which is fundamentally different from the objective of “follow the user’s instructions helpfully and safely.” This misalignment means that bigger models don’t inherently make them better at following user intent.

The Breakthrough

Researchers developed **InstructGPT**, a method that uses **reinforcement learning from human feedback (RLHF)** to align language models with user intentions. The approach fine-tunes GPT-3 models using human demonstrations and preferences, resulting in models that are significantly better at following instructions while being much smaller. Remarkably, a 1.3B parameter InstructGPT model outperforms the 175B GPT-3 model on instruction-following tasks despite having 100x fewer parameters.

How It Works

The three-step process starts with supervised fine-tuning on human demonstrations of desired behavior. Then, a reward model is trained on human comparisons between different model outputs. Finally, this reward model guides reinforcement learning optimization using PPO. The result is models that achieve **85% preference rate** over GPT-3 outputs and show **2x improvement** in truthfulness while reducing hallucinations from 41% to 21% on closed-domain tasks.

Why This Matters

This technique makes language models more helpful, honest, and harmless - three critical properties for safe AI deployment. The models generalize better to instructions in other languages and code-related tasks, even with minimal training data in those areas. Importantly, the approach

minimizes performance regressions on traditional NLP tasks, addressing the “alignment tax” problem.

The Business Opportunity

Human-aligned language models create new possibilities for customer service, content creation, education, and enterprise applications where reliability and safety are paramount. The method’s cost-effectiveness - requiring significantly less compute than training larger models - suggests a path toward more capable and trustworthy AI systems that can be deployed at scale with reduced risks.