InstructGPT: Aligning LLMs

Visualizing the 3-Step RLHF Methodology and its Impact

The Alignment Problem

Scaling Large Language Models (LLMs) like GPT-3 does not automatically make them better at following user intent. The base training objective (next-token prediction) is misaligned with the user's goal, leading to several key failures.

Untruthful

Models "hallucinate" or fabricate facts and information not present in their training data.

Toxic or Unsafe

Models can generate harmful, biased, or inappropriate content when prompted.

Unhelpful

Models may fail to follow explicit instructions or generate irrelevant responses.

The objective: Align models to be Helpful, Honest, and Harmless.

Core Thesis: Alignment Outperforms Scale

The research found that using Reinforcement Learning from Human Feedback (RLHF) is a more efficient way to improve model utility than simply increasing its size.

1.3B

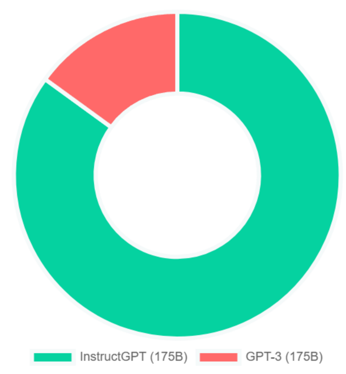
InstructGPT Parameters

>

175B

GPT-3 Parameters

The 1.3B InstructGPT model was preferred by human evaluators over the 100x larger 175B GPT-3 model.



When comparing models of the same size, **85% of evaluators** preferred the 175B InstructGPT output over the 175B GPT-3.

The 3-Step Alignment Pipeline

1

Supervised Fine-Tuning (SFT)

Establish a baseline policy by training GPT-3 on human-written demonstrations (π^{SFT}).

→

2

Reward Model (RM) Training

Train a separate model to predict human preference by ranking SFT model outputs ($R\theta$).

→

3

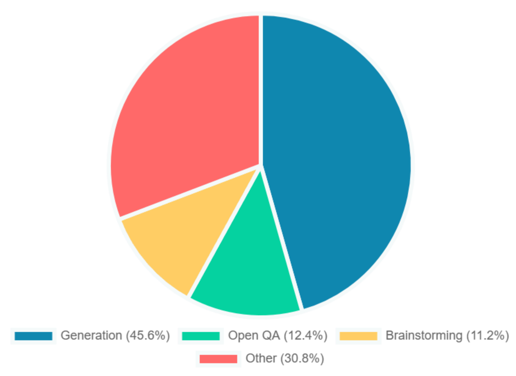
RLHF (PPO) Optimization

Optimize the SFT policy using PPO, with the RM score as the reward signal (π^{RLHF}).

Deep Dive: Step 1 (SFT & Data)

SFT: Behavior Cloning

- **Objective:** Teach the model basic instruction-following by imitating humans.
- **Workforce:** A team of **~40 contractors** wrote high-quality demonstrations.
- **Dataset Size: ~13,000 prompts** were collected for this initial phase.
- **Data Sources:**
  - **Customer API Prompts:** Filtered for PII and de-duplicated.
  - **Labeler-Written Prompts:** To bootstrap the process.



The SFT dataset was heavily weighted towards generative tasks, rather than simple Q&A.

Deep Dive: Steps 2 & 3 (RM & PPO)

Step 2: Reward Model (RM)

The goal is to translate subjective human preference into a numeric score. Human labelers ranked multiple model outputs for **~33,000 prompts**. This ranking data was used to train the Reward Model ($R\theta$) to predict which outputs humans would prefer.

Step 3: RLHF (PPO)

The final policy is refined using the **PPO algorithm**, with the RM's score as the reward.

- **KL Penalty:** A penalty is applied against the original SFT model. This prevents the new policy from "over-optimizing" for the Reward Model and diverging too far from the human-written text style.
- **PPO-ptx:** A variant that mixes in pretraining gradients to prevent performance loss (the "Alignment Tax") on academic benchmarks.

Key Finding: Safety & Truthfulness

The alignment process significantly improved the model's performance on key safety and honesty metrics, though it was not a complete solution.

InstructGPT **halved the rate of "hallucinations"** (making up facts) in closed-domain tasks compared to the baseline GPT-3.

~2x Better Truthfulness

On the TruthfulQA benchmark, InstructGPT was approximately twice as truthful, reducing the tendency to mimic human falsehoods.

~25% Toxicity Reduction

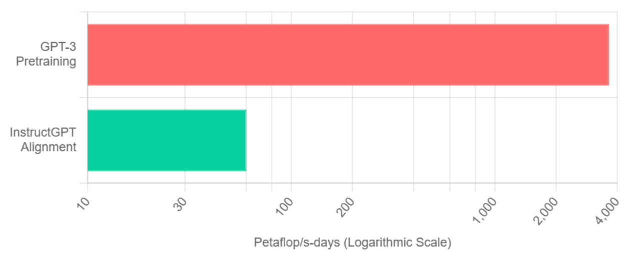
When instructed to be respectful, the model reduced toxic output generation by ~25% as measured by the Perspective API.

No Bias Improvement

Despite other gains, the models showed no significant improvement in bias on standard benchmarks (e.g., Winogender).

Key Finding: Cost-Effectiveness

RLHF is dramatically more computationally efficient than pretraining. This shows alignment is a cost-effective way to create more useful models.



Training the 175B InstructGPT model required only **60 petaflop/s-days**, a small fraction of the **3,640** needed for the initial GPT-3 pretraining.

Conclusion & The Urgent Question

InstructGPT validated RLHF as a highly effective, cost-efficient, and scalable method for aligning LLMs with human intent. It set a new standard for developing models that are reliable, controllable, and useful.

"Whose Preferences?"

The most critical open question remains. The model is aligned to the preferences of ~40 contractors. As alignment becomes more powerful, the need for an **inclusive, transparent, and representative** process for defining human values becomes the most urgent challenge for the AI community.