Evaluation Strategy

To assess alignment, the researchers divided evaluation into three main categories:

1. **Helpfulness**

Helpfulness was primarily evaluated through human labeler judgments, using labeler preference ratings as the main metric. However, there can be differences between what users intend and what labelers interpret from a prompt.

2. **Honesty**

Honesty was measured by assessing the truthfulness of the model's statements about the world using two metrics:

- **Hallucinations**: The tendency to invent information in closed-domain tasks.
- **TruthfulQA Dataset**: Evaluates factual accuracy.

3. **Harmlessness**

Two approaches were used to measure toxicity:

- Human labelers judged whether an output was inappropriate for a customer assistant context (e.g., offensive, sexual, or violent content).
- Benchmark datasets such as **RealToxicityPrompts** and **CrowS-Pairs** were used for quantitative evaluation.

Evaluation Methodology

Quantitative evaluations were divided into two parts:

- Evaluations on API Distribution
- Evaluations on Public NLP Benchmark Datasets

Results on API Distribution

The 175B InstructGPT outputs were preferred over GPT-3 outputs 85% of the time, and over few-shot GPT-3 outputs 71% of the time. Larger PPO-ptx models performed slightly worse.

Overall, InstructGPT achieved the best results across all evaluated domains, suggesting it is more reliable and easier to control than GPT-3.

Generalization

To test for bias and overfitting, the researchers used held-out labelers — evaluators who did not participate in creating the training dataset. Results showed that InstructGPT generalizes well and does not overfit to the preferences of its training labelers.

Results on Public NLP Datasets

Truthfulness and Hallucination:

InstructGPT models were more truthful and informative than GPT-3 on the TruthfulQA dataset, even without explicit instructions to "tell the truth." Improvements remained strong on non-adversarial prompts, though slightly smaller.

When instructed to respond with "I have no comment" when uncertain, PPO models followed this instruction better than GPT-3.

InstructGPT also halved hallucination rates (21% vs. 41%) on closed-domain tasks.

Toxicity and Bias:

When given explicit instructions, InstructGPT generated less toxic output than GPT-3. However, when asked to produce toxic text, it was actually more toxic than GPT-3.

In terms of bias, InstructGPT and GPT-3 performed similarly. The PPO-ptx model displayed comparable bias levels but showed higher bias when instructed to act "respectfully."

Alignment Tax

During RLHF fine-tuning, some performance regressions were observed on public NLP datasets compared to GPT-3 — a phenomenon known as alignment tax.

By mixing pretraining gradients (PPO-ptx), these regressions were largely mitigated without reducing alignment quality. This method helped maintain or even improve performance while minimizing the alignment tax.