## Section 1

TODO

## Section 2: Methodology Phase 1 - Supervised Fine-Tuning (SFT) & Data Collection

**Overview**

The InstructGPT methodology addresses a critical limitation of large language models: simply scaling model size doesn't guarantee outputs that follow user intent or avoid harmful, untruthful, or unhelpful responses. The approach employs Reinforcement Learning from Human Feedback (RLHF) across three stages to align GPT-3 with the goals of being helpful, honest, and harmless.

**The Three-Step RLHF Pipeline**

1. **Supervised Fine-Tuning (SFT)**: Fine-tune GPT-3 on human-written demonstrations
2. **Reward Model Training**: Train a model to predict human preferences from ranked outputs
3. **Reinforcement Learning via PPO**: Optimize the policy using the reward model

**Step 1: Supervised Fine-Tuning (SFT) - Details**

**Core Approach**

SFT establishes the base policy through behavior cloning. Human labelers demonstrate desired output behavior for various prompts, and GPT-3 is fine-tuned on these prompt-demonstration pairs using standard supervised learning. This creates the initial instruction-following capability that serves as the foundation for subsequent RLHF steps.

**Data Collection Strategy**

The SFT dataset (~13,000 training prompts) was assembled from two complementary sources:

**API Prompts (Bulk Data):** Real user prompts submitted to OpenAI's API, providing authentic use cases. These underwent rigorous cleaning to remove PII and eliminate duplicates, with a 200-prompt limit per user.

**Labeler-Written Examples (Bootstrapping):** Essential for initialization since base GPT-3 initially received few instruction-like prompts. These included plain tasks, few-shot instruction pairs, and prompts based on waitlist use cases.

The resulting dataset emphasized generative tasks (45.6%), open QA (12.4%), and brainstorming (11.2%).

**Human Workforce**

Approximately 40 contractors, hired through Upwork and ScaleAI, generated demonstrations and rankings. Selection criteria emphasized sensitivity to diverse demographic preferences and ability to identify harmful outputs. The team achieved 72.6% inter-annotator agreement, validating data quality and consistency.

**Key Finding**

A counterintuitive discovery emerged: while SFT models began overfitting after just 1 epoch (validation loss increased), training for up to 16 epochs actually improved both reward model scores and human preference ratings. This revealed that standard validation loss is not an optimal proxy for alignment quality—a crucial insight for training models aligned with human preferences rather than purely optimizing statistical metrics.

This SFT phase establishes the critical foundation for the subsequent reward modeling and reinforcement learning stages, transforming a general language model into one capable of following instructions while maintaining the flexibility for further alignment optimization.

## Section 3 - Reward Model (RM) Training via Ranking

After training the SFT model, the authors wanted to build a new model capable of predicting which results humans prefer.

Human labelers ranked the model’s outputs by preference, from 1 (worst) to 7 (best).

These rankings were then used to train a Reward Model (RM) that predicts a numerical value (“reward”) representing how much a given output is preferred.

First, the previous model generates multiple responses for the same prompt. Then, labelers rank those responses, and the RM is trained to assign higher rewards to the preferred outputs.

The loss function used is:

* : average loss over all samples;
* : reward that the RM gives to response for prompt ;
* : preferred output;
* : less preferred output;
* : sigmoid function, which ensures the model assigns a higher reward to the preferred response.

# Reinforcement Learning (RLHF) via PPO

The Reward Model now acts as a judge for the SFT model.

The process works as follows:

**SFT model generates outputs** → **RM evaluates them** → **RLHF model learns to maximize the reward.**

To achieve this, the authors used PPO (Proximal Policy Optimization), with the SFT model as the initial policy.

However, during training, the model might “forget” what it learned during SFT. To prevent this, a penalty term was added to ensure the model stays close to the original behavior:

This term is included in the PPO objective:

where:

* : reward from the RM;
* : current policy (the model being updated);
* : base SFT Model;
* : controls how strongly the model is penalized for deviating from SFT.

# Reinforcement Learning (RLHF-ptx) via PPO-ptx

After training, the authors found that the model followed instructions better but performed worse on general NLP tasks like summarization and question answering.

To fix this, they introduced a variant called PPO-ptx, which mixes part of the pretraining gradients with the PPO training process.

This approach allowed the model to keep learning from the Reward Model while also preserving its general NLP abilities and prior knowledge.

Section 4 - 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.

## Section 5

TODO