# 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.