InstructGPT Paper Presentation

Work Distribution Proposal

Student 1: The Motivator

Section 1. Introduction: The Misalignment Problem and InstructGPT Overview

Core Content Focus

This student must establish the necessity of the InstructGPT project. The focus is on the failure of large, traditionally trained language models (LLMs) like GPT-3, which are often *misaligned* with user intent, generating untruthful, toxic, or irrelevant outputs. The alignment goal is framed by the criteria of being **helpful, honest, and harmless (HOH)**. The student must introduce **InstructGPT** and highlight its main, counter-intuitive finding: alignment is often more effective than scale, demonstrated by a small InstructGPT model significantly outperforming the much larger GPT-3.

Student Deliverables: 1-Page Synthesis

A one-page text summarizing the motivation behind aligning LLMs, defining the HOH criteria, providing a high-level overview of Reinforcement Learning from Human Feedback (RLHF), and presenting the striking result comparing the 1.3B InstructGPT model to the 175B GPT-3 model.

Student Deliverables: Suggested Slides

Slide Number	Suggested Title
Slide 1	Title Slide: Training LMs to Follow Instructions with Human Feedback
Slide 2	The Problem: LLMs and Misalignment
Slide 3	Defining Alignment: The H-O-H Framework
Slide 4	InstructGPT: Solution via RLHF (High Level)
Slide 5	Key Insight: Alignment Trumps Scale (1.3B vs 175B)

Supporting Source Quotes

> Large language models (LLMs) such as GPT-3 often fail to follow user intent, producing untruthful, toxic, or irrelevant outputs. This paper presents **InstructGPT**, a family of models fine-tuned from GPT-3 using **reinforcement learning from human feedback (RLHF)** to align model behavior with human intent.
>
> Traditional language modeling trains models to predict the next token in web text, an objective **misaligned** with user goals such as following instructions or avoiding harmful behavior. Simply scaling

> models increases fluency but not alignment. The authors aim to train models that better follow **explicit** and **implicit** user intentions — being **helpful, honest, and harmless**.
>
> In human evaluations on our prompt distribution, outputs from the **1.3B parameter InstructGPT model are preferred to outputs from the 175B GPT-3**, despite having 100x fewer parameters.

Student 2: The Data Designer

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

Core Content Focus

This section focuses exclusively on the initial step of the three-stage methodology: **Supervised Fine-Tuning (SFT)**. The student must explain how data was sourced, emphasizing the role of the trained human labelers (about 40 contractors) and the origin of prompts (OpenAI API usage and labeler-written demonstrations). The diversity and volume of this initial training data (e.g., ~13k SFT prompts) should be highlighted.

Student Deliverables: 1-Page Synthesis

A concise explanation of Step 1 (SFT), detailing the necessity of human demonstrations, the process of recruiting and screening the ~40 contractors, and the distinct sources and categories of the initial prompt data.

Student Deliverables: Suggested Slides

Slide Number	Suggested Title
Slide 6	Methodology Overview (3 Steps)
Slide 7	Step 1: Supervised Fine-Tuning (SFT)
Slide 8	Data Collection: API Prompts and Labeler Demonstrations
Slide 9	The Human Labelers and Data Volume (~40 contractors, 13k SFT prompts)

Supporting Source Quotes

> The InstructGPT pipeline has **three main stages**... 1. **Supervised Fine-Tuning (SFT):** Human labelers write prompts and demonstrate desired responses. GPT-3 is fine-tuned on these demonstrations to create a supervised policy.
>
> Prompts were drawn from early **OpenAI API** usage and labeler-written examples.
>
> About **13k prompts** were used for supervised fine-tuning.
>
> To produce our demonstration and comparison data, and to conduct our main evaluations, we hired a team of about **40 contractors** on Upwork and through ScaleAI.

Student 3: The Architect

Section 3. Methodology Phases 2 & 3: Reward Modeling (RM) and RLHF (PPO)

Core Content Focus

This section details the two advanced, technical phases: **Reward Model (RM) Training** (Step 2) and **Reinforcement Learning (RLHF) via PPO** (Step 3). The student must explain how human ranking of model outputs is converted into a scalar reward function (RM). Subsequently, the student details how the PPO algorithm is used to fine-tune the SFT model based on the RM's feedback. Key technical concepts like the **KL penalty** (to prevent policy deviation) and the **PPO-ptx** variant (mixing pretraining gradients) must be explained clearly.

Student Deliverables: 1-Page Synthesis

A synthesis focusing on the mechanism of the feedback loop: how rankings lead to the RM (Step 2), and how PPO utilizes the RM output as a reward signal (Step 3). Clear definitions of PPO, KL penalty, and the purpose of the PPO-ptx gradient mixing should be included.

Student Deliverables: Suggested Slides

Slide Number	Suggested Title
Slide 10	Step 2: Reward Model (RM) Training via Ranking
Slide 11	The Reward Model Loss Function (Learning Human Preference)
Slide 12	Step 3: Reinforcement Learning (Proximal Policy Optimization - PPO)
Slide 13	The PPO Objective: KL Penalty & PPO-ptx

Supporting Source Quotes

> 2. **Reward Model (RM) Training:** Labelers compare several candidate model responses per prompt. A reward model is trained to predict which outputs humans prefer.
>
> The loss function for the reward model is:
>
> $$\text{loss}(\theta) = -\frac{1}{\binom{K}{2}}\mathbb{E}_{(x,y_w,y_l)\sim D}\left[\log\left(\sigma\left(r_\theta(x,y_w) - r_\theta(x,y_l)\right)\right)\right]$$
>
> where y_w is the preferred completion.
>
> 3. **Reinforcement Learning (RLHF) via PPO:** The SFT model is fine-tuned using **Proximal Policy Optimization (PPO)** to maximize the RM's reward signal. A **KL penalty** ensures the model stays close to the original GPT-3 distribution.
>
> A variant, **PPO-ptx**, mixes in pretraining gradients to preserve general NLP performance.

Student 4: The Results Analyst

Section 4. Key Findings: Performance Metrics and Alignment Tax Mitigation

Core Content Focus

This section details the quantitative improvements achieved by InstructGPT. Key results include the strong **human preference** for InstructGPT outputs (e.g., 175B InstructGPT preferred 85% of the time over 175B GPT-3), and the improvements in **truthfulness** (twice as truthful on TruthfulQA) and **safety** (halved hallucination rates, 25% less toxicity when prompted respectfully). The student must also explain the **"alignment tax"**—performance regressions on standard NLP benchmarks (SQuAD, DROP)—and how the **PPO-ptx** variant successfully minimized this tax.

Student Deliverables: 1-Page Synthesis

A summary of the core metrics: human preference results (with statistics), improvements in honesty (truthfulness/hallucinations), the persistent lack of improvement in raw bias metrics, and a thorough explanation of the alignment tax problem and its effective mitigation via PPO-ptx.

Student Deliverables: Suggested Slides

Slide Number	Suggested Title
Slide 14	Result: Human Preference (Overwhelming Win Rates)
Slide 15	Result: Truthfulness and Hallucination Reduction
Slide 16	Result: Toxicity and Bias (Mixed Safety Gains)
Slide 17	The Alignment Tax: Trade-offs on NLP Benchmarks
Slide 18	Mitigating the Tax: PPO-ptx Variant Success

Supporting Source Quotes

> InstructGPT outputs are significantly preferred over GPT-3, even with 100x fewer parameters. The 1.3B InstructGPT model is preferred over the 175B GPT-3 model. 175B InstructGPT is preferred over 175B GPT-3 **85%** of the time.
>
> InstructGPT roughly **halved hallucination rates** (21% vs. 41% on closed-domain tasks). It performed **twice as truthfully** as GPT-3 on the TruthfulQA benchmark.
>
> During RLHF fine-tuning, we observe performance regressions compared to GPT-3 on certain public NLP datasets... This is an example of an **“alignment tax”**.
>
> Mixing pretraining gradients (**PPO-ptx**) largely removed these regressions without reducing alignment quality.

Student 5: The Strategist/Critic

Section 5. Limitations, Generalization, and Broader Implications

Core Content Focus

This final section covers the model's performance boundaries and the societal context of the research. Topics include the successful **generalization** of instruction-following behavior to novel tasks like code summarization and non-English instructions; remaining limitations (e.g., following instructions with **false premises** or excessive **hedging**); a critical discussion on **labeler bias** (i.e., whose preferences are encoded, typically **~40 English-speaking contractors** and OpenAI researchers); and the **cost-effectiveness** of RLHF compared to pretraining.

Student Deliverables: 1-Page Synthesis

A critical synthesis addressing the successful generalization found in the model, detailing the specific failures or simple mistakes observed, questioning whose values the model is truly aligned to, and summarizing the broader implications for cost, scalability, and future alignment research.

Student Deliverables: Suggested Slides

Slide Number	Suggested Title
Slide 19	Generalization: Non-English Instructions and Code
Slide 20	Persistent Limitations (False Premises & Hedging)
Slide 21	Ethical Critique: Whose Preferences are Encoded?
Slide 22	Broader Impact: Cost-Effectiveness and Scalability
Slide 23	Conclusion and Future Alignment Directions

Supporting Source Quotes

> Models generalized to **held-out labelers** and **novel tasks**, such as code summarization and non-English instructions.
>
> **Limitations:** Models can still fabricate facts, hedge excessively, or obey harmful instructions. When given an instruction with a **false premise**, the model sometimes incorrectly assumes the premise is true.
>
> Alignment reflects the preferences of **~40 English-speaking contractors** and OpenAI researchers, not a global user base.
>
> The cost of increasing model alignment is modest relative to pretraining. RLHF is **computationally inexpensive** compared to pretraining (60 vs. 3,640 petaflop/s-days for GPT-3).
>
> Full alignment requires addressing **whose preferences** are encoded and how to represent diverse value systems.