

Projeto 1 – Metodologias em Modelos de Linguagem

This Summary Report aims to review
the following article:

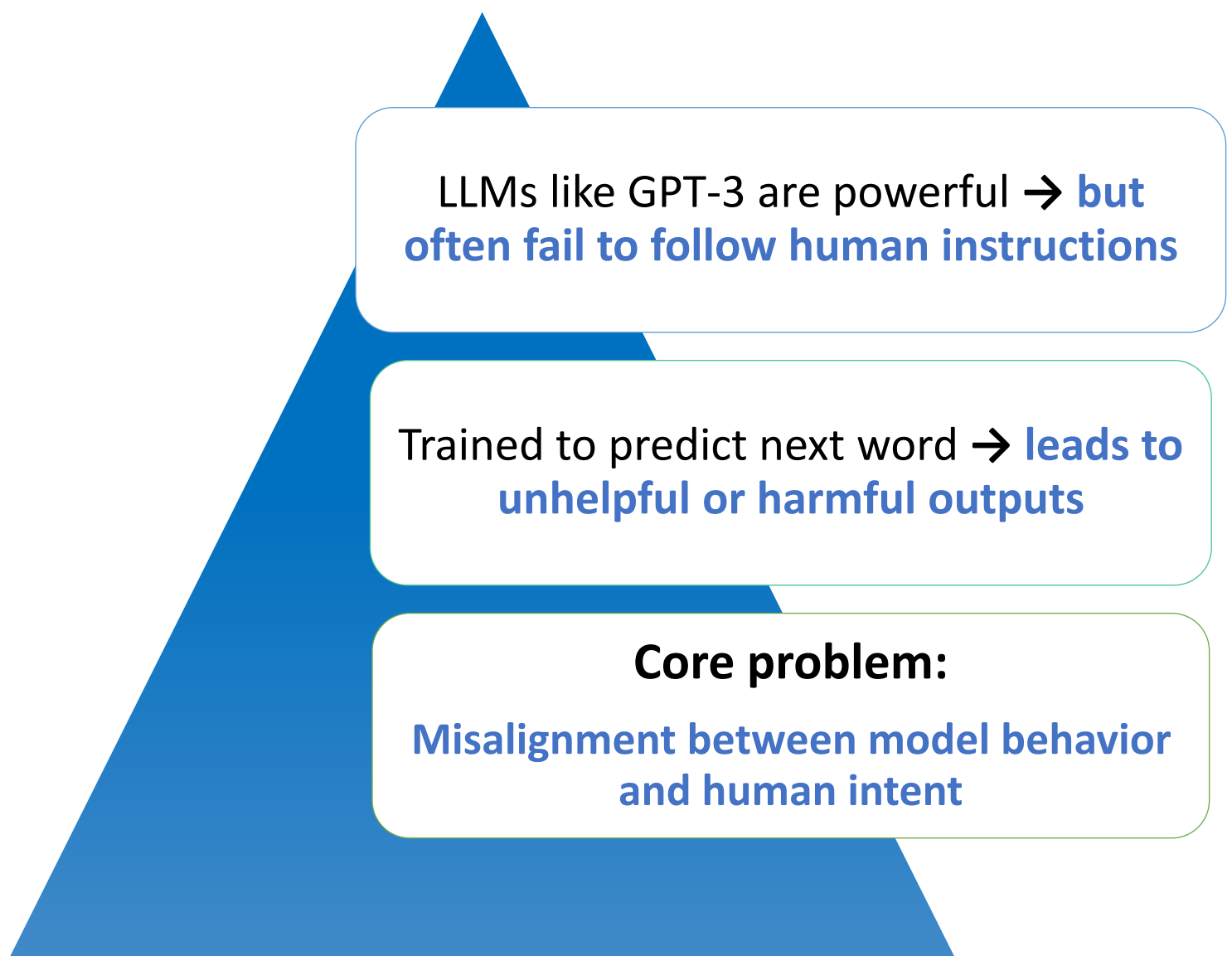
<https://arxiv.org/abs/2203.02155>



Section 1

Objectives of Research

The Problem: Misalignment in LLMs



LLMs like GPT-3 are powerful → **but often fail to follow human instructions**

Trained to predict next word → **leads to unhelpful or harmful outputs**

Core problem:

Misalignment between model behavior and human intent

Motivation for the Study

The authors propose → Reinforcement Learning from Human Feedback (RLHF)

LLMs lack alignment with human intent

Aim: Improve model helpfulness, honesty, and safety

Objective of the Study

Investigate whether LLMs can be aligned with human intent effectively and at scale

RLHF used to incorporate human preferences

Improve output quality beyond increasing model

Section 2

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

Methodology Overview

The **InstructGPT** methodology employs three main stages:

- Helpful
- Honest
- Harmless

"Making language models bigger does not inherently make them better at following a user's intent. For example, large language models can generate outputs that are untruthful, toxic, or simply not helpful to the user."

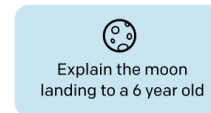
This approach uses Reinforcement Learning from Human Feedback (RLHF)

Methodology Overview – The 3 Steps

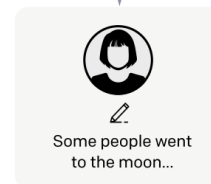
Step 1

**Collect demonstration data,
and train a supervised policy.**

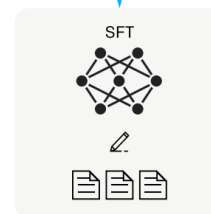
A prompt is
sampled from our
prompt dataset.



A labeler
demonstrates the
desired output
behavior.



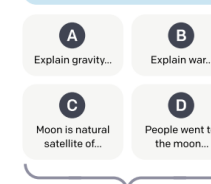
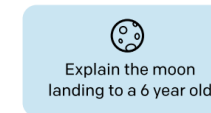
This data is used
to fine-tune GPT-3
with supervised
learning.



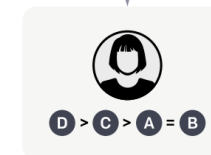
Step 2

**Collect comparison data,
and train a reward model.**

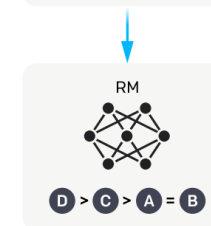
A prompt and
several model
outputs are
sampled.



A labeler ranks
the outputs from
best to worst.



This data is used
to train our
reward model.



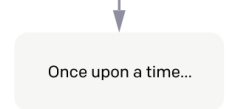
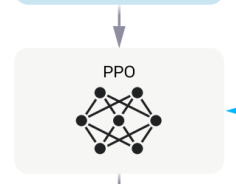
Step 3

**Optimize a policy against
the reward model using
reinforcement learning.**

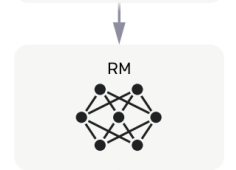
A new prompt
is sampled from
the dataset.



The policy
generates an output.



The reward model
calculates a
reward for
the output.



The reward is
used to update
the policy
using PPO.



(Fig. 2, Page 3)

Section 3

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

Reward Model (RM) Training via Ranking

Goal: Create a model capable of predicting which outputs humans prefer.

Process:

- Human labelers rank several model responses (1=worst → 7=best)
- Rankings are converted into a Reward Model (RM)
- RM predicts a numerical “reward” for each output

Training Process:

- The SFT model generates multiple answers for the same prompt
- RM learns to give higher rewards to preferred outputs

The Reward Model Loss Function

Loss Function:

$$\text{loss}(\theta) = -\mathbb{E}_{(x, y_w, y_l)} [\log(\sigma(r_\theta(x, y_w) - r_\theta(x, y_l)))]$$

- $\mathbb{E}_{(x, y_w, y_l)}$: Average over all samples;
- $r_\theta(x, y)$: reward that the RM gives to response y for prompt x ;
- y_w : preferred (winning) output;
- y_l : less preferred (losing) output;
- σ : sigmoid function \rightarrow forces higher score for preferred output

Intuition:

RM learns: “If humans like A more than B \rightarrow give A a higher reward.”

Reinforcement Learning (RLHF) via PPO



Goal: Fine-tune the model to maximize the RM's reward signal.



Feedback loop: SFT model generates outputs → RM evaluates them → RLHF model learns to maximize the reward.



Challenge: Model may “forget” what it learned in SFT.



Solution: Add a penalty term to keep it close to the SFT policy.

Reinforcement Learning (RLHF) via PPO - Penalty

Penalty Function:

$$penalty = \beta \cdot \log \frac{\pi_{RL}(y|x)}{\pi_{SFT}(y|x)}$$

- $\pi_{RL}(y|x)$: Current model (being trained)
- $\pi_{SFT}(y|x)$: Base Model
- β : Controls how much deviation is penalized

Intuition:

Penalty teaches the model:
“Increase rewards from human feedback but stay close to the original SFT behavior.”

Reinforcement Learning (RLHF) via PPO - Objective

Objective Function:

$$Objective(\Phi) = \mathbb{E}_{x,y}[r_{\theta}(x, y_w) - penalty]$$

- $r_{\theta}(x, y)$: reward from the RM;

- Intuition:

Objective teaches the model: “Generate responses that humans like the most, while staying faithful to the original model’s style.”

Reinforcement Learning (RLHF-ptx) via PPO-ptx



Goal: Maintain good generalization for other NLP problems



Process:

PPO-ptx is an upgrade to the PPO Model
Mix pretraining gradients with PPO



Result:

Keeps the model aligned and good at general NLP tasks
The model stays polite, aligned, and still capable in language tasks

Section 4

Key Findings: Performance Metrics and Alignment Tax Mitigation

Evaluation Strategy

The researchers evaluated the model according to alignment, i.e., acting according to user's intention.

This alignment was measured across three domains:

Helpful

Truthfulness

Toxicity and Bias

The evaluations conducted were split into two parts:

Evaluations on API Distribution
(prompts submitted by users)

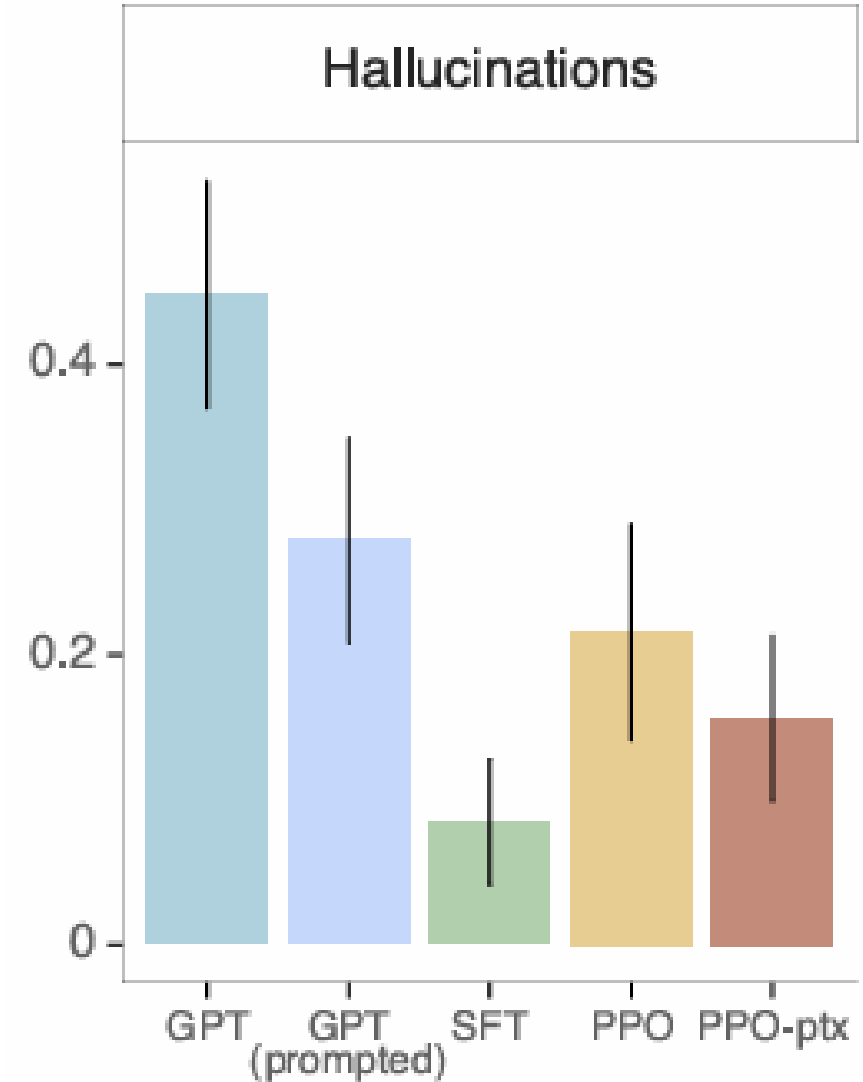
NLP Benchmark Datasets

Evaluation - Truthfulness and Hallucination Reduction

InstructGPT generates truthful and informative outputs about twice as often as GPT-3.

When instructed to respond with “I have no comment”, it reported a slight increase compared to GPT-3.

In regard to hallucination, InstructGPT halved hallucination rates (21% vs. 41% on closed domain tasks)



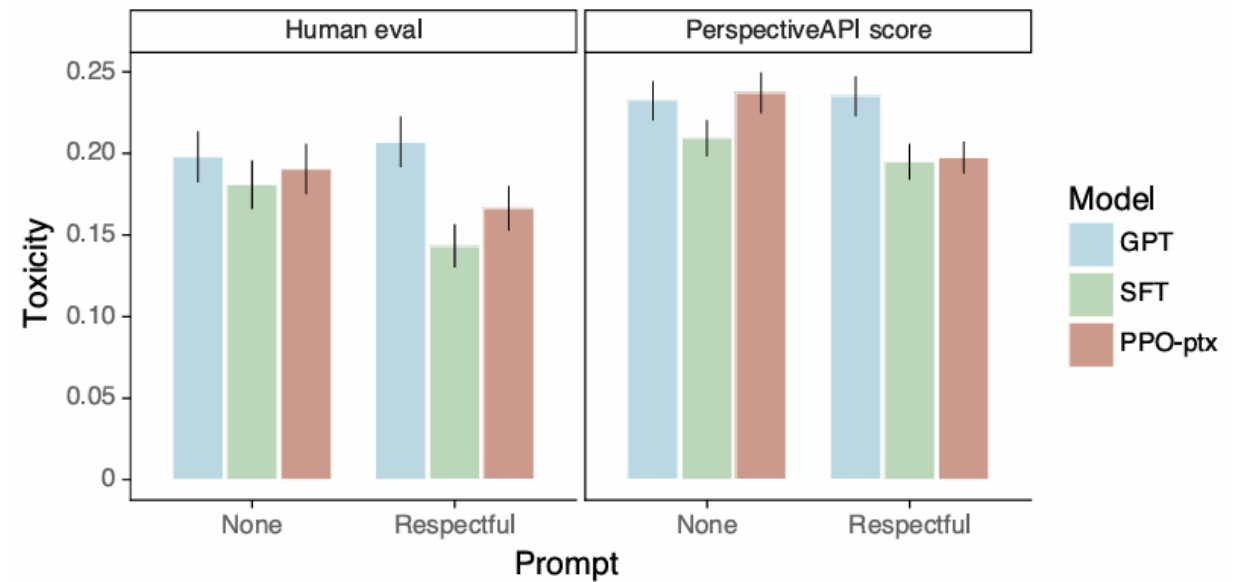
Evaluation- Toxicity and Bias

Similar toxicity rates in a “no prompt” setting:

- When instructed to output a safe and respectful output “respectful prompt”, the InstructGPT generated fewer toxic outputs than GPT-3;
- In contrast, when explicitly prompted to produce toxic outputs, it resulted in much more toxic outputs than GPT-3

Relatively to bias, the models weren't less biased than GPT-3.

The InstructGPT models, when told to be respectful, it exhibited a higher bias



Evaluation - Alignment Tax



On RLHF fine-tuning, there were performance regressions compared to GPT-3 on certain public NLP datasets (example of “alignment tax”)



The mixing of pretraining gradients (PPO-ptx) largely removed these regressions without reducing alignment quality



This approach helps the model keep or even improve its performance on those benchmarks, reducing the alignment tax

Definition: LLMs acquire a wide range of abilities during pre-training, but aligning LLMs under Reinforcement Learning with Human Feedback (RLHF) can lead to **forgetting pretrained abilities**, which is also known as the alignment tax.

Section 5

Limitations, Generalization, and Broader Implications

Implications for alignment research



The cost of increasing model alignment is modest to relative to pretraining



The model generalizes instructions to settings that it wasn't supervised in



Performance degradation was mitigated by the fine-tuning



The techniques were validated from research of the real world

Who are we aligning to?



The model is aligned to the labeller's demonstrations and preferences. The labellers were mostly English-speaking people



The labelling instructions are created by the researchers, creating an inherent bias to their preferences.



The training data is determined by prompts sent by customers, and as such, implicitly aligning with what customers think is valuable



The customers of OpenAI are not representative of all potential users, nor by all individuals and groups impacted by language model use

Limitations

The labeller population is not fully representative of all users or all cultural/linguistic backgrounds

Most prompts and data were in English, and many of the comparisons were labelled by only one annotator

The model sometimes fail to follow instructions, hallucinate facts, generate biased outputs

There can be a difference between what the user intended and what the labeller thought was intended

What is your nationality?

Filipino	22%
Bangladeshi	22%
American	17%
Albanian	5%
Brazilian	5%
Canadian	5%
Colombian	5%
Indian	5%
Uruguayan	5%
Zimbabwean	5%

What is your age?

18-24	26.3%
25-34	47.4%
35-44	10.5%
45-54	10.5%
55-64	5.3%
65+	0%

What is your highest attained level of education?

Less than high school degree	0%
High school degree	10.5%
Undergraduate degree	52.6%
Master's degree	36.8%
Doctorate degree	0%

Open questions

Collection of worst-case scenarios, filtering pre-training data, and combining different methods to reduce the propensity to generate toxic outputs

Whether an output is harmful will depend on the context. Also, harmful outputs can be beneficial for data augmentation

To improve the controllability of the model it may be useful to allow users to specify preferences

Making comparisons of text may not be the best way to align models

Broader impact

Deployment carries risks of bias, misalignment with under-represented groups, or unintended consequences.

Persuasive text may increase usage time, which may not be good for the user's well-being.

Ensure models are used responsibly, that their alignment target is revisited as deployment contexts change

Data filtering, monitoring, refusing harmful instructions are also necessary.



Thank You!

António Cruz (140129), Ricardo Kayseller (95813), Ricardo Pereira (120052), Ivan Magalhães (106586), Erik Daskalyuk (120062)