VARIABLZ Academy

REINFORCEMENT LEARNING WITH HUMAN FEEDBACK

For LLM Engineers

RLHF

Reinforcement Learning from Human Feedback

Paper published by Open AI by 2022

RLHF

VARIABLZ Academy

Open AI

Anthrop\c

DeepMind

Meta AI

X AI

Industry Standard for Fine Tuning LLMs

Why Fine-Tuning

LLMs?

- Scaling alone doesn’t ensure alignment
- Misalignment Issues


Harmful


Unhelpful

1. Followed Human Instructions
2. Followed Human Preference

Reinforcement Learning from Human Feedback (RLHF)

1. Supervised Fine-Tuning (SFT)
2. Reward Modeling (RM)
3. Reinforcement Learning (PPO)
4. Alignment Tax
5. Reinforcement Learning (PPO-ptx)

1. Supervised Fine-Tuning (SFT)

GPT-3

40 Human
Labellers

Prompts (13,000)

1. Text Generation
2. Q&A
3. Brainstorming
4. Chat
5. Summarization
6. Rewriting
7. Extraction, etc.

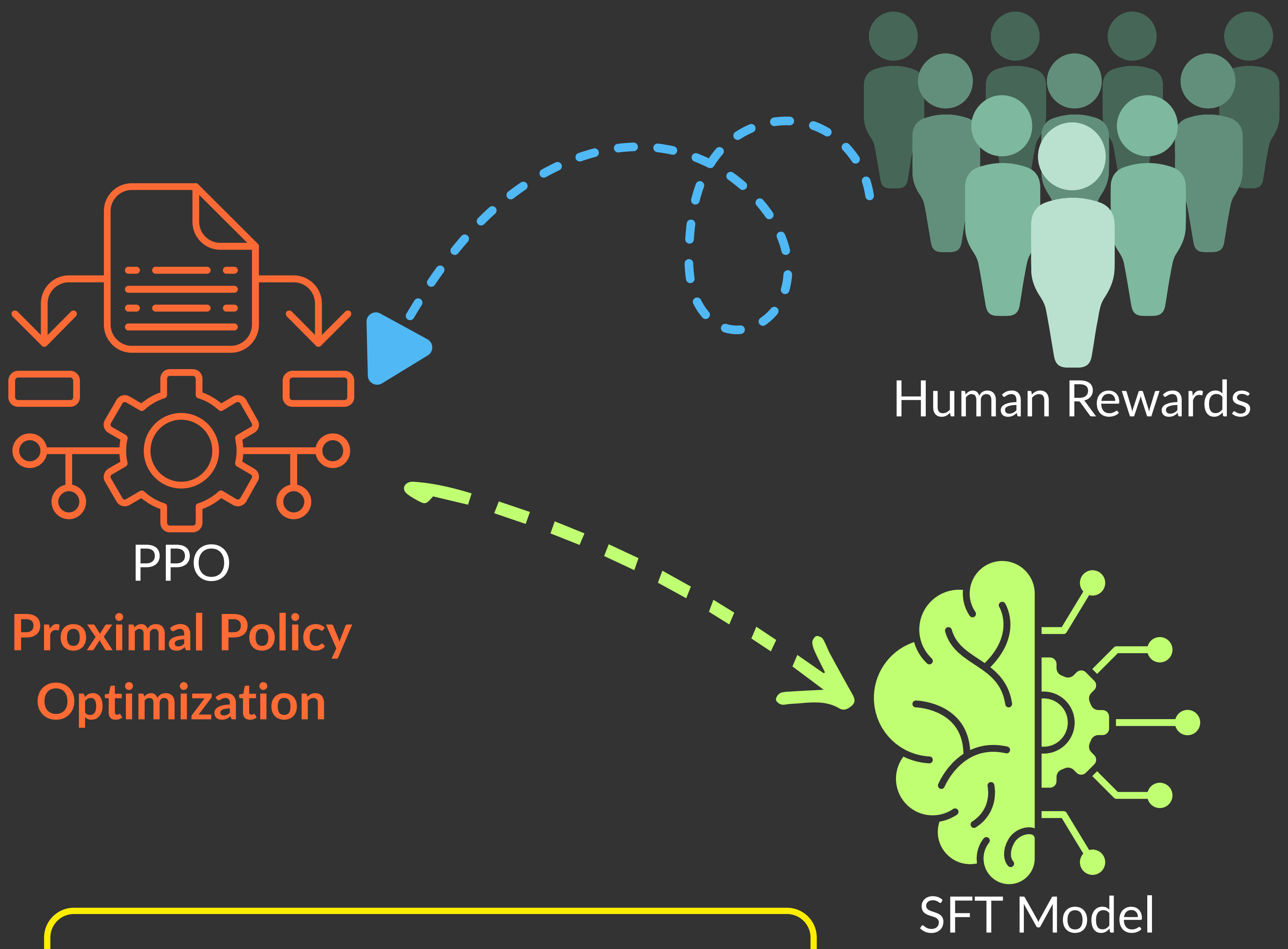
SFT

1. Helpful
2. Truthful
3. Harmless

Fine tuned model generated appropriate, relevant and constraint following outputs, and less hallucination and toxic contents

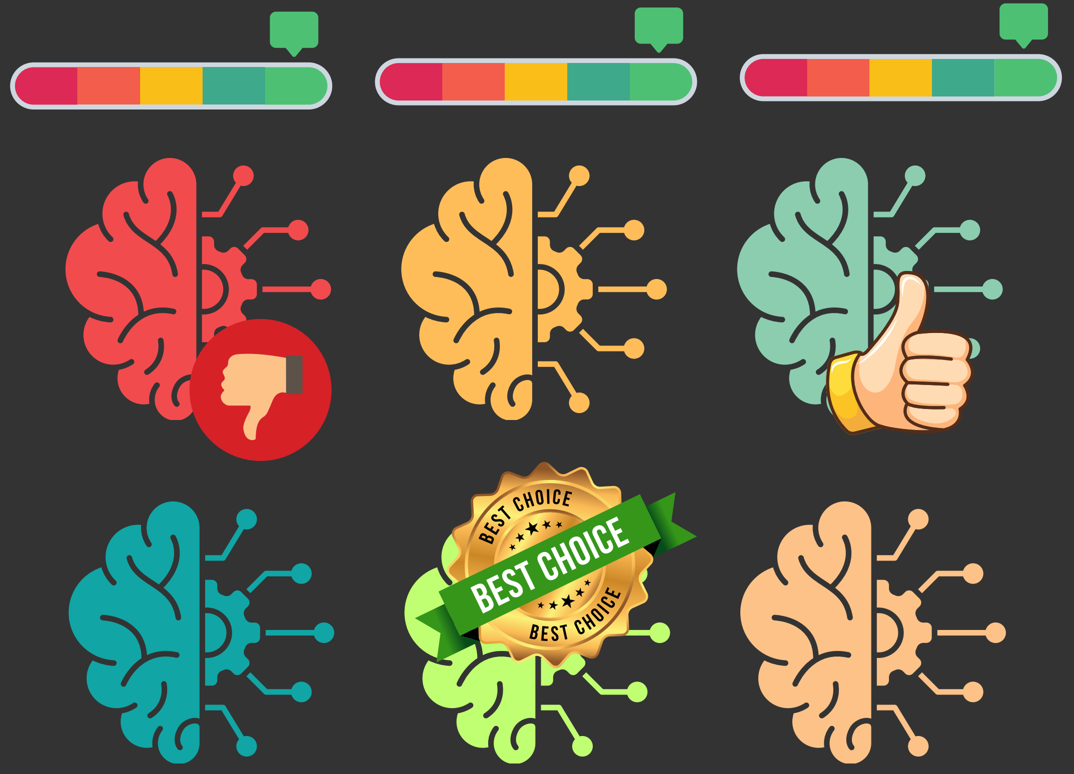
1. Followed Human Instructions
2. Followed Human Preference

How a model follows Human Preference?



Directly involving humans in every training step would be impossibly slow and expensive

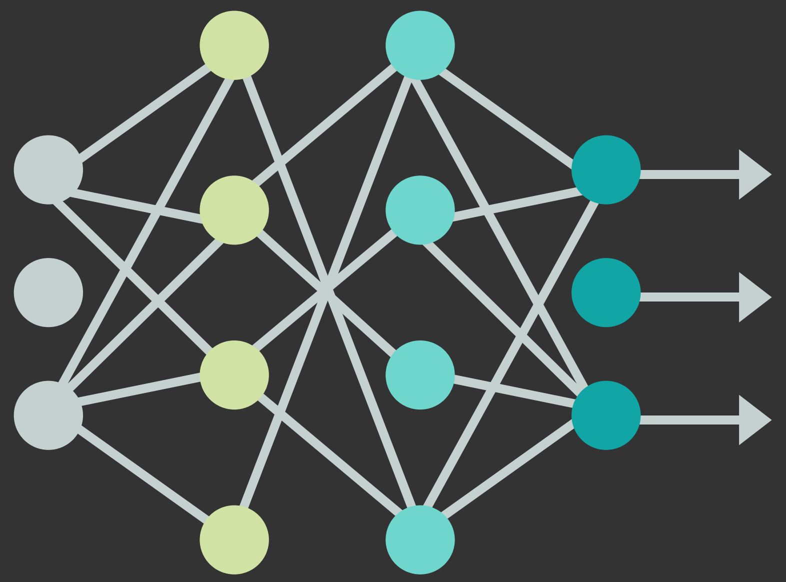
2. Reward Model (RM)



SFT Models

Human Raters

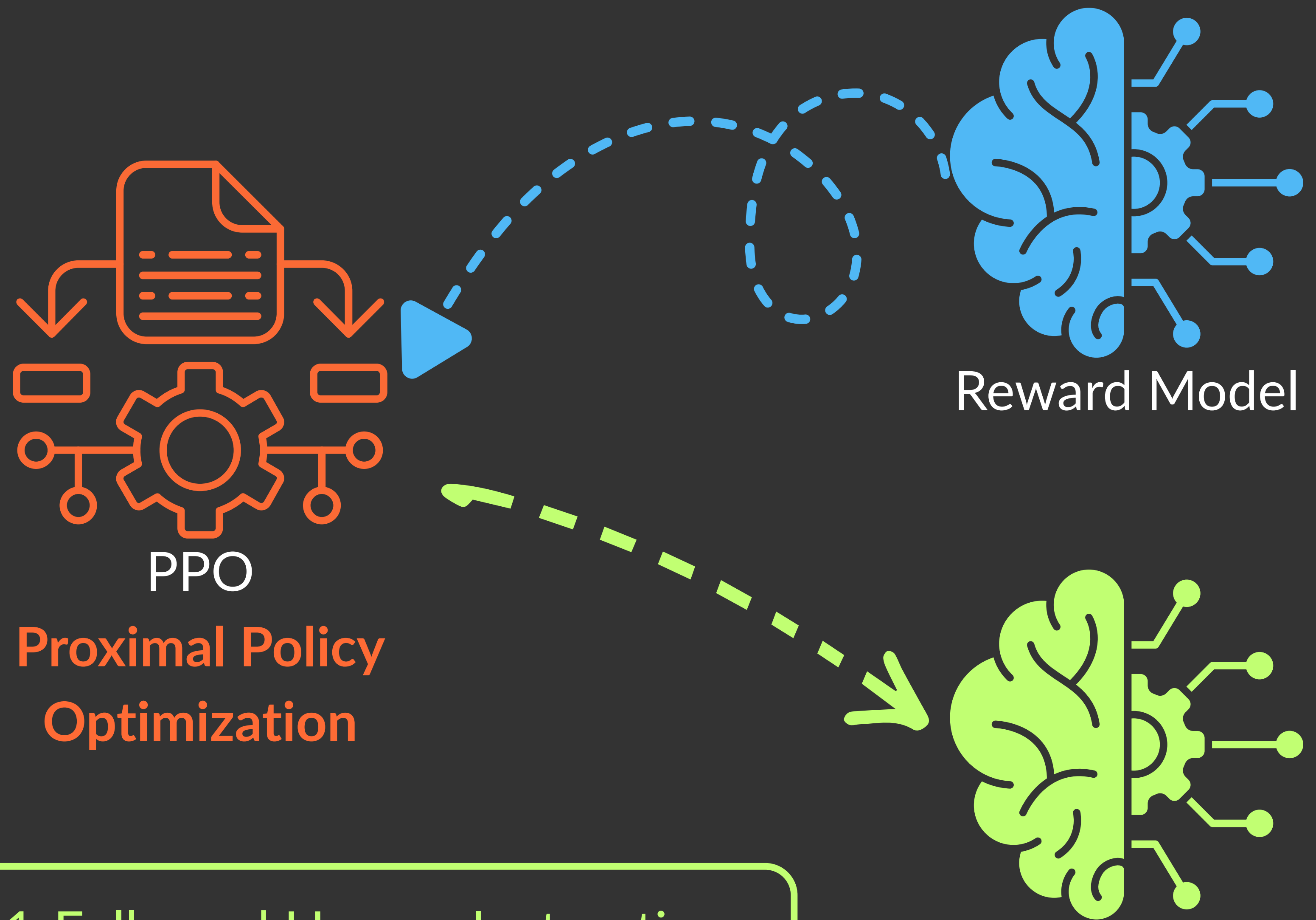
The reward model is then trained on a dataset to predict which response a human would prefer.



Neural Network

Model to Judge Human Preference

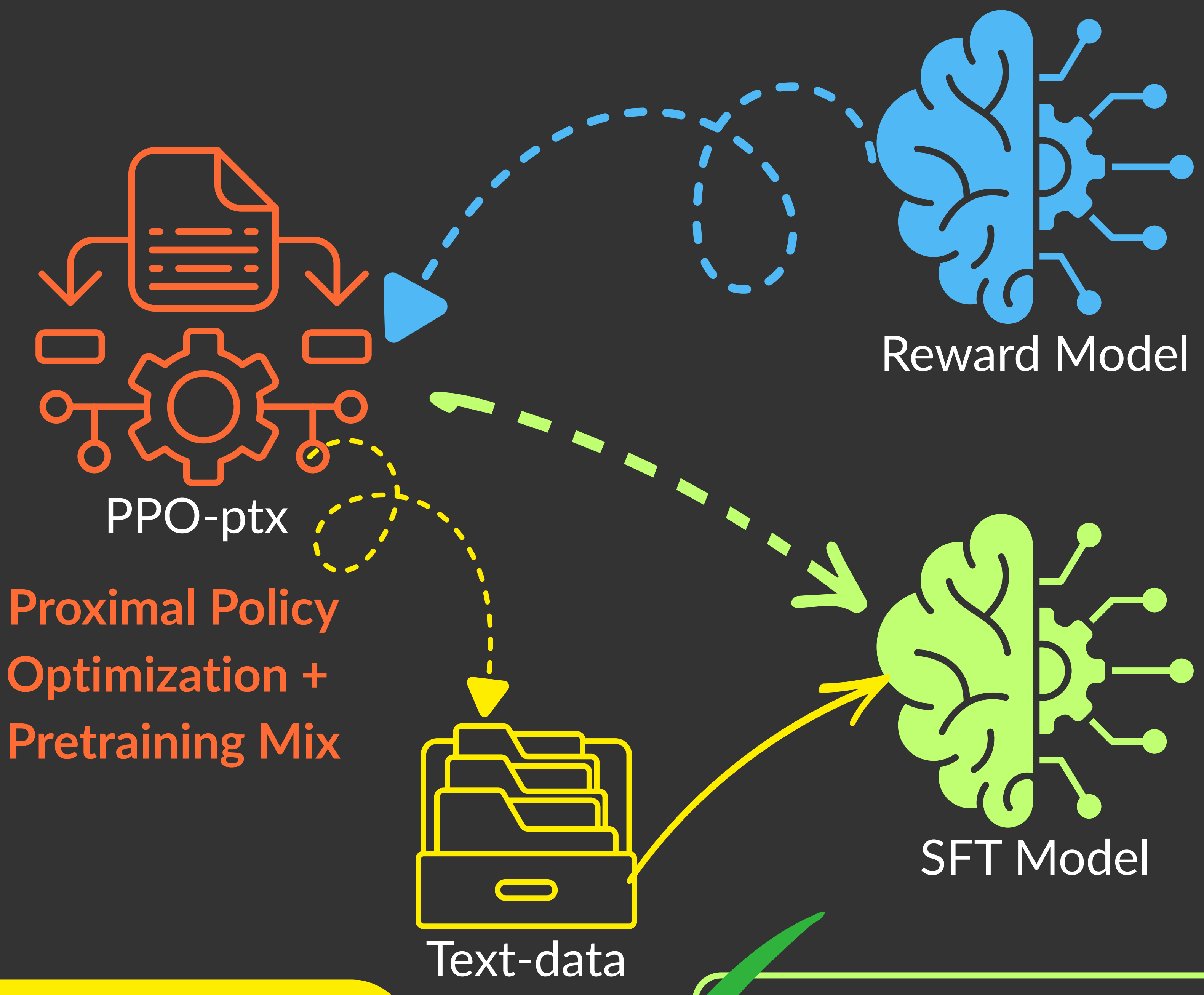
3. Reinforcement Learning (PPO)



1. Followed Human Instructions
2. Followed Human Preference
3. Lost Language Capabilities

Alignment Tax

4. Reinforcement Learning (PPO-ptx)



RLHF

1. Followed Human Instructions
2. Followed Human Preference
3. Retains Language Capabilities

VARIABLZ
Academy

THANK YOU!

If you found this very useful,

Share Your Love ❤️ with a 5⭐ Review.

- Parthiban Kannan
Variablz Academy