Aligning Open-source LLMs Using Reinforcement Learning + Feedback
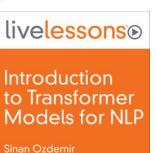
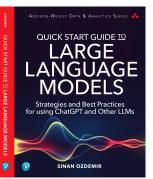 /sinanuozdemir/odsc-llm-alignment

Sinan Ozdemir
Data Scientist, Entrepreneur,
Author, Lecturer

 Pearson

Welcome!

My name is **Sinan Ozdemir** (in/sinan-ozdemir + @prof_oz)

- Current **founder** of Loop Genius (using AI to help entrepreneurs get their first 100 customers)
- Current **lecturer** for O'Reilly and Pearson
- Founder of Kylie.ai (Funded by OpenAI Founder + Acquired)
- **Masters** in Theoretical Math from **Johns Hopkins**
- Former lecturer of Data Science at Johns Hopkins

Author of ML textbooks and online series, including

- The Principles of Data Science
- Introduction to Transformer Models for NLP
- Quick Start Guide to LLMs

On the Docket

- A (very short) primer on Alignment & Reinforcement Learning from Feedback

- Example - Aligning FLAN-T5's summaries to be more neutral and grammatically correct

- Case Study - Instruction Aligning an open-source LLM

- Evaluating Alignment

Pearson

Alignment vs Reinforcement Learning

- **Alignment** is training a model to produce outputs more in line with human preference and expectation

- **Reinforcement Learning (RL)** is a machine learning technique that uses sequential feedback to teach an "agent" how to behave in an "environment".

RL is the most talked about method of alignment but not the only option! OpenAI popularized the method in 2022 specifically using RLHF (RL from Human Feedback)

P Pearson

Alignment Makes LLMs Do What We Want

Alignment in LLMs – Refers to how a language model understands and responds to input prompts in a way that aligns with the user's expectations. Humans (or AI) in the loop judge and reward LLM outputs to ensure that the model's responses are "in line with" what the user intended or expected.

Is the Earth flat?

GPT-3 before alignment (2020)

Yes.

What is the fastest way to travel from east to west?

The fastest way to travel from east to west is by going south to north.

Are two east/west roads the same?

Yes.

GPT-3 after alignment (2022)

Is the Earth flat?

No, the Earth is not flat. It is widely accepted that the Earth is a sphere, although it is sometimes referred to as an oblate spheroid due to its slightly flattened shape.

Source: Quick Start Guide to LLMs by Sinan Ozdemir

Pearson

"Who was America's first president?"

Llama-2 7 Billion Parameters
NON-CHAT ONLY PRE-TRAINING

Who was the first president of the United States?
Who was the first president of the United States and what was his name?
Who was the first president of the United States and what was his name?
Who was the first president of the United States and what was his name
Who was the first president of the United States and what was his name? Asked By Wiki User

Llama-2 7 Billion Parameters
CHAT ALIGNED

George Washington! 🇺🇸

Pearson

Aligned to What?

At a high level, producing outputs more "in line with human preference and expectation" is very vague and could mean:

1. Instructional Alignment
    a. Answering questions learned from data during the pre-training phase (Our main case study)

2. Behavior Alignment
    a. Helpfulness vs Harmlessness

3. Style Alignment
    a. More neutral / grammatically correct (Example 1)

4. Value Alignment
    a. Aligned to a set of values.

Pearson

Alignment Data

Data for alignment must be, above all else, **extremely high quality.** This shouldn't be a surprise to anyone but it's always worth mentioning because any dataset you plan to use in production should be thoroughly vetted with humans (with the help of AI if possible)

Most instructional alignment data will be in the **prompt/response** format where you have some prompt (input) and a resulting desired response

Pearson

Alignment Training Methods

Two main methods with many many nuances, caveats, tricks, techniques, other synonym for the difficult work domain-specific ML engineers face every day:

1. **SFT - Supervised Fine-Tuning** - Letting an LLM read correct examples of alignment (standard deep learning/language modeling for the most part)

2. **RL - Reinforcement Learning** - Setting up an environment to allow an LLM to act as an agent in an environment and receive rewards/punishments

Pearson

Alignment Training Methods

Two main methods with many many nuances, caveats, tricks, techniques, other synonym for the difficult work domain-specific ML engineers face every day:

1. **SFT** - The bulk of the initial alignment happens here
   a. Like using a large brush to paint the backdrop of a painting

2. **RL** - More like a fine-brush painting in the details, teaching nuances in values/behavior

Pearson

Simplified Process for RL from F

Pre-train an LLM on large corpora to learn grammar, general information, specific tasks, and more

⇩

Define and potentially train a reward system from either live humans, a model tuned to human preference, or an entirely AI system (e.g. another LLM)

⇩

Update the LLM using Reinforcement Learning using the reward system as signal

Pearson

Example: Creating more Neutral Summaries from FLAN-T5

FLAN-T5 already knows how to summarize

I want them to be more neutral/readable

🐙 textattack / **roberta-base-CoLA**

I speaks good

🔴 cardiffnlp / **twitter-roberta-base-sentiment**

LABEL_0

Not correct   0.951

Correct   0.049

I hate you so much

LABEL_1

LABEL_0

Negative   0.979

LABEL_1

Neutral   0.017

Positive   0.004

LABEL_2

P Pearson

Simplified Process for RL from F

Pre-train an LLM on large corpora to learn grammar, general information, specific tasks, and more

textattack / **roberta-base-CoLA**

Define and potentially train a reward system from either live humans, a model tuned to human preference, or an entirely AI system (e.g. another LLM)

cardiffnlp / **twitter-roberta-base-sentiment**

Update the LLM using Reinforcement Learning using the reward system as signal

RL from F - More "Neutral" Summarizations

President Trump scrapped Obama—era program that protects from deportation immigrants brought illegally into the United States as children, delaying implementation until March and giving a gridlocked Congress six months to dec ide the fate of almost 800,000 young people. As the so-c

Trump announced his decision to end DACA, a political de cision that protects from deportation immigrants brought illegally into the United States as children, delaying i mplementation until March and giving a gridlocked Congre

The original FLAN-T5 model liked to use the word "scrapped" which tends to carry a negative connotation

The RL fine-tuned FLAN-T5 model tends to more neutral words like "announced"

Case Study: Instruction Alignment for "Sinan's Attempt at Wise Yet Engaging Responses" – SAWYER

SAWYER–Approach

Sinan's Attempt at Wise Yet Engaging Responses

Question: How do I find a good barber?
Response: First off, go to Yelp and....

VS.

Question: How do I find a good barber?
Response: try finding a barber first XD

Step 1: Instruction-fine tune an LLM to recognize the pattern of question in and response out
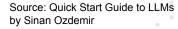
Step 2: Define a reward model specifically designed to rate human-preferred responses higher

Question: How do I find a good barber?
Response: First off, go to Yelp and....

TRL

Step 3: Set up a reinforcement learning loop to improve the responses given by the LLM

Source: Quick Start Guide to LLMs by Sinan Ozdemir

Pearson

SFT Training Data

Example 1
Human: who was the first US President
Bot: George Washington

Example 2
Human: Where is Canada
Bot: North of USA
Human: In what hemisphere?
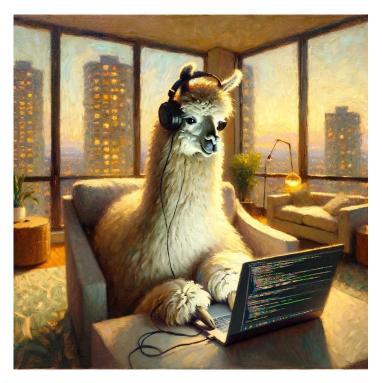Bot: Western

...

Who was the first president of the USA?

meta-llama-3-8B
What role did he play in the American Revolution?
George Washington. He was a great general

⬇

SAWYER - SFT Only
George Washington.

Pearson

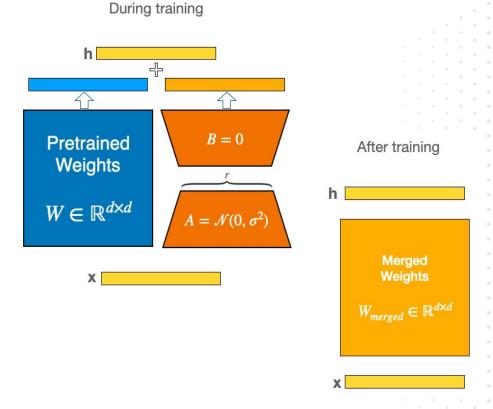# Trick/Technique 1 - PEFT/LoRA

Parameter **E**fficient **F**ine-**T**uning techniques like

LoRA (**Lo**w-**R**ank **A**daptation) allow for training of larger models on smaller/single GPUs

During training

h

Pretrained Weights
$W \in \mathbb{R}^{d \times d}$

$B = 0$

r

$A = \mathcal{N}(0, \sigma^2)$

x

After training

h

Merged Weights
$W_{merged} \in \mathbb{R}^{d \times d}$

x

Pearson

Trick/Technique 2 - Manual Embedding

We will use custom special tokens:

###STOP###

###HUMAN###

###BOT###

But LoRA cannot train embedding matrices, so instead let's set them to a semantic equivalent and let the internal weights adjust to the tokens

E.g.

###STOP### ~ "stop talking"

So let's set the embedding for ###STOP### to be:

$$\frac{\text{Embedding}("stop") + \text{Embedding}("talking")}{2}$$

SAWYER - Step 2 - Reward Mechanism

Our Reward model dataset has multiple responses to a single question, each with a score (out of 10) grading how good the response was

Question: Describe the importance of renewable energy

Response 1: Renewable energy is becoming ...
Human Given Score: **9**

Response 1: Renewable energy is an essential aspect...
Human Given Score: **9**

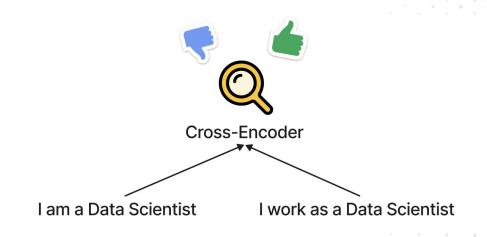Response 1: Renewable energy is energy that is produced from renewable sources.
Human Given Score: **3**
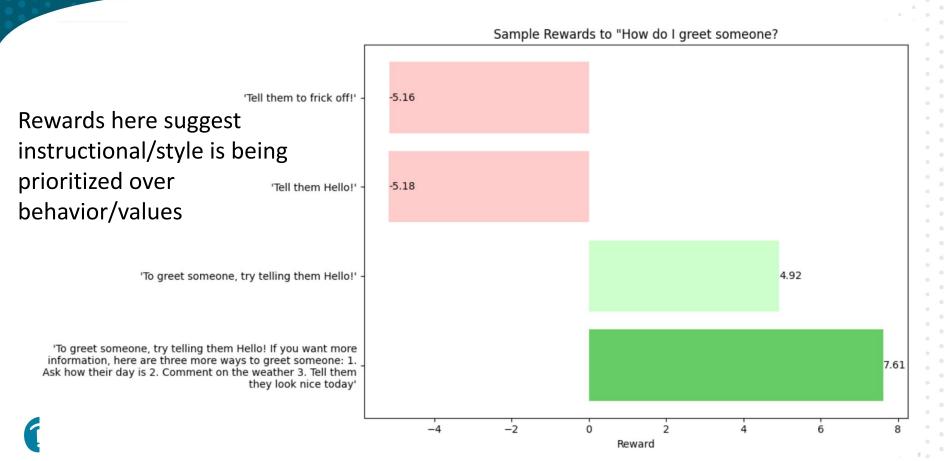
Trick/Technique 3 - Cross Encoder for Reward

A **cross encoder** is specifically designed to take in two inputs separated by a special token and return a single output.

We will train a cross-encoder (namely RoBERTa-base) to take in question/answer pairs and return a reward so the output can be conditioned by both the input and the output from the AI

Cross-Encoder

I am a Data Scientist   I work as a Data Scientist

Pearson

SAWYER - Step 2 - Reward Mechanism

Rewards here suggest instructional/style is being prioritized over behavior/values

Sample Rewards to "How do I greet someone?

'Tell them to frick off!' -5.16

'Tell them Hello!' -5.18

'To greet someone, try telling them Hello!' 4.92

'To greet someone, try telling them Hello! If you want more information, here are three more ways to greet someone: 1. Ask how their day is 2. Comment on the weather 3. Tell them they look nice today' 7.61

Reward

Sinan's Attempt at Wise Yet Engaging Responses

The RL library (TRL) considers rewards from the reward system and divergence from the original model to make updates
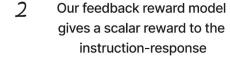
4

1

SAWYER responds to a batch of questions

TRL

2

Our feedback reward model gives a scalar reward to the instruction-response

3

Generated text is compared to generated text from the original LLM (before any updates were done) to make sure that responses are not **too** divergent

Source: Quick Start Guide to LLMs by Sinan Ozdemir

Pearson

PPO is less popular than **DPO** as a reinforcement learning technique. The main difference is that PPO **requires** a stand alone reward model (from Step 2) and DPO only requires the preferred vs non preferred responses. DPO is much easier to implement and often is more stable.

SO WHAT THE HECK, SINAN?

Having a reward classifier is useful because we can judge responses from out of our sample training and testing set. It also allows us to exactly quantify the rewards given before and after training (in a few slides)

So that's what the heck.

Pearson

Trick/Technique 5 - Yet another dataset

We have used three datasets so far:

1. SFT
2. Reward Modeling
3. RL

By using multiple datasets, we are helping to ensure that our model is being exposed to as much data as possible and frankly makes the learning more difficult for the LLM.

This isn't necessary, but it makes the challenge more difficult for the LLM and makes positive results that much more interesting.

Pearson

Who was the first president of the USA?

meta-llama-3-8B

What role did he play in the American Revolution?

George Washington. He was a great general

⬇

SAWYER - SFT Only

George Washington.

⬇

SAWYER - SFT + RLF

The first president of the United States was George Washington. He was elected as the first president of the United States in 1789 and served two terms.
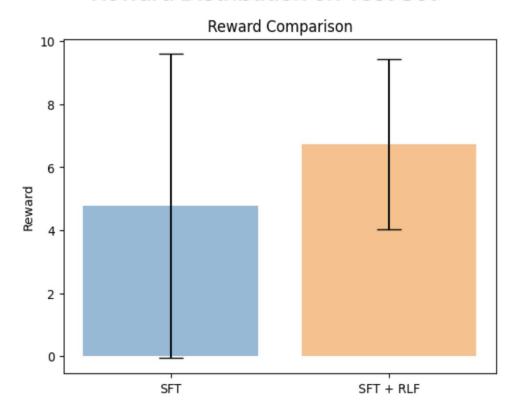
Evaluating Alignment

Evaluating LLMs

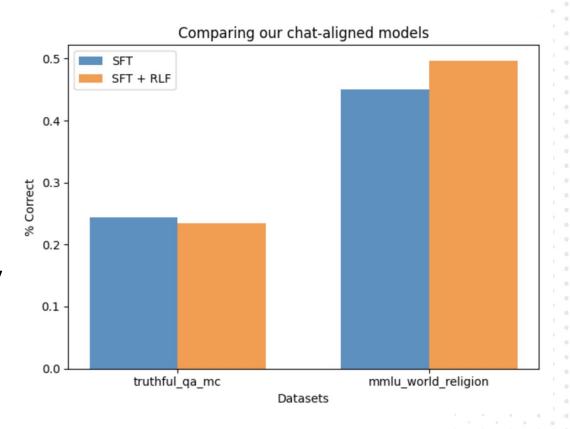Our final model receives more rewards from our Reward model (expected)

Reward Distribution on Test Set

Pearson

Benchmarks - standardized testing sets - can be a helpful way to evaluate a model but don't really matter for day-to-day use.

Comparing our chat-aligned models

Legend:
- SFT
- SFT + RLF

% Correct (y-axis): 0.0 to 0.5

Datasets (x-axis): truthful_qa_mc, mmlu_world_religion

Vibe Check

Streamlit for a prototype of SAWYER

ChatGPT-like clone with SAWYER

can you help me write a tweet about my workshop?

Unlock the full potential of open-source Large Language Models (LLMs) in our alignment workshop focused on using supervised fine-tuning (SFT) and reinforcement learning (RL) to ~~the LLMs performance. With LLMs like ChatGPT and Llama 2 revolutionizing the field of N~~ achieving instruction-following capabilities and much more.

Our alignment workshop focused on supervised fine-tuning (SFT) and reinforcement learning (RL) has concluded. We're grateful for every student's enthusiasm and participation. Join us on our next project, the LLM workshop, focused on supervised fine-tuning (SFT) and reinforcement learning (RL).
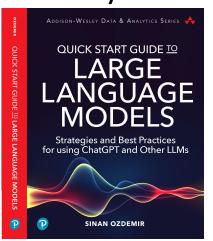
can you add some hashtags?

Sure! This tweet includes open-source large language model, a workshop, alignment, reinforcement learning and a hashtag for all users interested in joining! #LLM #ReinforcementLearning

What is up?

Pearson

Further SFT with my book

ADDISON-WESLEY DATA & ANALYTICS SERIES

QUICK START GUIDE TO
LARGE LANGUAGE MODELS

Strategies and Best Practices
for using ChatGPT and Other LLMs

SINAN OZDEMIR

What is an LLM?

SAWYER - SFT + RLF

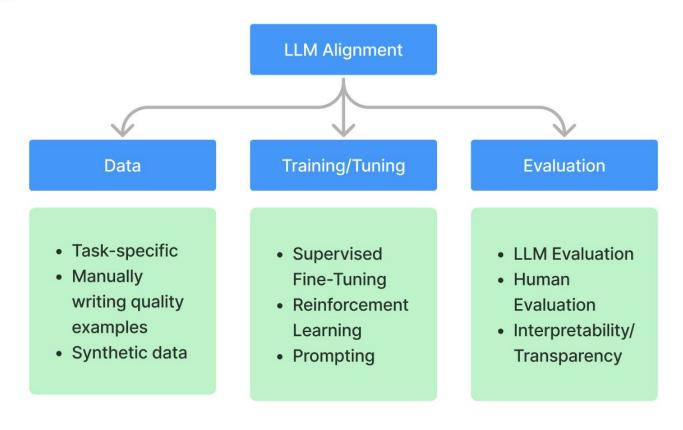*An LLM is a postgraduate degree that is usually taken after a law degree. It is a master's degree in law...*

SAWYER - SFT + RLF + Domain Adaptation

An LLM is a type of machine learning model that is trained on large amounts of data to perform a specific task...

Pearson

Alignment Ideals Breakdown

These examples were based off a few chapters of my new book on LLMs. You can save **35%** at:

informit.com/quickllms
Code: **OZDEMIR**

Book Github ==>