Module 10 - Exercise

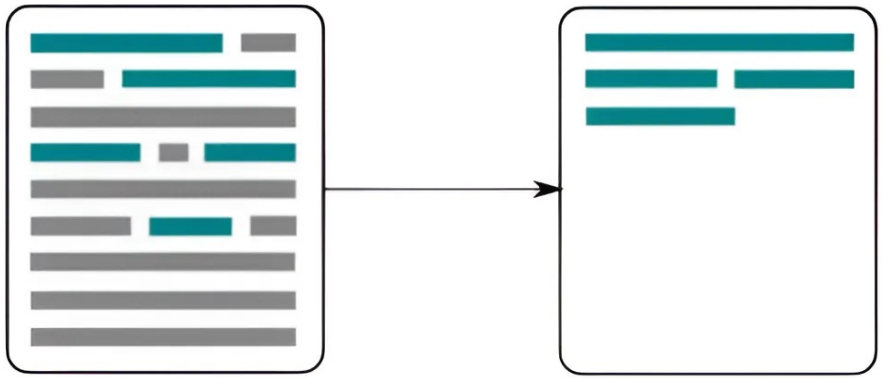
Text Summarization with Human Feedback (RLHF)

AI VIET NAM
Nguyen Quoc Thai

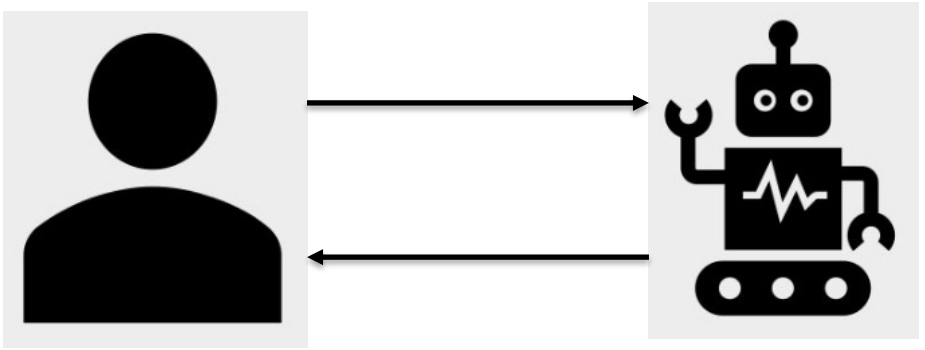
Objective

- ChatGPT

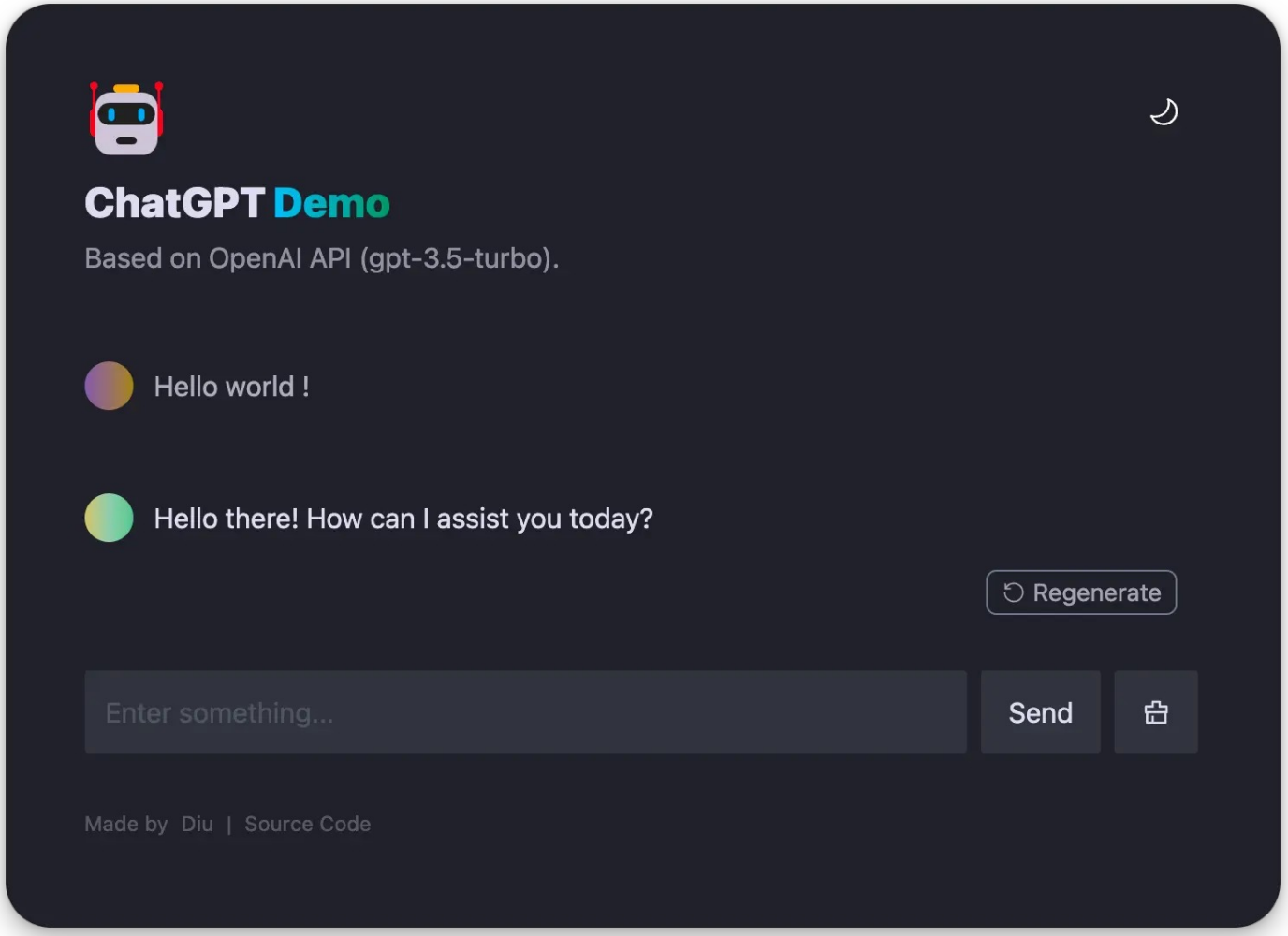
- Text Summarization



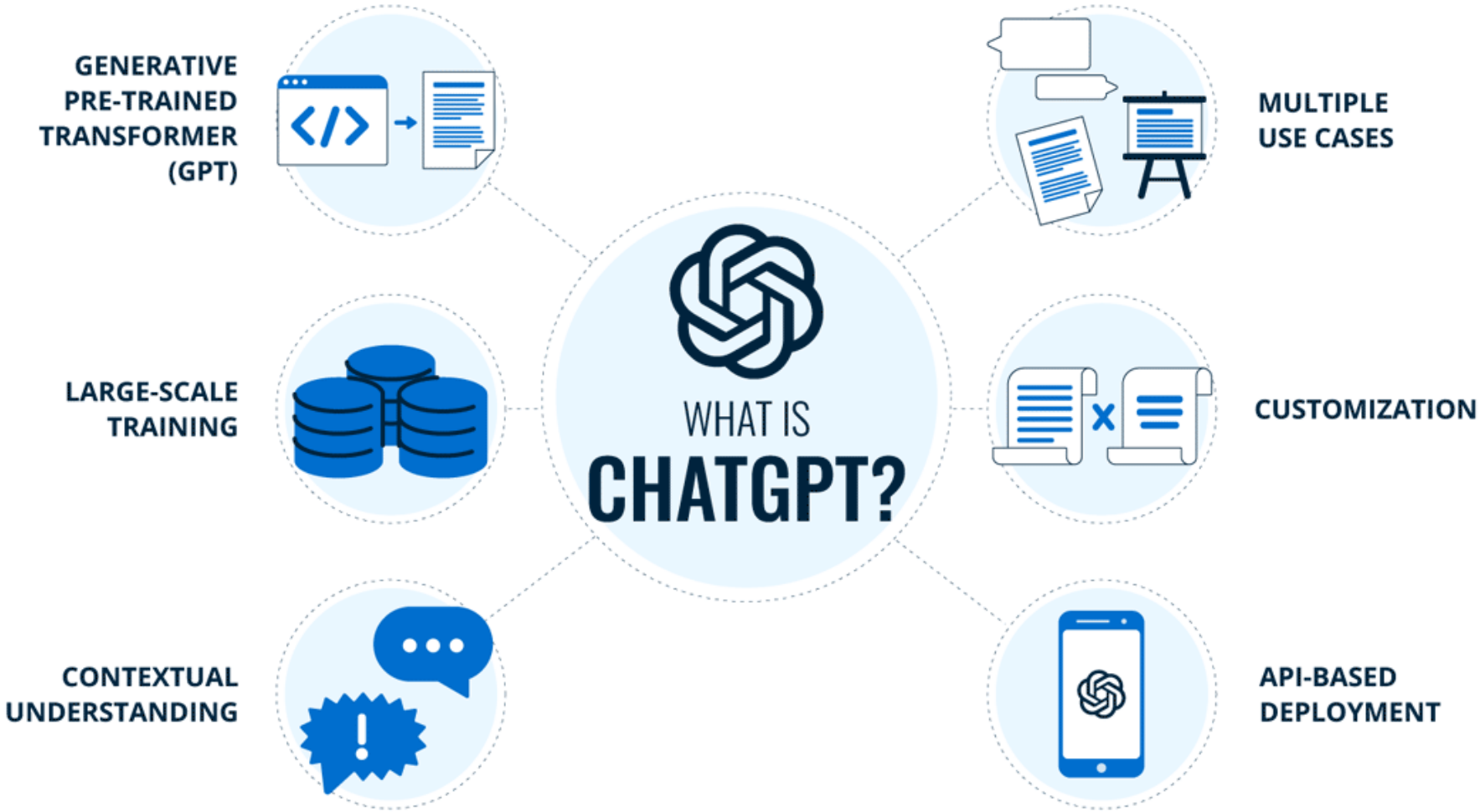
- Human Feedback



ChatGPT


ChatGPT Demo
Based on OpenAI API (gpt-3.5-turbo).
Hello world !
Hello there! How can I assist you today?
Regenerate
Enter something...
Send
Made by Diu | Source Code

ChatGPT

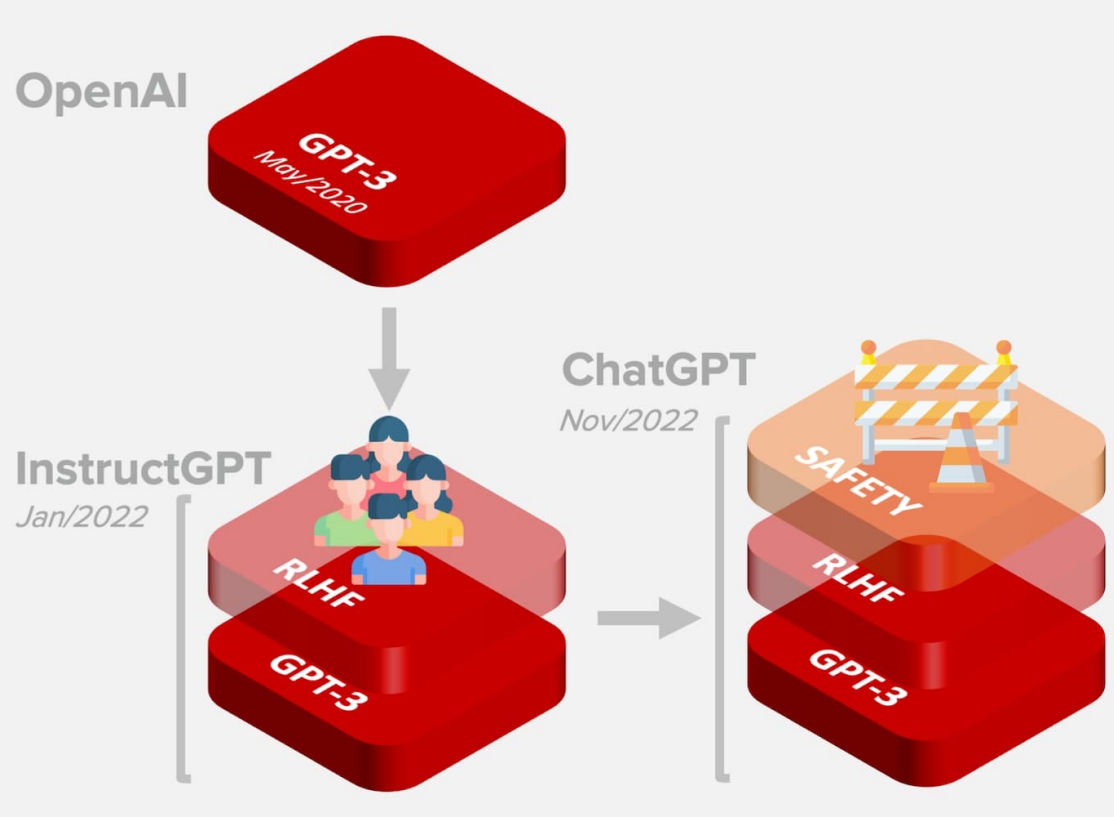

GENERATIVE PRE-TRAINED TRANSFORMER (GPT)
LARGE-SCALE TRAINING
CONTEXTUAL UNDERSTANDING
WHAT IS CHATGPT?
MULTIPLE USE CASES
CUSTOMIZATION
API-BASED DEPLOYMENT

Outline

- Large Language Models
- InstructGPT (RLHF)
- Text Summarization using RLHF

Large Language Models

ChatGPT



84%
More truthful

1.5 years
More knowledge
To June 2021

HHH: Helpful, honest, harmless
New alignment objective to be useful, truthful, and careful

InstructGPT
OpenAI

1.9x
Larger context window
(from 2,048 to 4,000 tokens)

58%
Less hallucinative

Large Language Models

Large Language Models

- Medium-sized models: BERT/RoBERTa models (100M or 300M), T5 models (220M, 700M, 3B, 11B)
- “Very” large LMs: models of 100+ billion parameters

  GPT3 (175B), BLOOM (176B), PaLM (540B), GLaM (1200B)…
- Larger model sizes => Larger compute, more expensive during inference

Large Language Models

Large Language Models

- Data scale: usually in the order of trillions of tokens

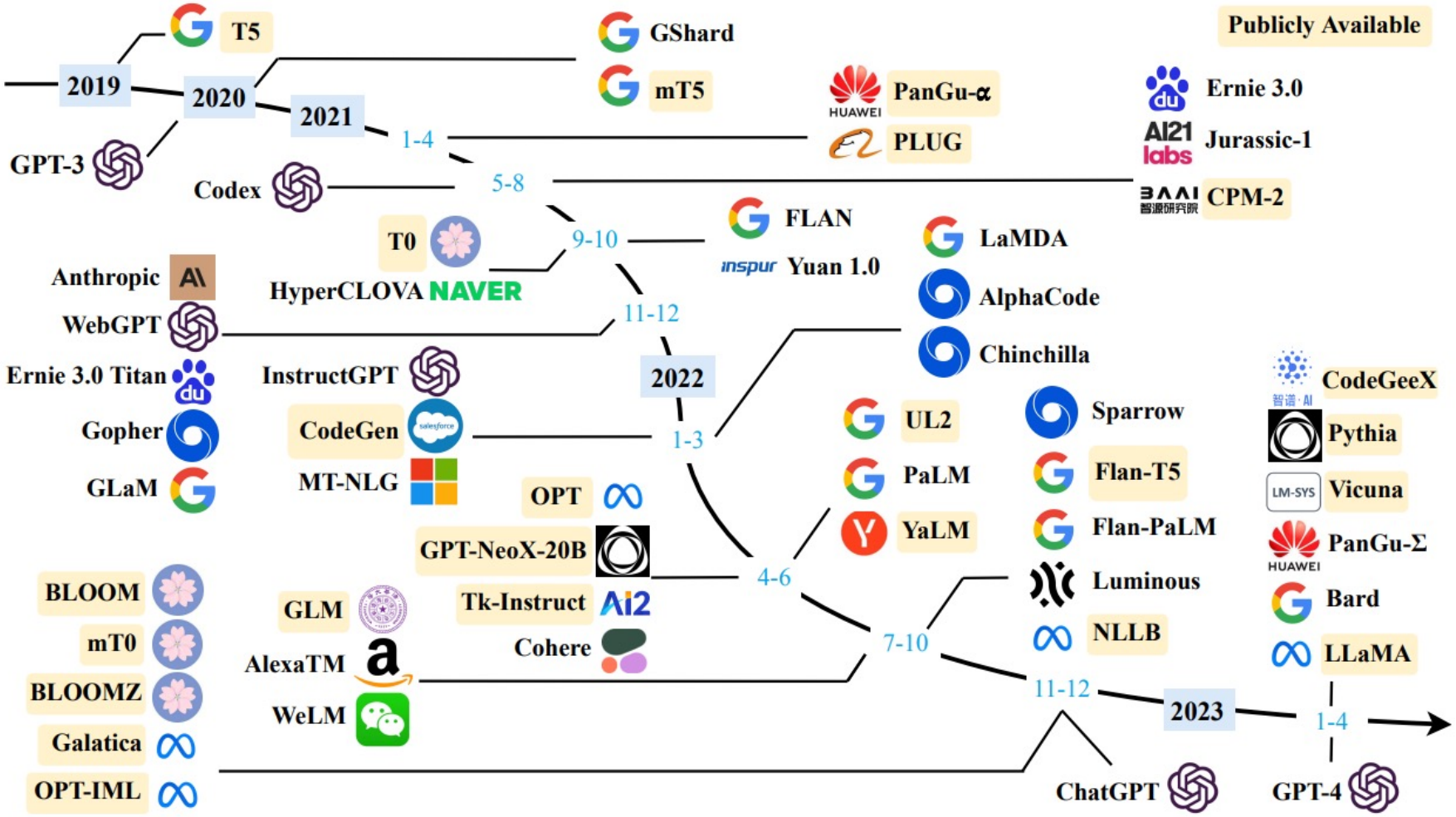
  GPT3 (0.5 trillion tokens), LLaMA (1.4 trillion tokens), …

- Training data: Low-quality data

Corpora	Size	Source	Latest Update Time
BookCorpus [122]	5GB	Books	Dec-2015
Gutenberg [123]	-	Books	Dec-2021
C4 [73]	800GB	CommonCrawl	Apr-2019
CC-Stories-R [124]	31GB	CommonCrawl	Sep-2019
CC-NEWS [27]	78GB	CommonCrawl	Feb-2019
REALNEWs [125]	120GB	CommonCrawl	Apr-2019
OpenWebText [126]	38GB	Reddit links	Mar-2023
Pushift.io [127]	2TB	Reddit links	Mar-2023
Wikipedia [128]	21GB	Wikipedia	Mar-2023
BigQuery [129]	-	Codes	Mar-2023
the Pile [130]	800GB	Other	Dec-2020
ROOTS [131]	1.6TB	Other	Jun-2022

Large Language Models

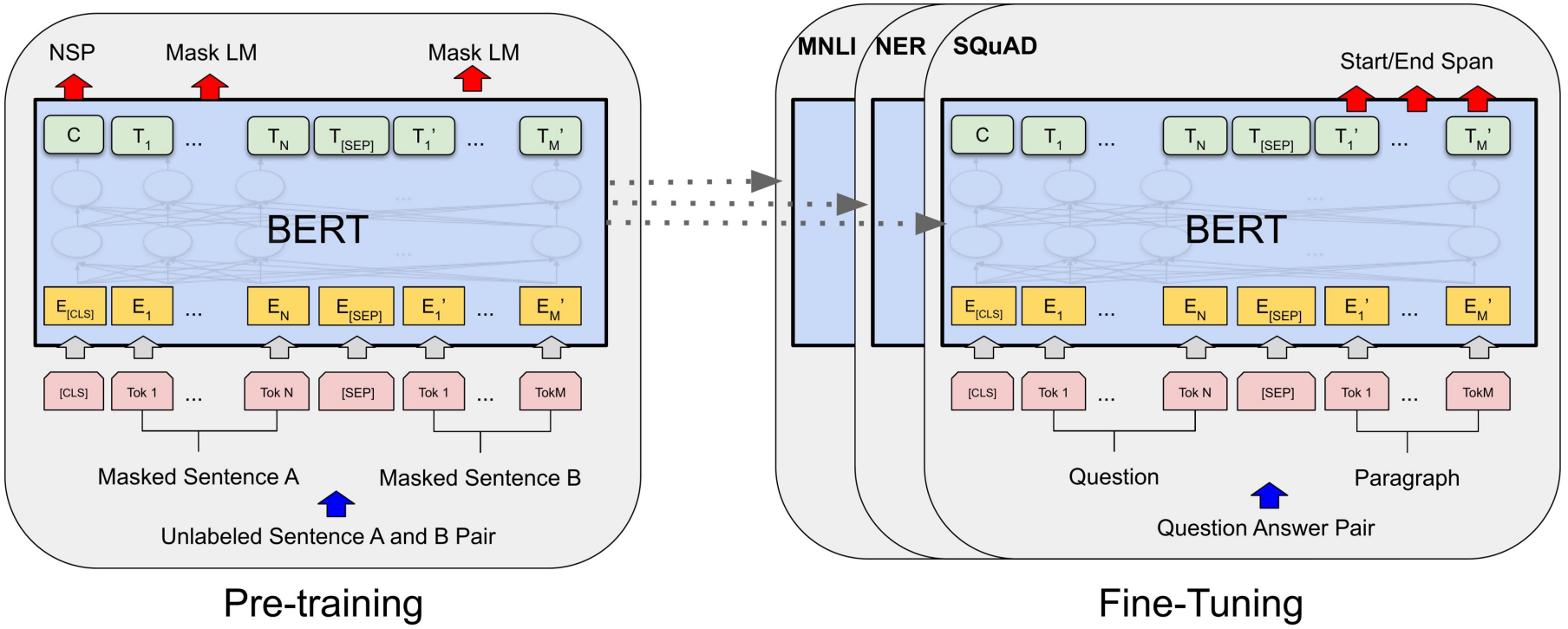
Large Language Models



Large Language Models

Large Language Models

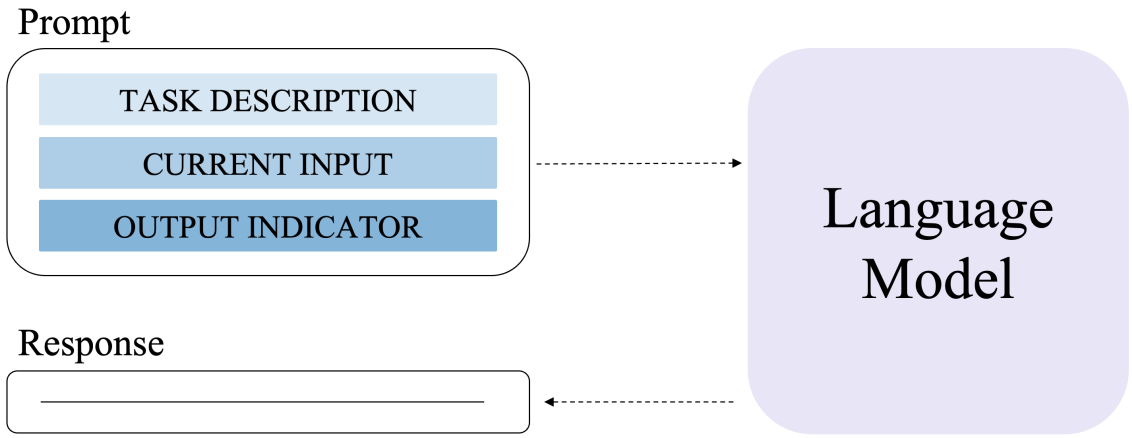
- Pre-training: trained on huge amounts of unlabeled text using “self-supervised” training objective
- Adaptation: how to use a pre-trained model for downstream task?



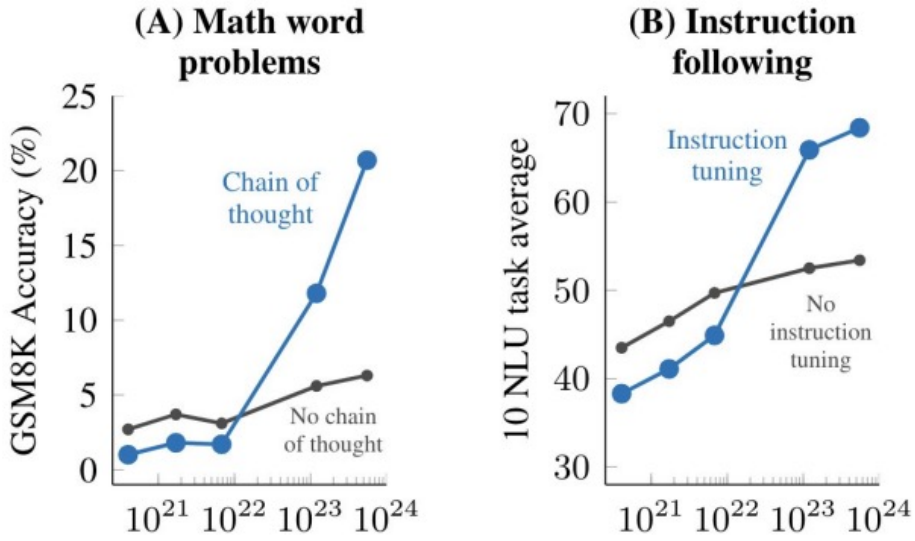
Large Language Models

Large Language Models

- The promise: one single model to solve many NLP tasks



- Emergent properties in LLMs



Large Language Models

Prompts

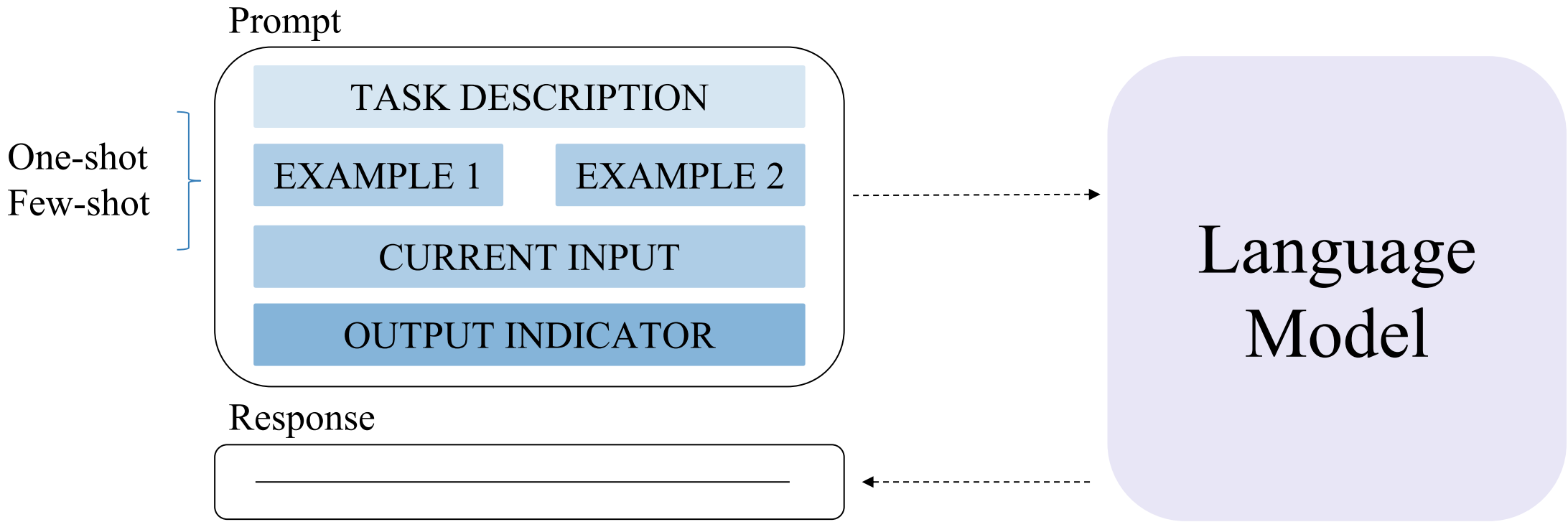
- Prompts involve instructions and context passed to a language model to achieve a desired task
- Prompt engineering is the practice of developing and optimizing prompts to efficiently use language models (LMs) for a variety of applications

```
What is prompt engineering?

Prompt engineering is a process of creating a set of prompts,
or questions, that are used to guide the user toward a desired
outcome. It is an effective tool for designers to create user
experiences that are easy to use and intuitive. This method is
often used in interactive design and software development, as
it allows users to easily understand how to interact with a
system or product..
```

Large Language Models

Elements of Prompts



Large Language Models

Three setting for In-Context-Learning

Few-shot

In addition to the task description, the model sees a few examples of the task. No gradient updates are performed.

```
1  Translate English to French:        ← task description
2  sea otter => loutre de mer          ← examples
3  peppermint => menthe poivrée        ←
4  plush girafe => girafe peluche      ←
5  cheese =>                           ← prompt
```

Zero-shot

The model predicts the answer given only a natural language description of the task. No gradient updates are performed.

```
1  Translate English to French:        ← task description
2  cheese =>                           ← prompt
```

One-shot

In addition to the task description, the model sees a single example of the task. No gradient updates are performed.

```
1  Translate English to French:        ← task description
2  sea otter => loutre de mer          ← example
3  cheese =>                           ← prompt
```

Outline

- Large Language Models
- InstructGPT (RLHF)
- Text Summarization using RLHF

InstructGPT (RLHF)

Overview

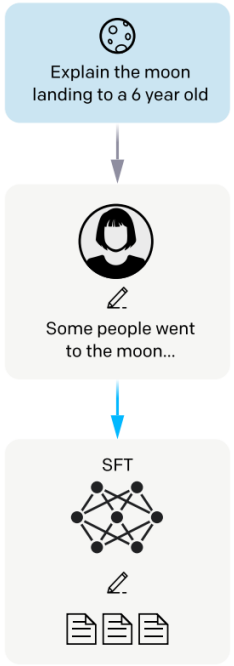
Step 1

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.



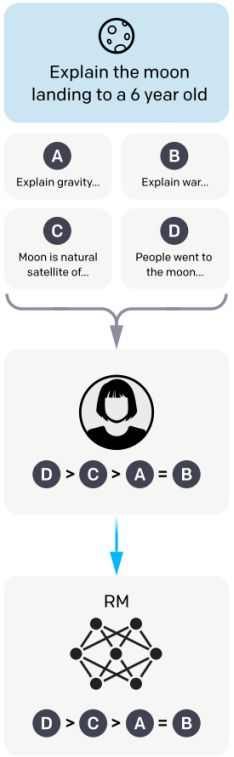
Step 2

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

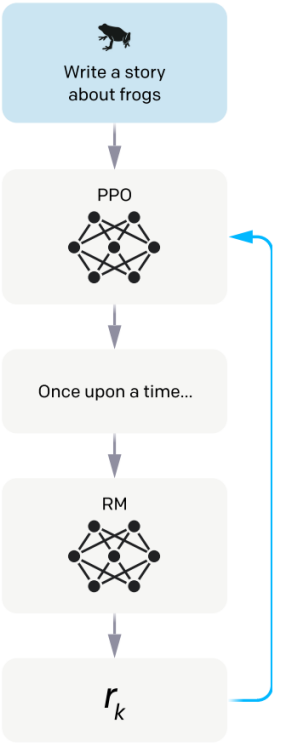
Optimize a policy against the reward model using reinforcement learning.

A new prompt is sampled from the dataset.

The policy generates an output.

The reward model calculates a reward for the output.

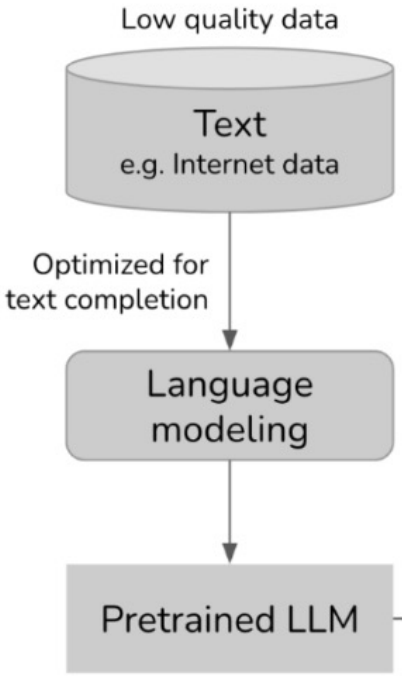
The reward is used to update the policy using PPO.



InstructGPT (RLHF)

Pre-training LLMs

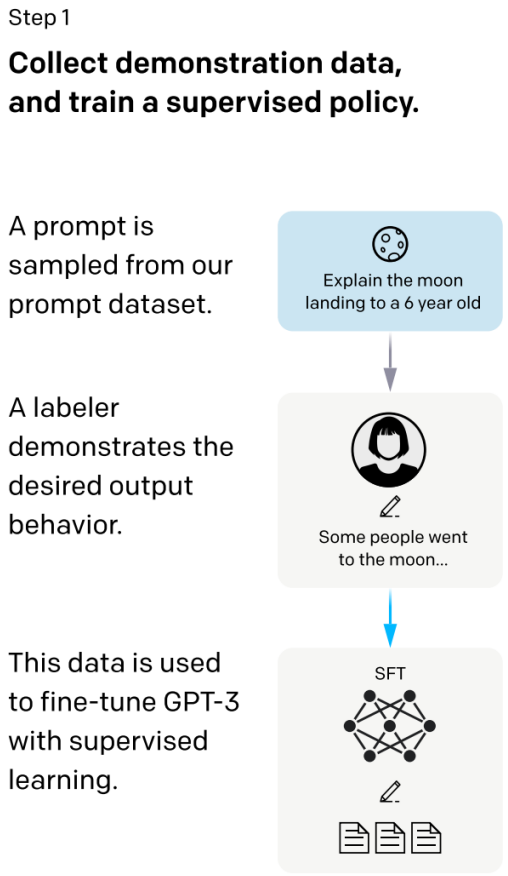
- LLM_φ: the language model being trained, parameterized by φ. The goal is to find φ for which the cross entropy loss is minimized.
- $[T_1, T_2, \ldots, T_V]$: vocabulary – the set of all unique tokens in the training data.
- V: the vocabulary size.
- $f(x)$: function mapping a token to its position in the vocab. If x is T_k in the vocab, $f(x) = k$.
- Given the sequence $(x_1, x_2, \ldots, x_n)$, we'll have n training samples:
  - Input: $x = (x_1, x_2, \ldots, x_{i-1})$
  - Ground truth: x_i
- For each training sample (x, x_i):
  - Let $k = f(x_i)$
  - Model's output: $LLM(x) = [\bar{y}_1, \bar{y}_2, \ldots, \bar{y}_V]$. Note: $\sum_j \bar{y}_j = 1$
  - The loss value: $CE(x, x_i; \varphi) = -\log \bar{y}_k$
- Goal: find φ to minimize the expected loss on all training samples. $CE(\varphi) = -E_x \log \bar{y}_k$



Scale May '23	**>1 trillion** tokens
Examples **Bolded**: open sourced	GPT-x, Gopher, **LLaMa** **Pythia**, **Bloom**, **StableL**

InstructGPT (RLHF)

Supervised Fine-Tuning (SFT)

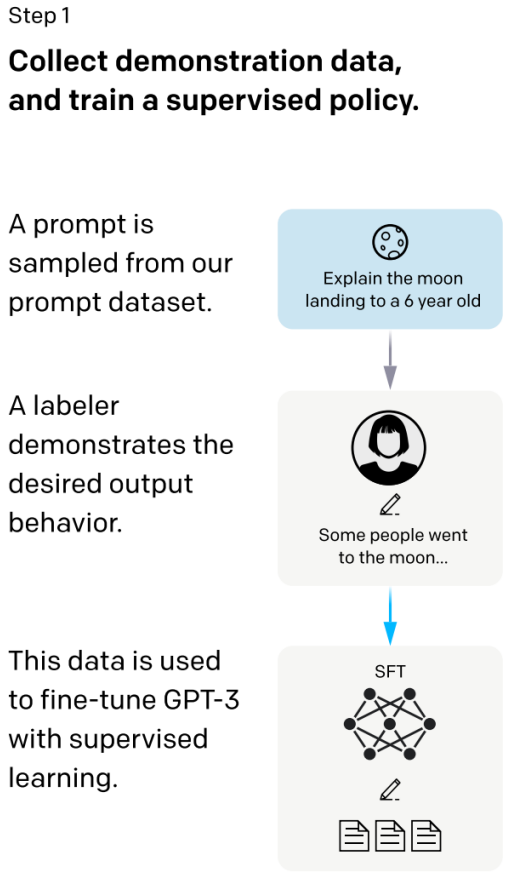


- Goal: optimize the LLM to generate the response that users are looking for

Prompt	Response
Serendipity means the occurrence and development of events by chance in a happy or beneficial way. Use the word in a sentence.	Running into Margaret and being introduced to Tom was a fortunate stroke of serendipity.
ELI5: What's the cause of the "anxiety lump" in our chest during stressful or disheartening experiences?	The anxiety lump in your throat is caused by muscular tension keeping your glottis dilated to maximize airflow. The clenched chest or heartache feeling is caused by the vagus nerve which tells the organs to pump blood faster, stop digesting, and produce adrenaline and cortisol.

InstructGPT (RLHF)

Supervised Fine-Tuning (SFT)

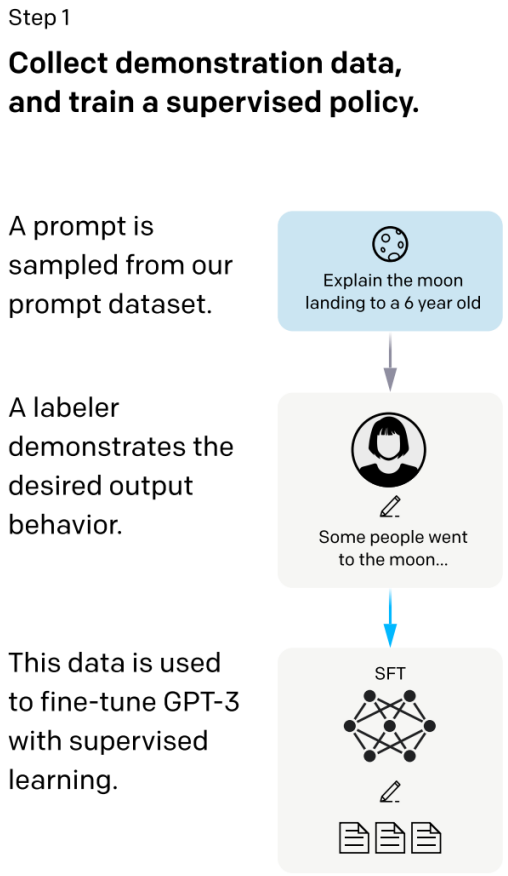


- Goal: optimize the LLM to generate the response that users are looking for

Prompt	Response
Serendipity means the occurrence and development of events by chance in a happy or beneficial way. Use the word in a sentence.	Running into Margaret and being introduced to Tom was a fortunate stroke of serendipity.
ELI5: What's the cause of the "anxiety lump" in our chest during stressful or disheartening experiences?	The anxiety lump in your throat is caused by muscular tension keeping your glottis dilated to maximize airflow. The clenched chest or heartache feeling is caused by the vagus nerve which tells the organs to pump blood faster, stop digesting, and produce adrenaline and cortisol.

InstructGPT (RLHF)

Supervised Fine-Tuning (SFT)

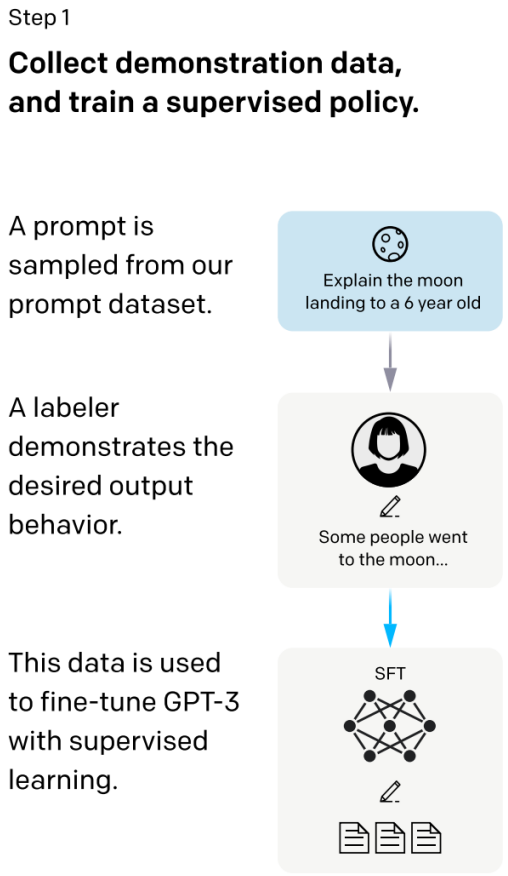


- A large collections of prompts:
  - Labeler-written prompts

Plain	Labelers to come up with an arbitrary task, while ensuring diversity of tasks
Few-shot	Labelers to come up with an instruction and multiple query/response pairs for that instruction "Given the sentiment for a tweet"
User-based	Collect use-cases stated in applications to the OpenAI API. Labelers to come up with prompts correspponding to these use-cases

InstructGPT (RLHF)

Supervised Fine-Tuning (SFT)

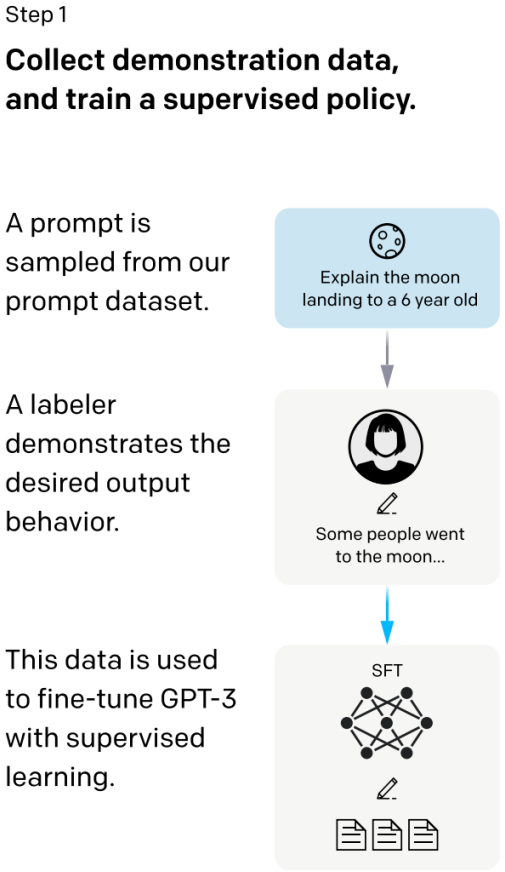


- A large collections of prompts:
  - Labeler-written prompts
  - API user prompts (From OpenAI GPT3 Playground)

- 200 prompts / per organization
- 10 use cases

InstructGPT (RLHF)

Supervised Fine-Tuning (SFT)



- A large collections of prompts:

Table 1: Distribution of use case categories from our API prompt dataset.

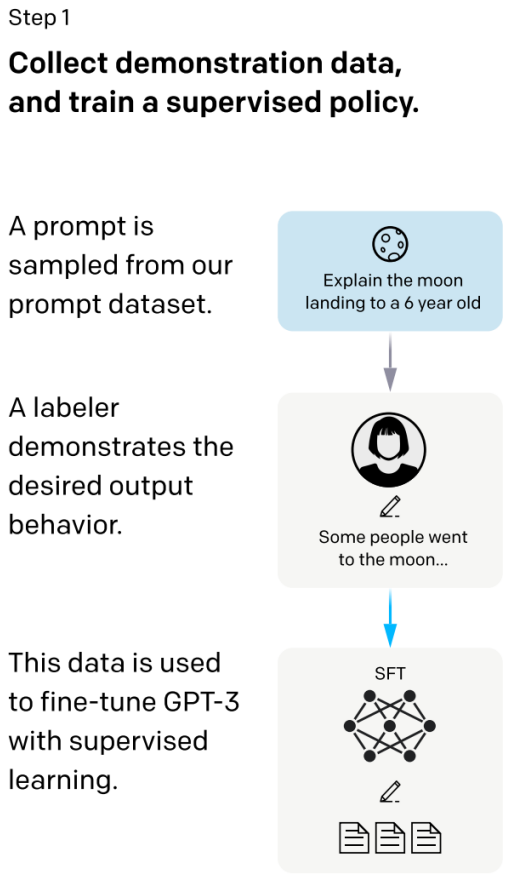
Use-case	(%)
Generation	45.6%
Open QA	12.4%
Brainstorming	11.2%
Chat	8.4%
Rewrite	6.6%
Summarization	4.2%
Classification	3.5%
Other	3.5%
Closed QA	2.6%
Extract	1.9%

Table 2: Illustrative prompts from our API prompt dataset. These are fictional examples inspired by real usage—see more examples in Appendix A.2.1.

Use-case	Prompt
Brainstorming	List five ideas for how to regain enthusiasm for my career
Generation	Write a short story where a bear goes to the beach, makes friends with a seal, and then returns home.
Rewrite	This is the summary of a Broadway play: """ {summary} """ This is the outline of the commercial for that play: """

InstructGPT (RLHF)

Supervised Fine-Tuning (SFT)

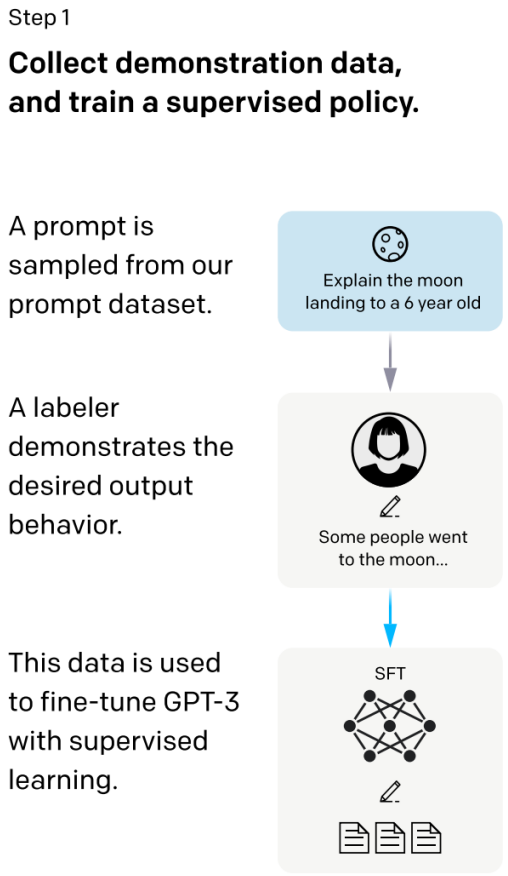


- A large collections of prompts:

SFT Data		
split	source	size
train	labeler	11,295
train	customer	1,430
valid	labeler	1,550
valid	customer	103

InstructGPT (RLHF)

Supervised Fine-Tuning (SFT)

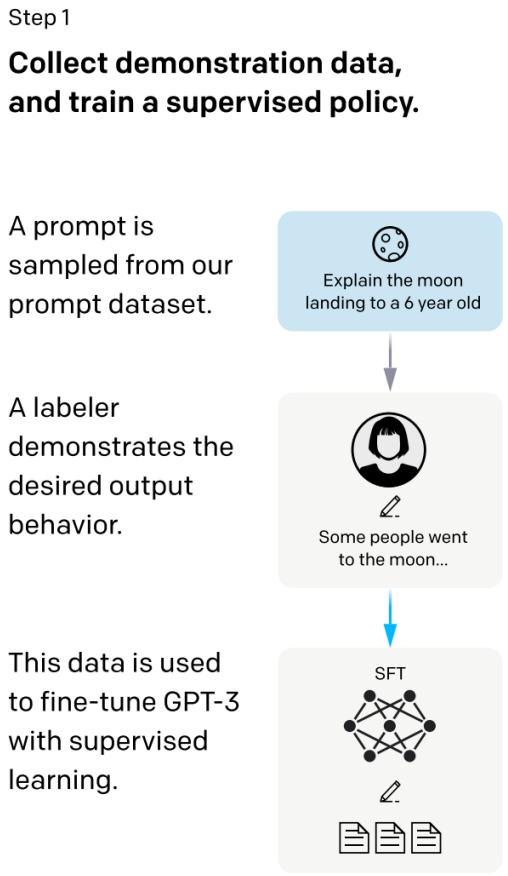


- Fine tune the model, call this model SFT Model
  - Initialized with pretrained GPT3 175B model
  - Trained for 16 epochs on demonstration data
  - Notation:

$$\pi^{SFT}$$

InstructGPT (RLHF)

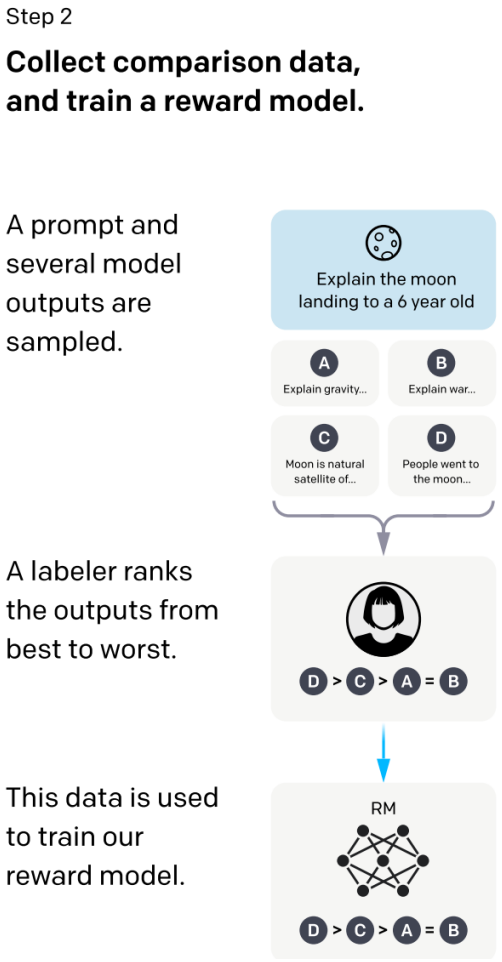
Supervised Fine-Tuning (SFT)



- Task: language modeling
- Training data: high-quality in the format of (prompt, response)
- Data scale: 10,000 – 100,000 (prompt, response) pairs
- Model: LLMs
  - Input: prompt
  - Output: response for this prompt
- Loss: cross entropy

InstructGPT (RLHF)

Reward Modeling



- Training a model to output a score on a given input (a pair of prompt – response)
- A classification or regression task

InstructGPT (RLHF)

Reward Modeling

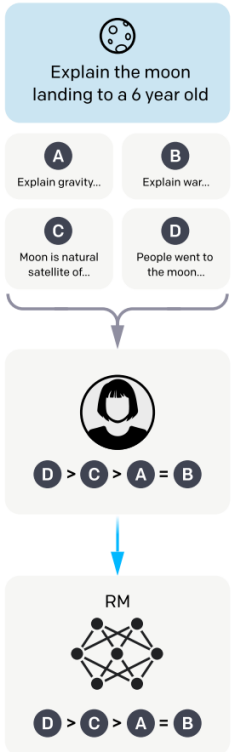
Step 2

Collect comparison data, and train a reward model.

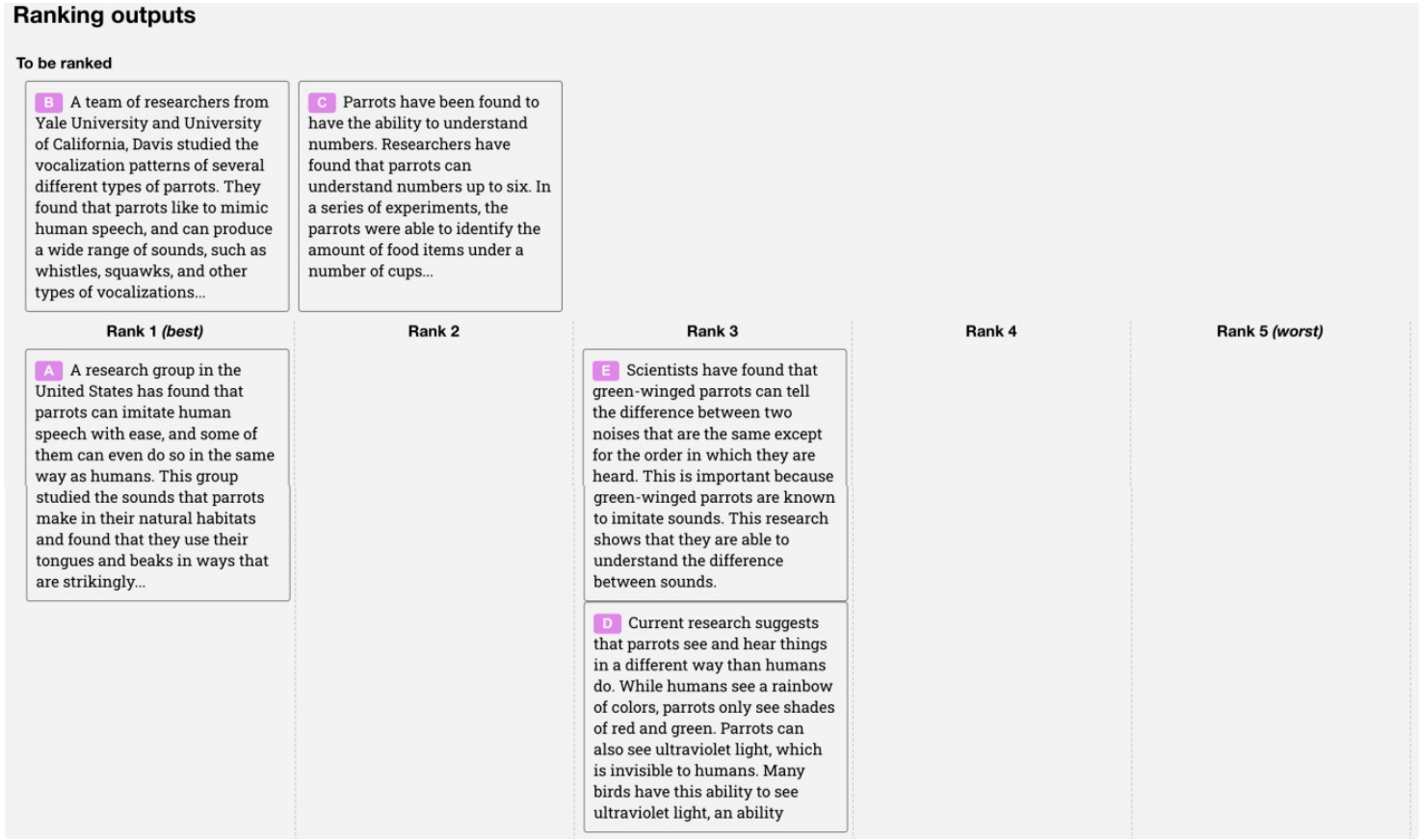
A prompt and several model outputs are sampled.

A labeler ranks the outputs from best to worst.

This data is used to train our reward model.

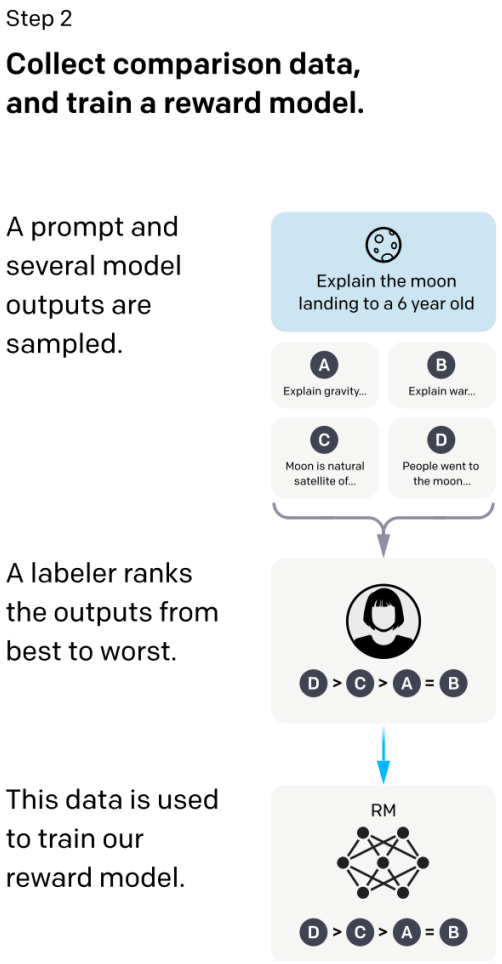


- Given K = 4 to 9 outputs to rank for each prompt



InstructGPT (RLHF)

Reward Modeling

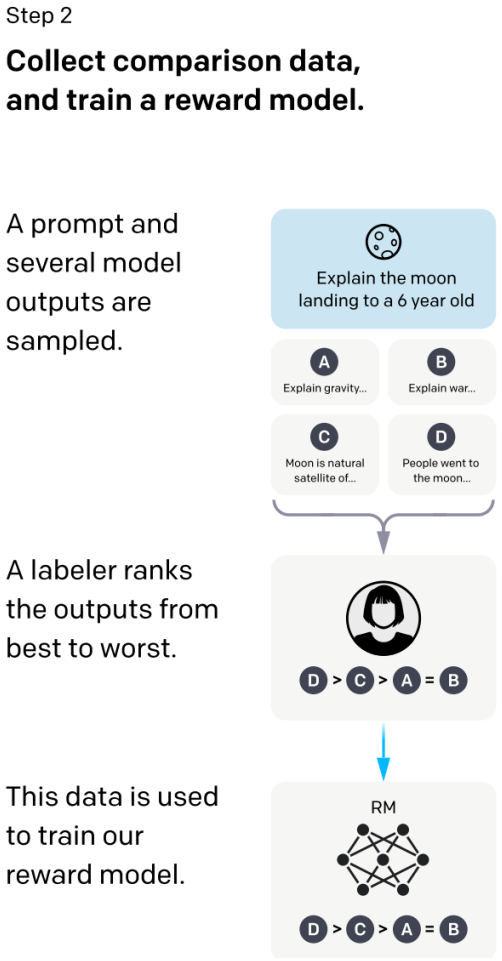


- Given K = 4 to 9 outputs to rank for each prompt
- For 4 ranked responses: D > C > A = B

=> 5 ranked pairs: (D > C), (D > A), (D > B), (C > A), (C > B)

InstructGPT (RLHF)

Reward Modeling



- The reward model: r_θ

x: the prompt, y_w: the better completion, y_l: the worse completion

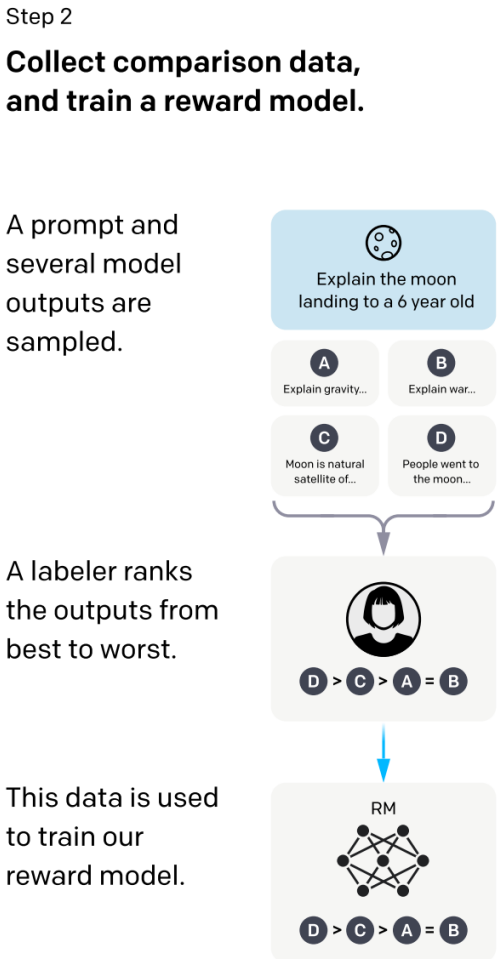
$$\text{loss}(\theta) = \mathrm{E}_{(x,\mathrm{y_w},\mathrm{y_l})\sim D}\left[\log\left(\sigma\left(\mathrm{r}_\theta(\mathrm{x},\mathrm{y_w}) - \mathrm{r}_\theta(\mathrm{x},\mathrm{y_l})\right)\right)\right]$$

Reward on better completion

Reward on worse completion

InstructGPT (RLHF)

Reward Modeling



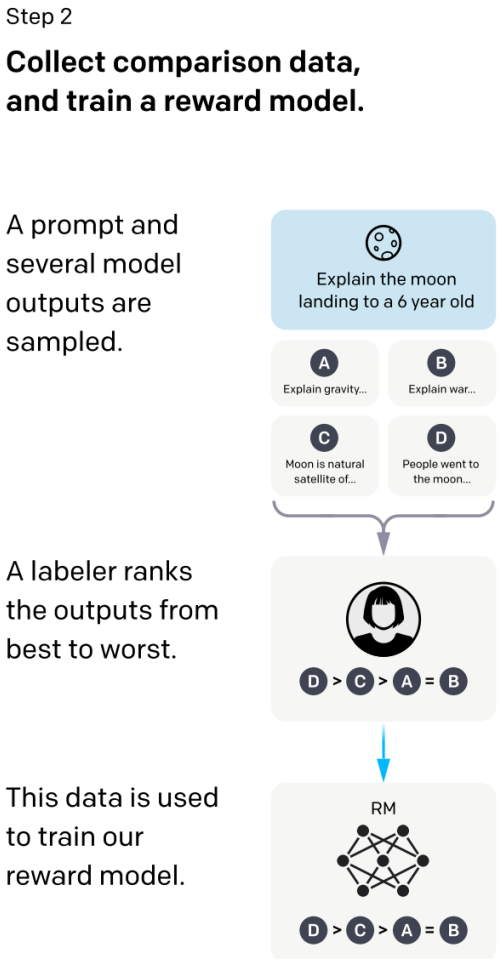
- The reward model: r_θ
- Overfitting problem

Each prompt has K completions => K choose 2 pairs to compare

Each completion can appear in K-1 gradient updates

InstructGPT (RLHF)

Reward Modeling



- The reward model: r_θ
- Overfitting problem

Each prompt has K completions => K choose 2 pairs to compare

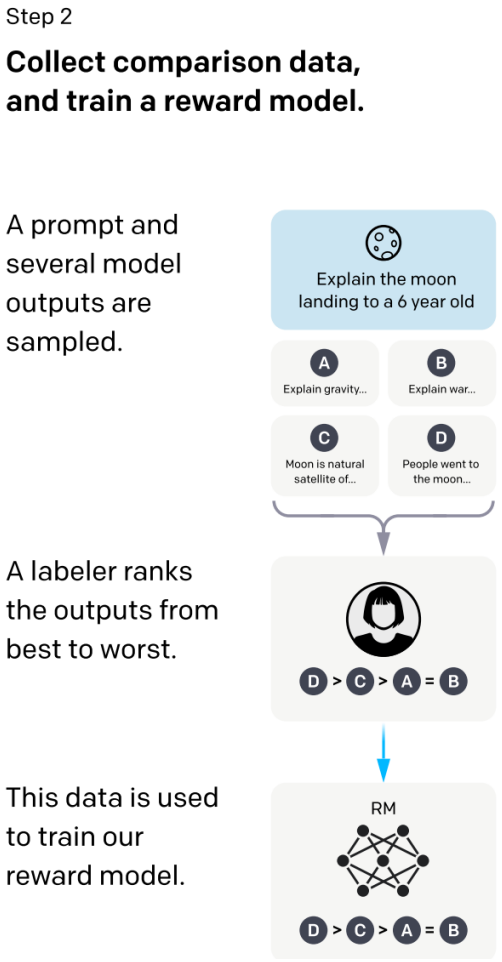
Each completion can appear in K-1 gradient updates

- Solution: train on all comparisons from each prompt as a single batch element
- Normalization in loss with -1/(K choose 2):

$$\text{loss}(\theta) = -\frac{1}{\binom{K}{2}} E_{(x,y_w,y_l)\sim D}\left[\log\left(\sigma\left(r_\theta(x,y_w) - r_\theta(x,y_l)\right)\right)\right]$$

InstructGPT (RLHF)

Reward Modeling



- The reward model: r_θ
- Training data: high-quality data

x: the prompt, y_w: the better completion, y_l: the worse completion

- Data scale: 100K – 1M examples

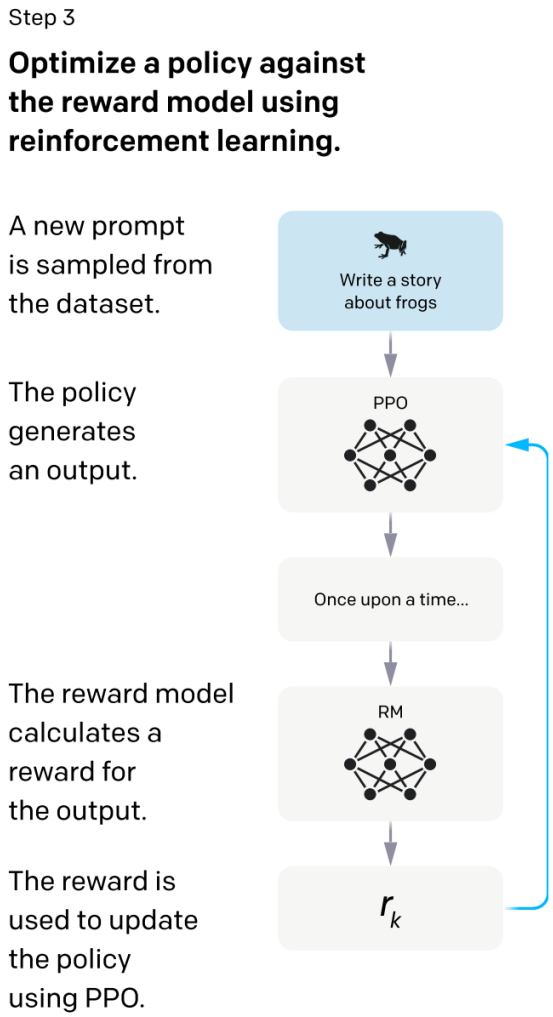
InstructGPT: 50,000 prompts (each prompt: 4 to 9 responses) => 300K to 1.8M training examples

- Training sample (x, y_w, y_l)

$$\text{loss}(\theta) = -\frac{1}{\binom{K}{2}} E_{(x,y_w,y_l)\sim D}\left[\log\left(\sigma\left(r_\theta(x, y_w) - r_\theta(x, y_l)\right)\right)\right]$$

InstructGPT (RLHF)

Reinforcement Learning

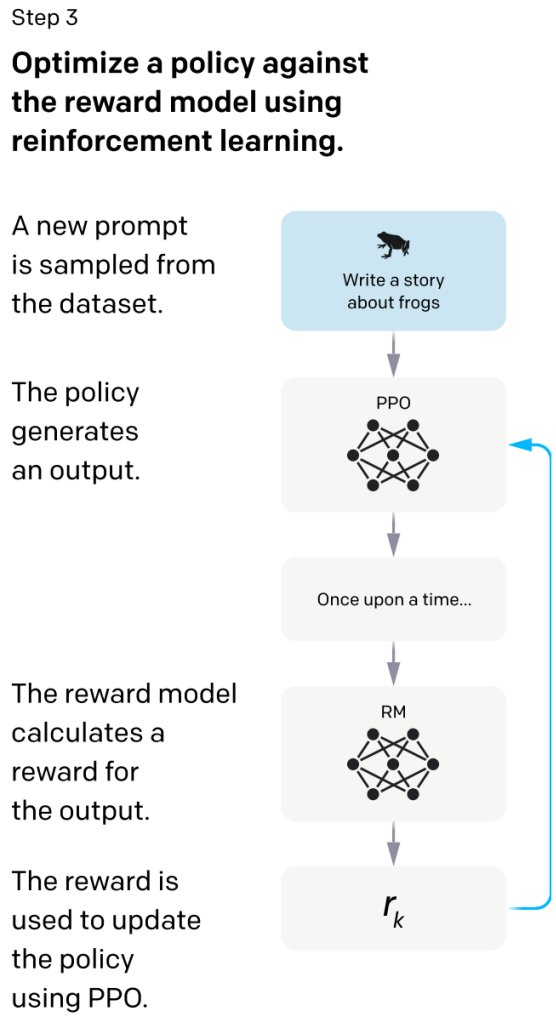


- Goal: train the SFT model to generate output responses that will maximize the scores by the RM model
- Training data: randomly selected prompts
- Data sacle: 10,000 – 100,0000 prompts

PPO Data		
split	source	size
train	customer	31,144
valid	customer	16,185

InstructGPT (RLHF)

Reinforcement Learning

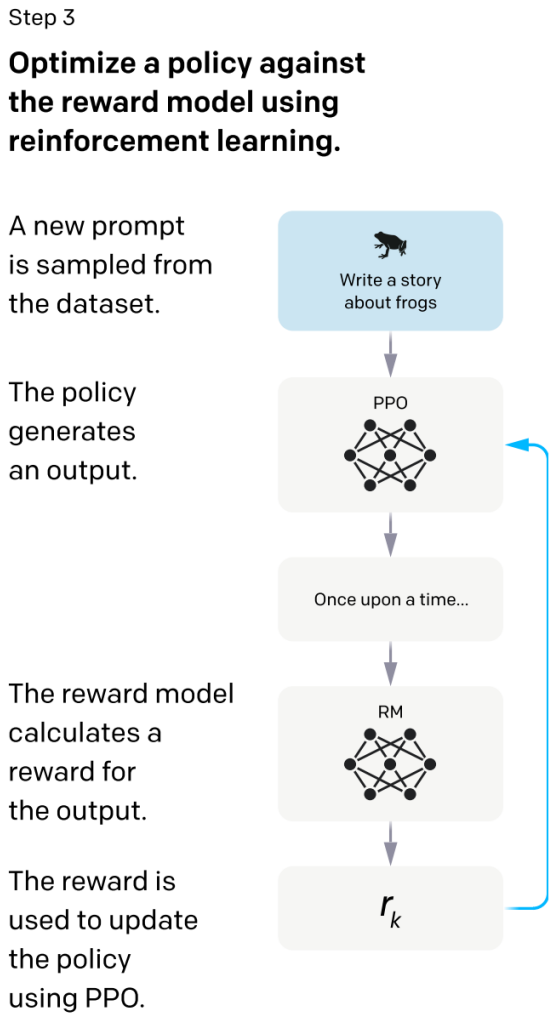


ML Task: Reinforcement Learning

- Action space: the vocabulary of tokens the LLM uses. Taking action means choosing a token to generate
- Observation space: the distribution over all possible prompts
- Policy: the probability distribution over all actions to take (all tokens to generate) given an observation (a prompt)
- Reward function: the reward model from stage 2

InstructGPT (RLHF)

Reinforcement Learning



- D_{RL}: the distribution of prompts used for RL model
- LLM π_ϕ^{RL}: the model being trained with RL, parameterized by ϕ
- For each x from D_{RL}: y: $\pi_\phi^{RL}(x)$

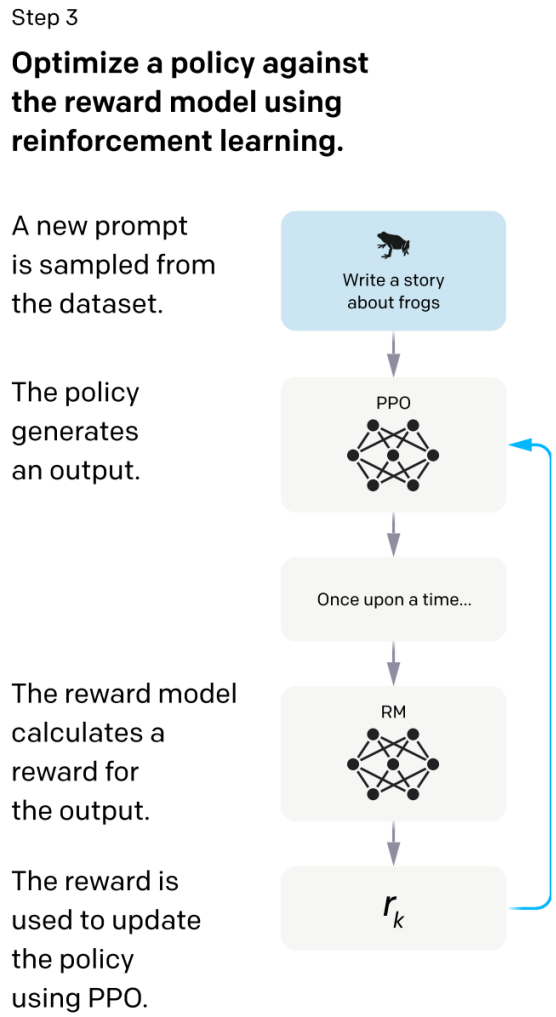
$$\text{objective}_1(x, y; \phi) = \boldsymbol{r_\theta}(x, y)$$

- For all training data DRL

$$\text{objective}_1(x, y; \phi) = E_{(x,y)\sim D_{\pi_\phi}RL}\boldsymbol{r_\theta}(x, y)$$

InstructGPT (RLHF)

Reinforcement Learning

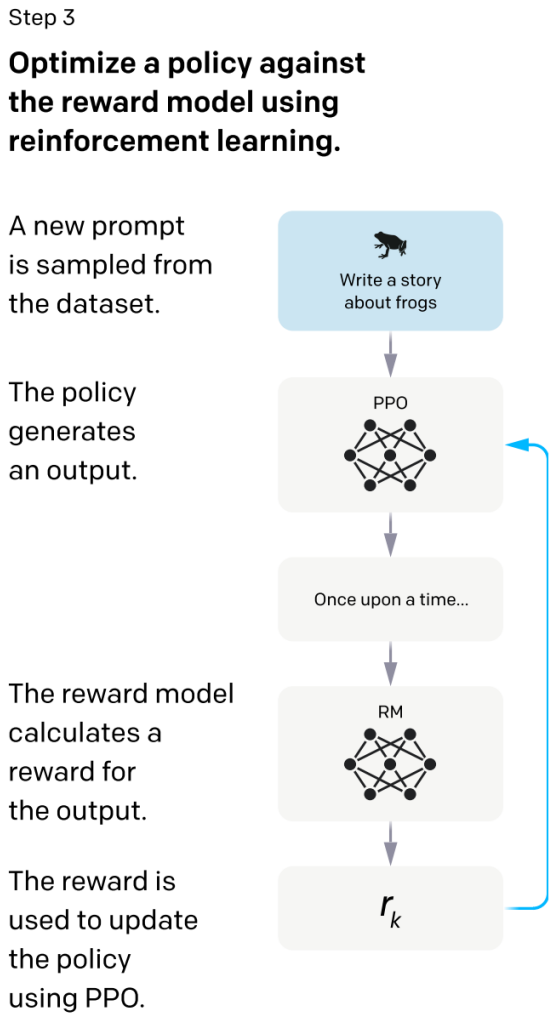


- **Worse reward esimates**: as RLHF is updated, its outputs become very different from what the RM was trained on
- Solution: add a KL penalty that makes sure PPO model output does not deviate too far from SFT model

$$\text{objective}_1(\text{x}, \text{y}; \phi) = \text{E}_{(x,y)\sim \text{D}_{\pi_\phi} RL}\left[\boldsymbol{r_\theta}(\text{x}, \text{y}) - \beta \log \frac{\pi_\phi^{\text{RL}}(y|x)}{\pi^{\text{SFT}}(y|x)}\right]$$

InstructGPT (RLHF)

Reinforcement Learning

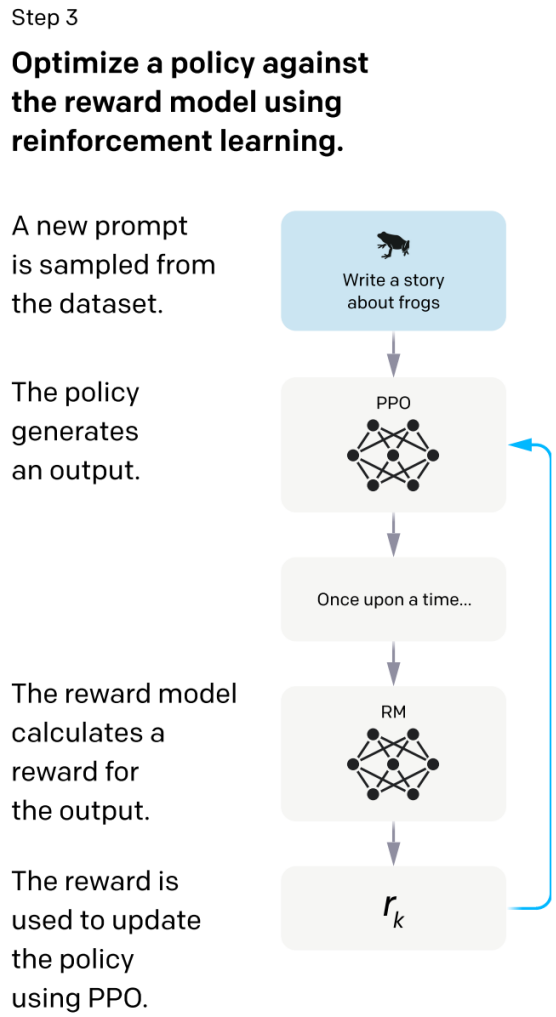


- **Just using RL objective leads to performance degradation on many NLP tasks**
- Solution: add a auxiliary LM ojective on the pretraining data. Call this variant PPO-ptx
- $D_{pretrain}$: the distribution of the pretraining data for the pretrain model

$$\text{objective}_2(x_{pretrain}; \phi) = \gamma E_{x \sim D_{pretrain}} \left[\log \left(\pi_{\phi}^{RL}(x) \right) \right]$$

InstructGPT (RLHF)

Reinforcement Learning



- **Maximize the objective function in RL training:**

$$\text{objective}(\boldsymbol{\phi}) = \text{objective}_1(\text{x}, \text{y}; \phi) + \text{objective}_2\left(\text{x}_{\text{pretrain}}; \phi\right)$$

$$\text{objective}(\boldsymbol{\phi}) = \text{E}_{(x,y)\sim \text{D}_{\pi_\phi} RL}\left[\boldsymbol{r_\theta}(\text{x}, \text{y}) - \beta \log \frac{\pi_\phi^{\text{RL}}(y|x)}{\pi^{\text{SFT}}(y|x)}\right] + \gamma \text{E}_{\text{x}\sim \text{D}_{\text{pretrain}}}\left[\log\left(\boldsymbol{\pi_\phi^{\text{RL}}}(\text{x})\right)\right]$$

InstructGPT (RLHF)

Summary

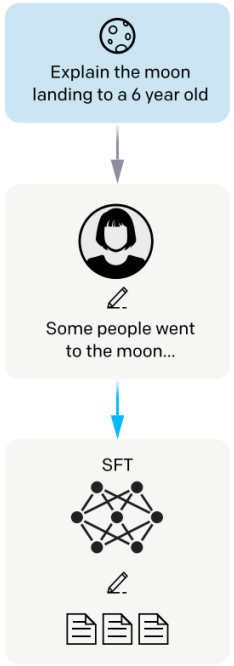
Step 1

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.



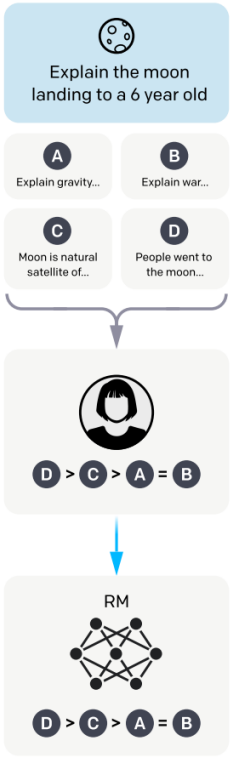
Step 2

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

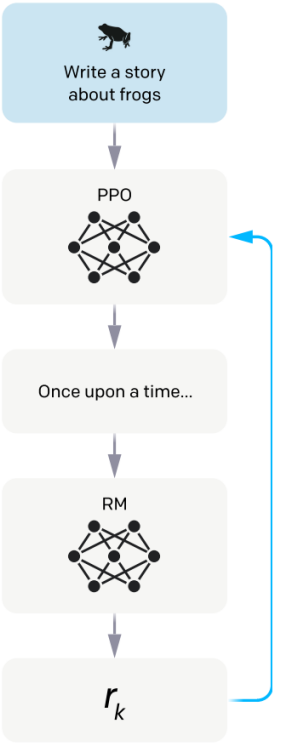
Optimize a policy against the reward model using reinforcement learning.

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.

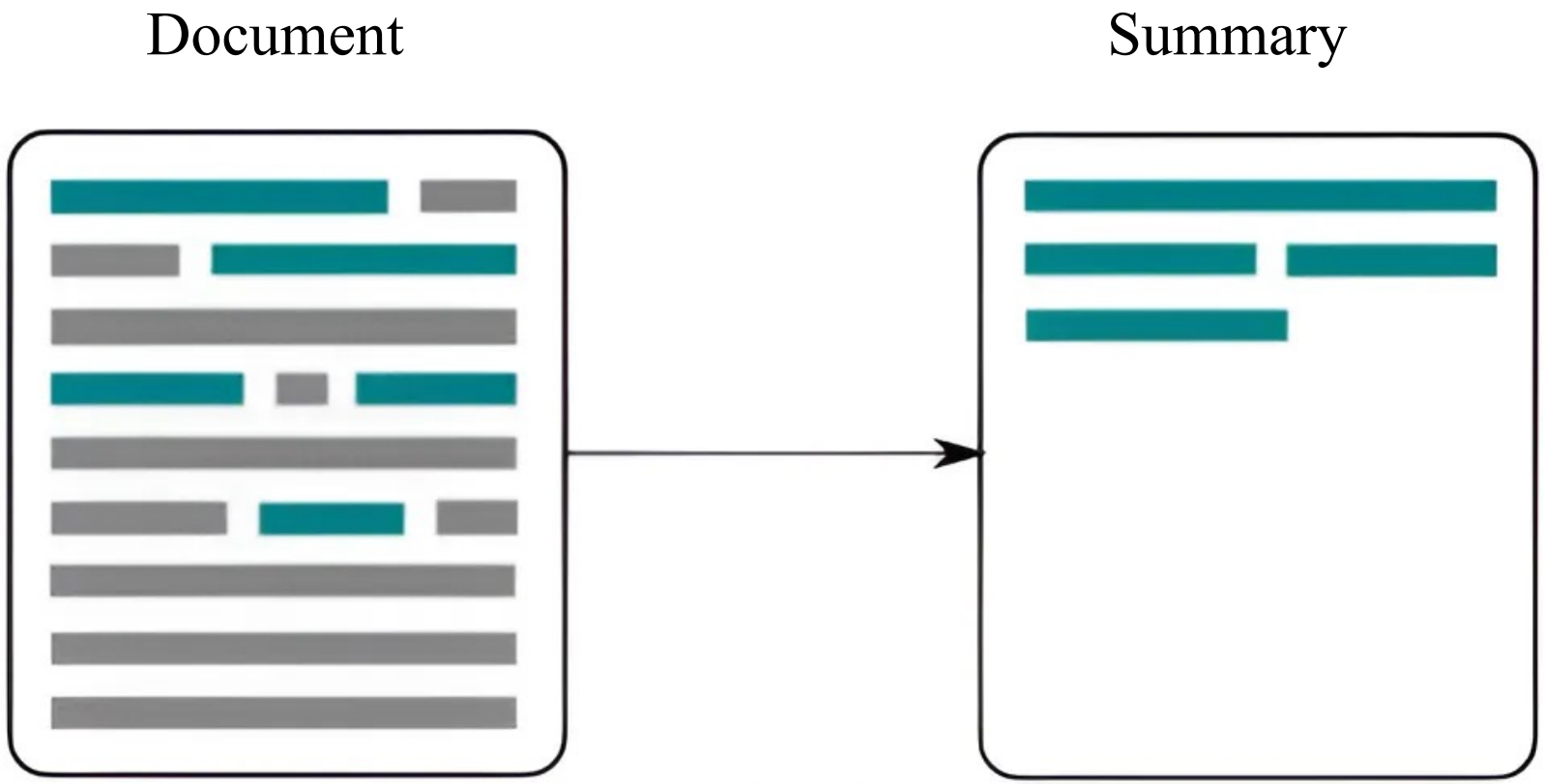


Outline

- Large Language Models
- InstructGPT (RLHF)
- Text Summarization using RLHF

Text Summarization using RLHF

Text Summarization



Text Summarization using RLHF

Text Summarization

SUBREDDIT: r/relationships TITLE: I (f/22) have to figure out if I want to still know these girls or not and would hate to sound insulting POST: Not sure if this belongs here but it's worth a try. Backstory: When I (f/22) went through my first real breakup 2 years ago because he needed space after a year of dating roand it effected me more than I thought. It was a horrible time in my life due to living with my mother and finally having the chance to cut her out of my life. I can admit because of it was an emotional wreck and this guy was stable and didn't know how to deal with me. We ended by him avoiding for a month or so after going to a festival with my friends. When I think back I wish he just ended. So after he ended it added my depression I suffered but my friends helped me through it and I got rid of everything from him along with cutting contact. Now:

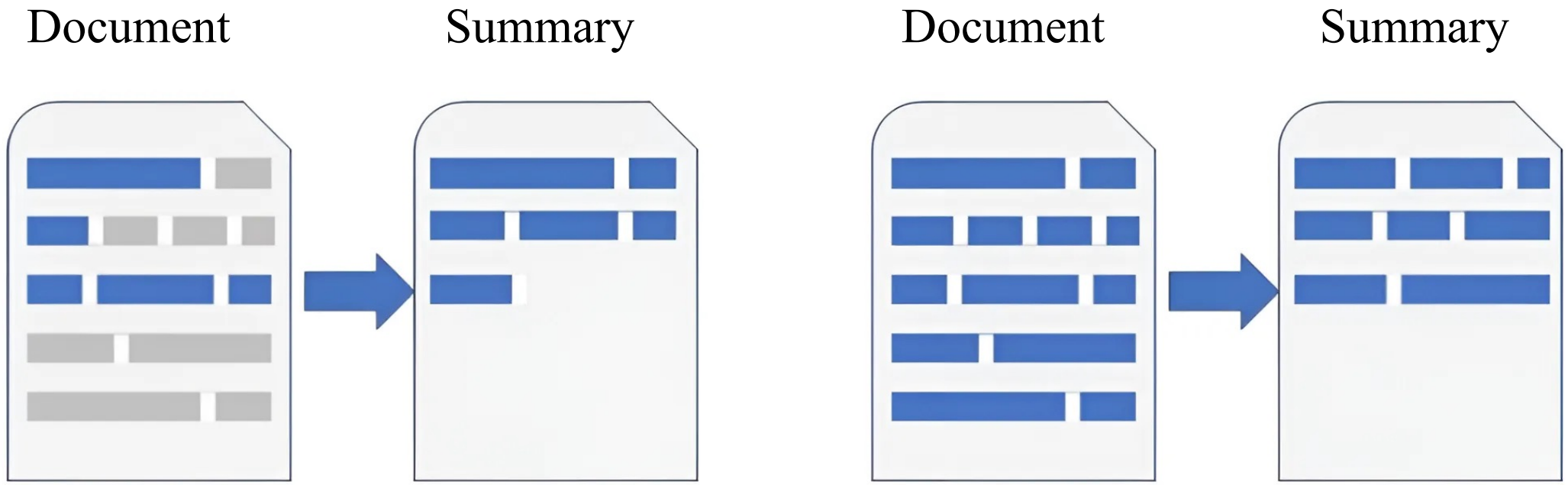
→

I still have contact with an old ex's friends but can't stand to see or talk to him. His friends are really nice ,so how do I tell them I possibly want to unfriend them on Facebook because of him?

Text Summarization using RLHF

Types of Text Summarization

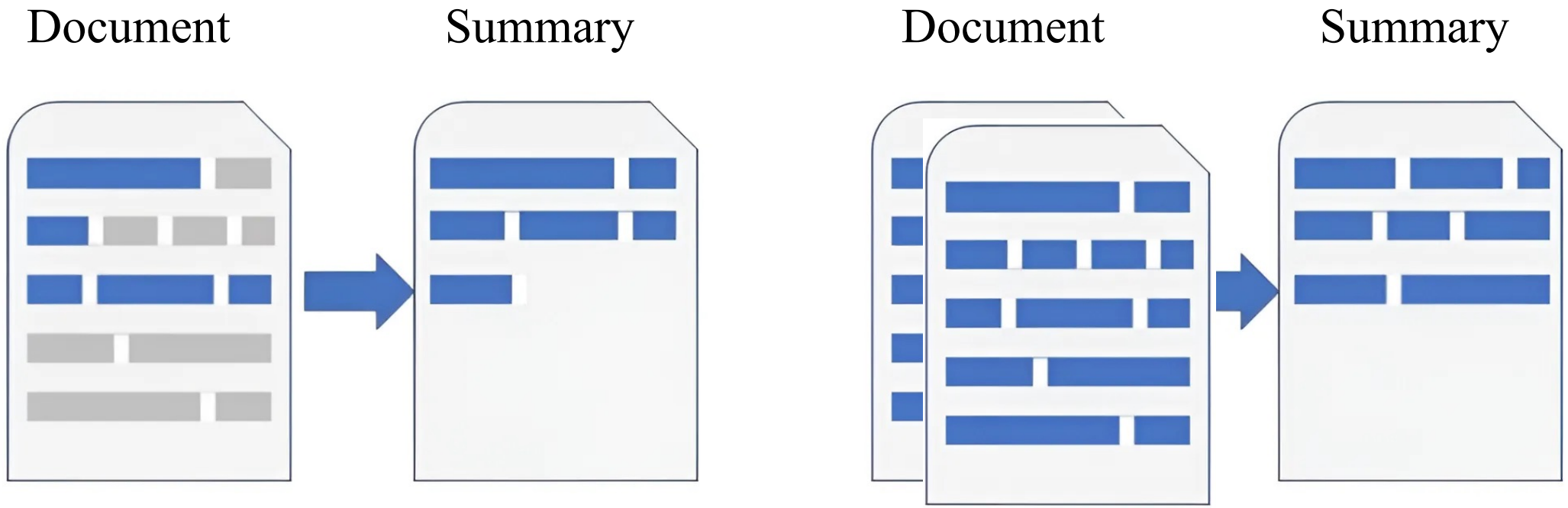
- Based on Output Type



Text Summarization using RLHF

Types of Text Summarization

- Based on Input Type



Text Summarization using RLHF

Dataset

CarperAI/openai_summarize_tldr

prompt string · *lengths* 1.75k↔1.99k 16.1%	label string · *lengths* 196↔217 4.7%
SUBREDDIT: r/relationships TITLE: I (f/22) have to figure out if I want to still know these girls or not and would hate to sound insulting POST: Not sure if this belongs here but it's worth a try. Backstory: When I (f/22) went through my first real breakup 2 years ago because he needed space after a year of dating roand it effected me more than I thought. It was a horrible time in my life due to living with my mother and finally having the chance to cut her out of my life. I can admit because of it was an emotional wreck and this guy was stable and didn't know how to deal with me. We ended by him avoiding for a month or so after going to a festival with my friends. When I think back I wish he just ended. So after he ended it added my depression I suffered but my friends helped me through it and I got rid of everything from him along with cutting contact. Now:	I still have contact with an old ex's friends but can't stand to see or talk to him. His friends are really nice ,so how do I tell them I possibly want to unfriend them on Facebook because of him?

‹ Previous 1 2 3 ... 1,168 Next ›

Text Summarization using RLHF

Dataset

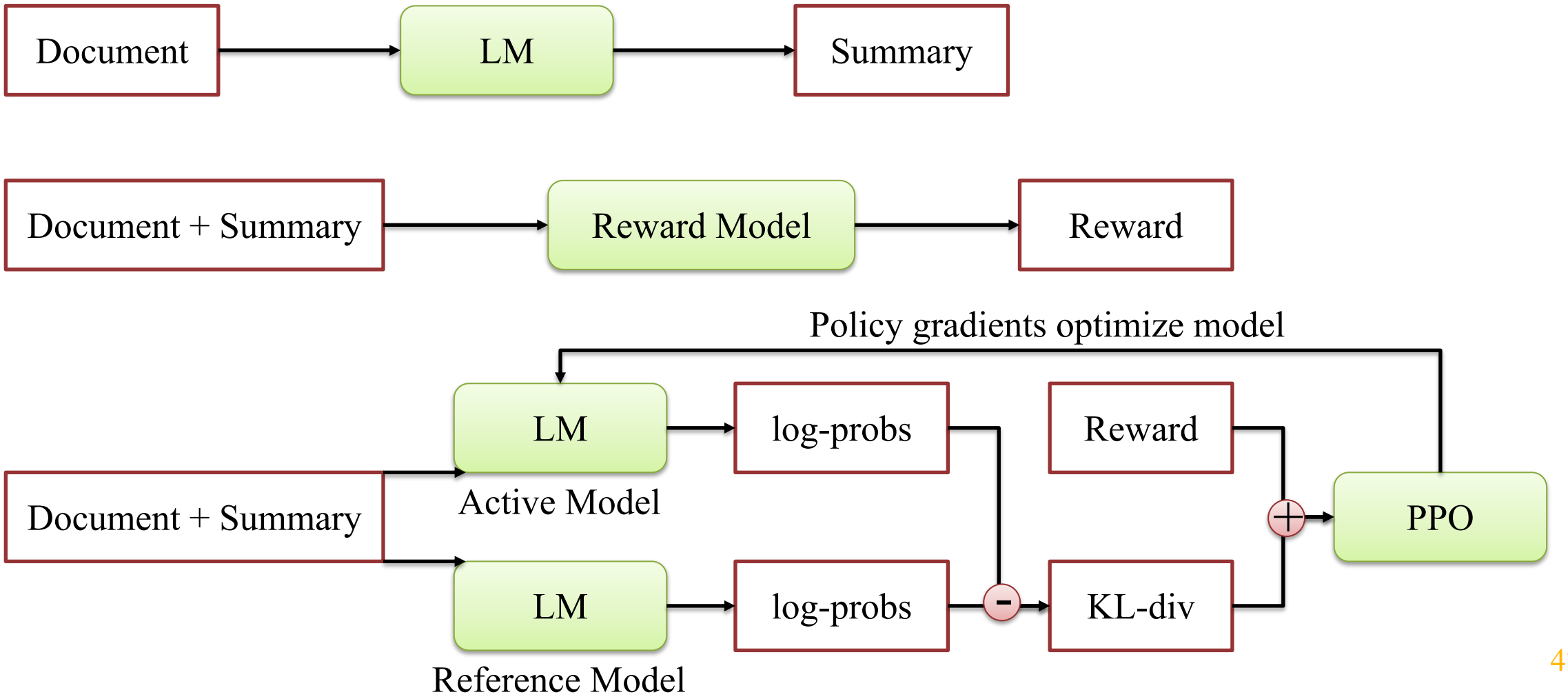
CarperAI/openai_summarize_comparisions

prompt string · *lengths* 1.22k↔1.45k 17.2%	chosen string · *lengths* 141↔198 49.3%	rejected string · *lengths* 138↔195 41.9%
SUBREDDIT: r/relationships TITLE: My [21/M] girlfriend [19/F] broke up with me after she went through my Facebook without my permission. POST: My girlfriend and I had been dating for 15 months. **Last week my girlfriend went onto my Facebook account and read through my message history with a couple of girls.** She was **searching for a specific girl that I used to flirt with in the past, and she found it.** We had fought one time before about me flirting with this girl, and I stopped talking to her entirely for a couple of months (obviously she didn't believe I	TL;DR: My Girlfriend of 15 months went through my Facebook messages without my permission and found old conversations of me flirting with a girl. She broke up with me and went no contact.	TL;DR: My girlfriend and I broke up after she went through my Facebook account without my permission.<|endoftext|>Citizens for the Republic

‹ Previous 1 2 3 ... 837 Next ›

Text Summarization using RLHF

Pipeline



Text Summarization using RLHF

Pipeline – Supervised Fine-Tuning

SUBREDDIT: r/relationships TITLE: I (f/22) have to figure out if I want to still know these girls or not and would hate to sound insulting POST: Not sure if this belongs here but it's worth a try. Backstory: When I (f/22) went through my first real breakup 2 years ago because he needed space after a year of dating roand it effected me more than I thought. It was a horrible time in my life due to living with my mother and finally having the chance to cut her out	I still have contact with an old ex's friends but can't stand to see or talk to him. His friends are really nice ,so how do I tell them I possibly want to unfriend them on Facebook because of him?

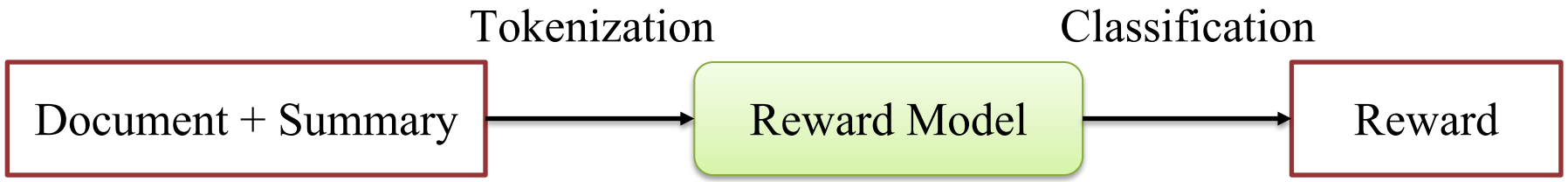
Dataset: CarperAI/openai_summarize_tldr

Task: Text Generation

Metric: ROUGE

Text Summarization using RLHF

Pipeline – Reward Modeling



	Chosen: 1	Rejected: 0
SUBREDDIT: r/relationships TITLE: My [21/M] girlfriend [19/F] broke up with me after she went through my Facebook without my permission. POST: My girlfriend and I had been dating for 15 months. **Last week my girlfriend went onto my Facebook account and read through my message history with a couple	TL;DR: My Girlfriend of 15 months went through my Facebook messages without my permission and found old conversations of me flirting with a girl. She broke up with me and went no contact.	TL;DR: My girlfriend and I broke up after she went through my Facebook account without my permission.<|endoftext|>Citizens for the Republic

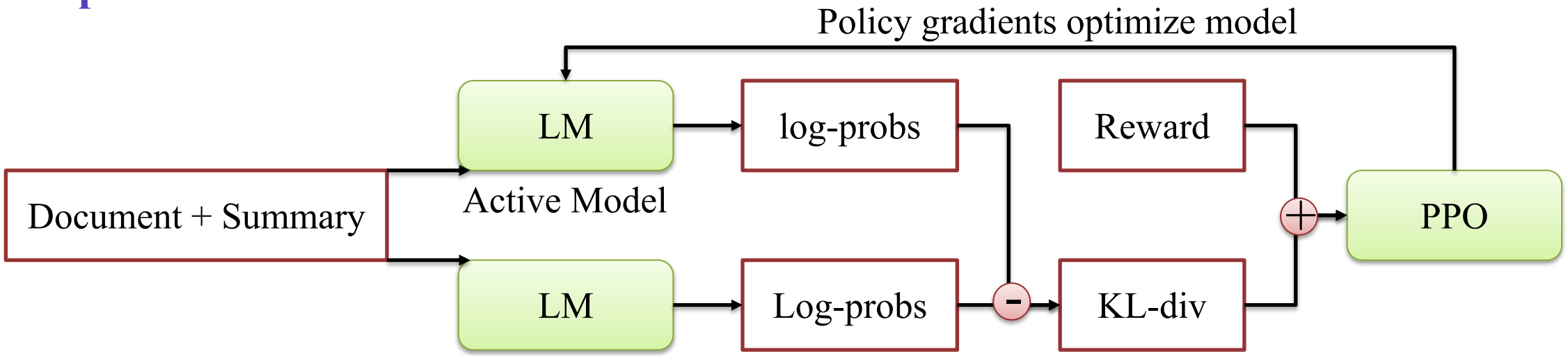
Dataset:CarperAI/openai_summarize_comparisions

Task: Text Classification (Document-Level)

Metric: Accuracy

Text Summarization using RLHF

Pipeline – PPO



SUBREDDIT: r/relationships TITLE: I (f/22) have to figure out if I want to still know these girls or not and would hate to sound insulting POST: Not sure if this belongs here but it's worth a try. Backstory: When I (f/22) went through my first real breakup 2 years ago because he needed space after a year of dating roand it effected me more than I thought. It was a horrible time in my life due to living with my mother and finally having the chance to cut her out	I still have contact with an old ex's friends but can't stand to see or talk to him. His friends are really nice ,so how do I tell them I possibly want to unfriend them on Facebook because of him?

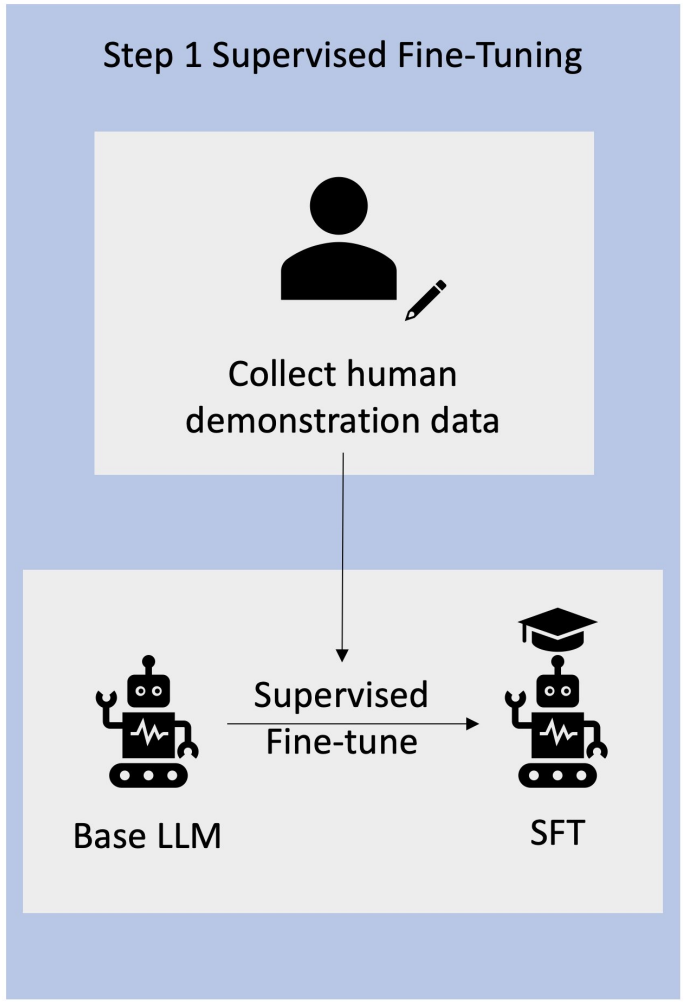
Dataset: CarperAI/openai_summarize_tldr

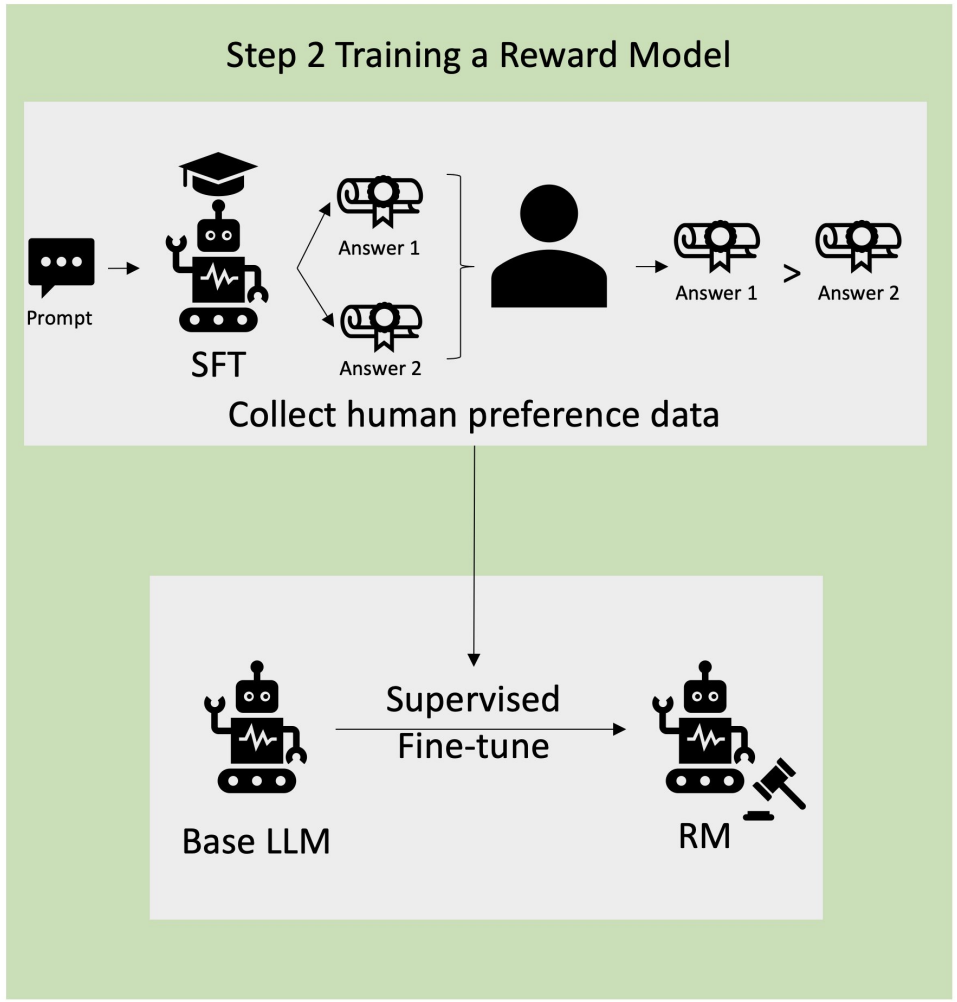
Task: Reinforcement Learning

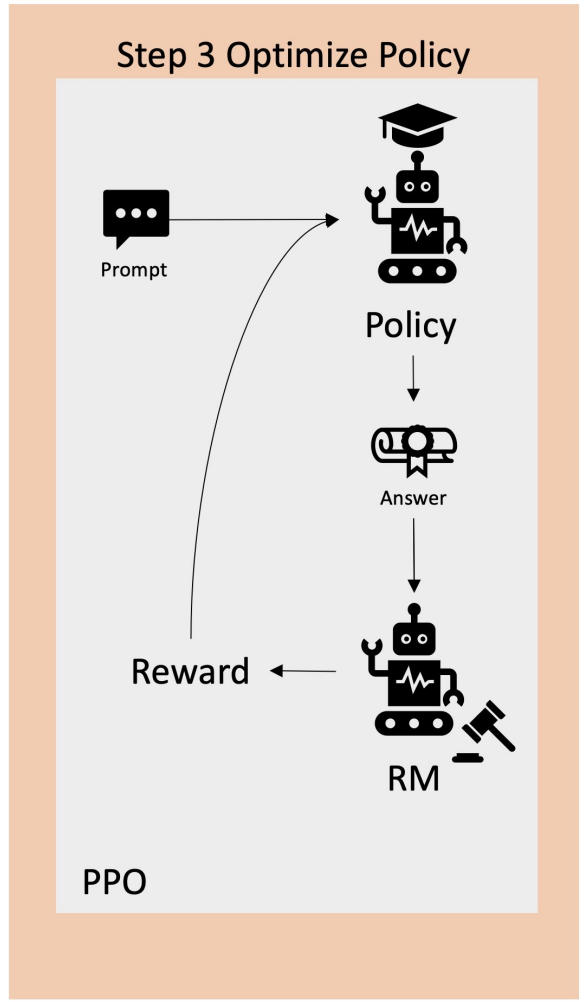
Text Summarization using RLHF

Experiment

Summary


Step 1 Supervised Fine-Tuning
Collect human demonstration data
Supervised Fine-tune
Base LLM
SFT


Step 2 Training a Reward Model
Prompt
SFT
Answer 1
Answer 2
Answer 1
>
Answer 2
Collect human preference data
Supervised Fine-tune
Base LLM
RM


Step 3 Optimize Policy
Prompt
Policy
Answer
Reward
RM
PPO

Thanks!

Any questions?