GPT3 & instruct GPT

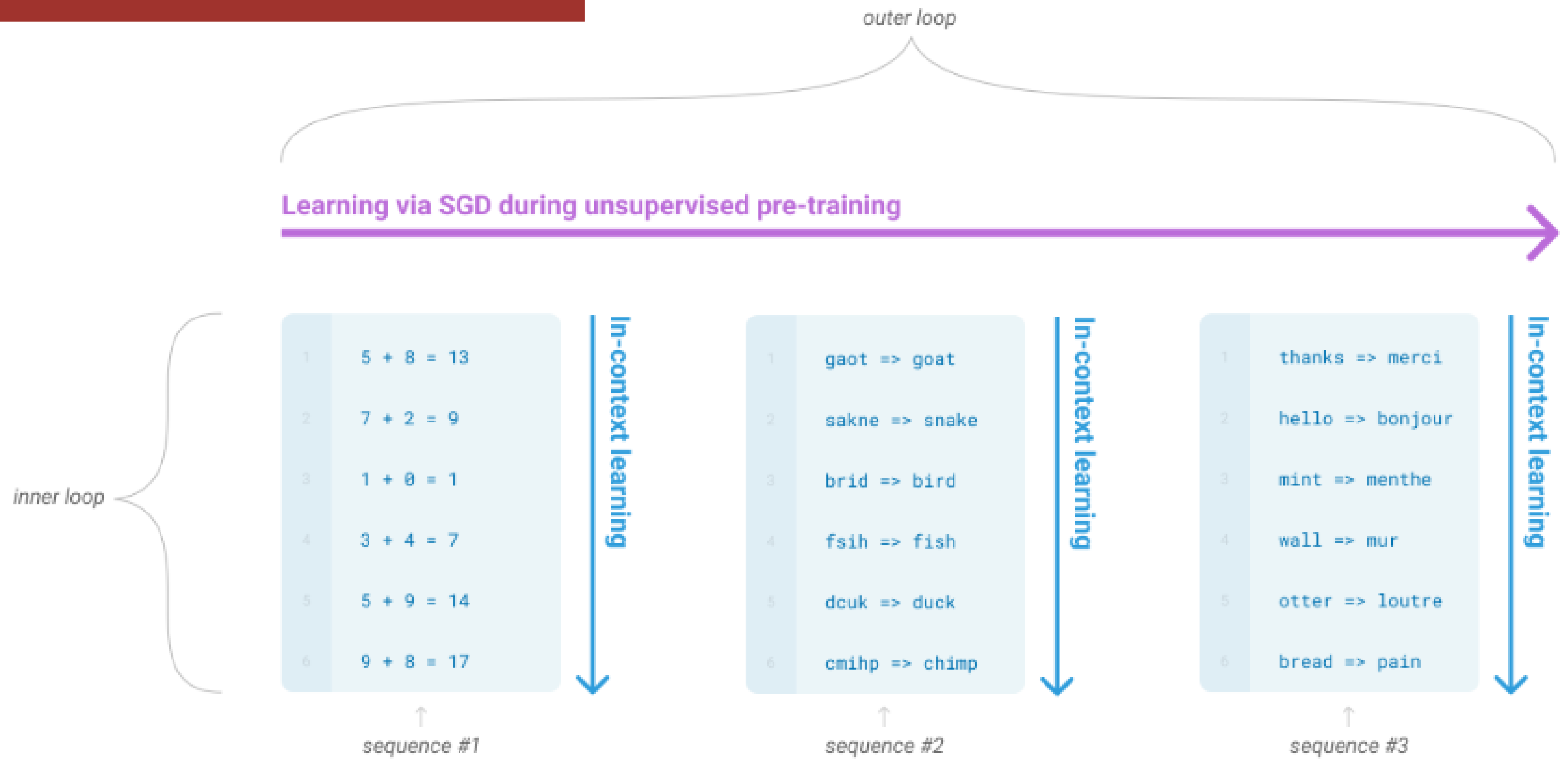
Kang Byeon Jin

Study Group NLP

TRAIN AND TEST

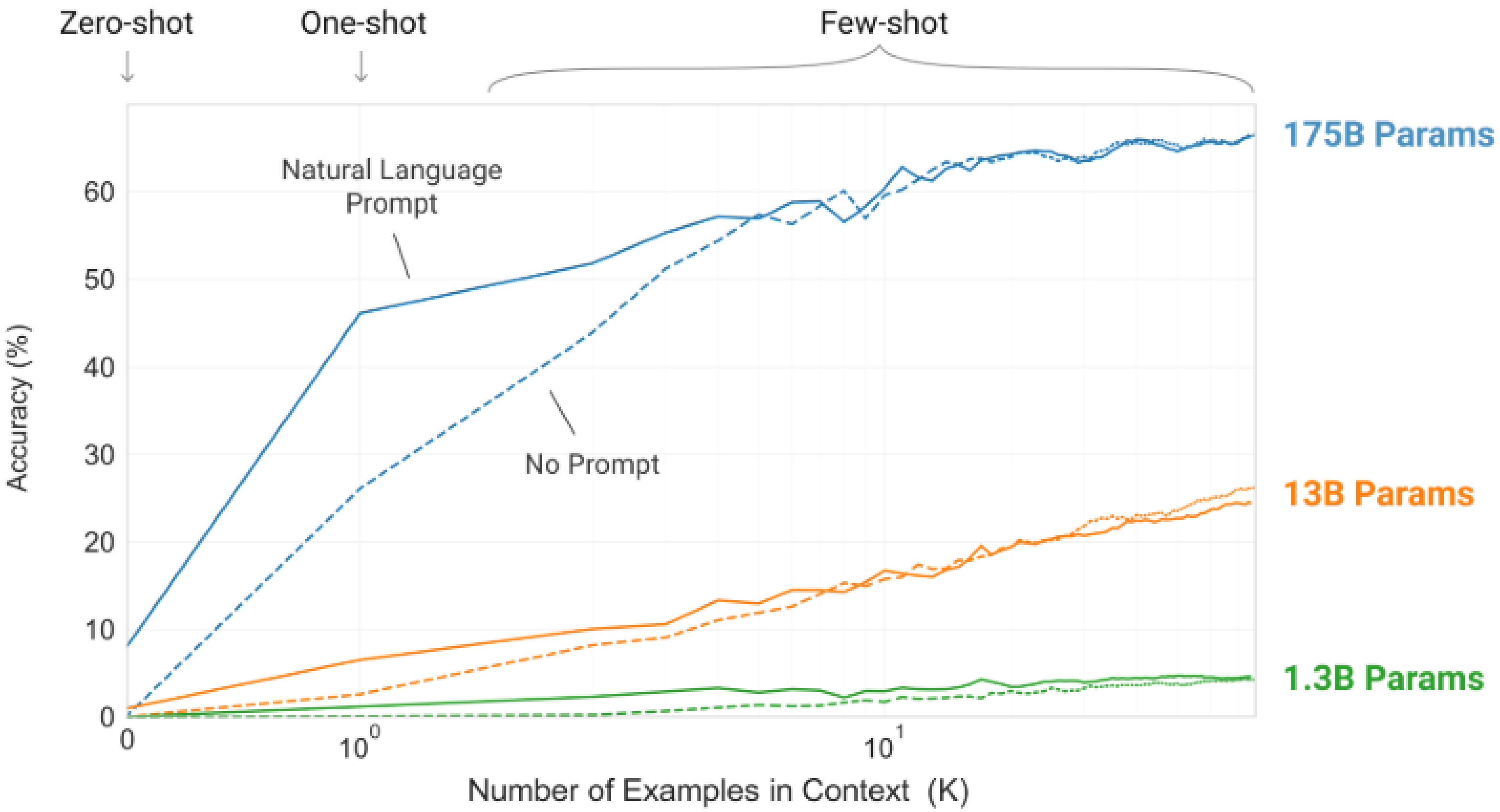
Language Models are Few-Shot Learners

Language model meta-learning



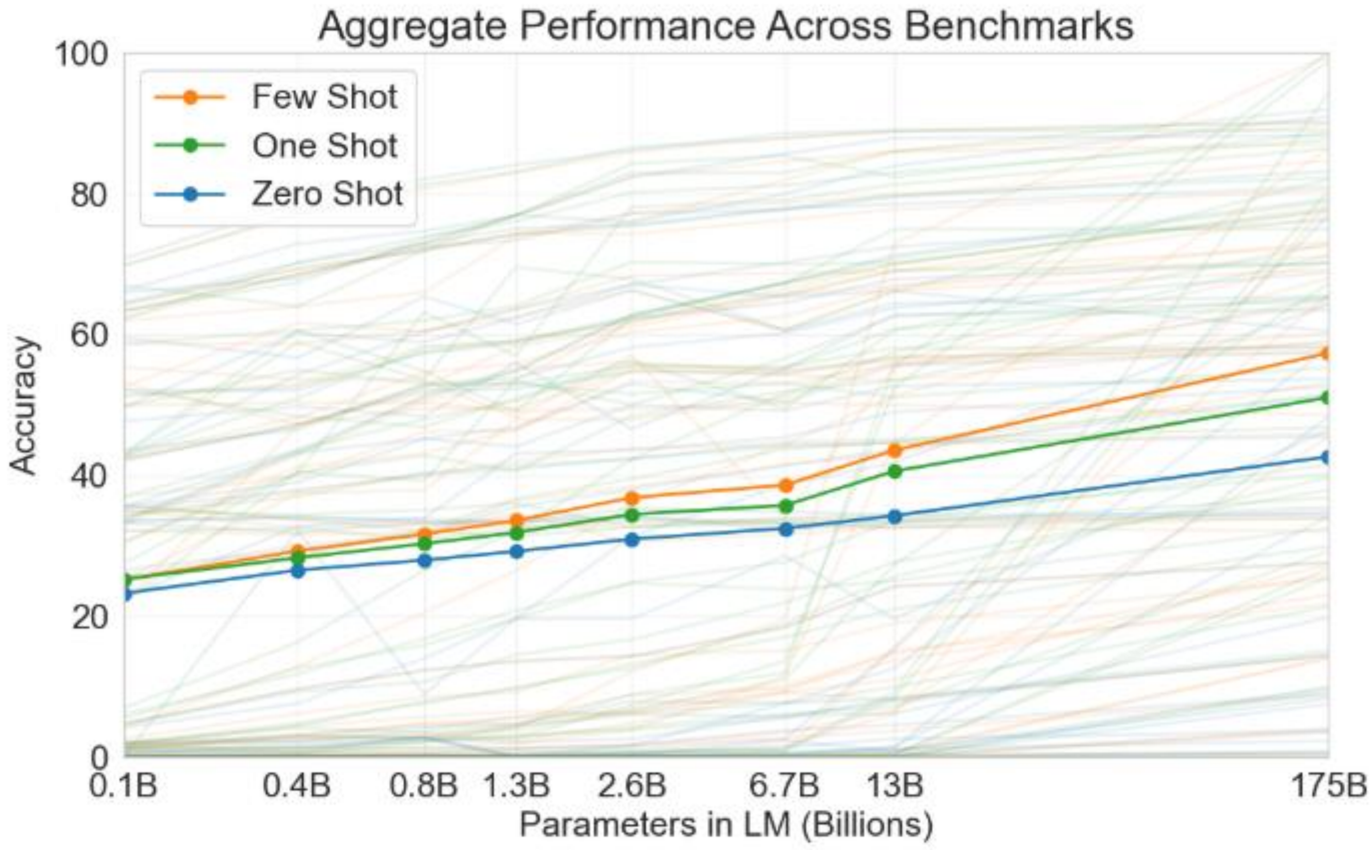
Language Models are Few-Shot Learners

Larger models make increasingly efficient use of in-context information



Language Models are Few-Shot Learners

Aggregate performance for all 42 accuracy-denominated benchmarks



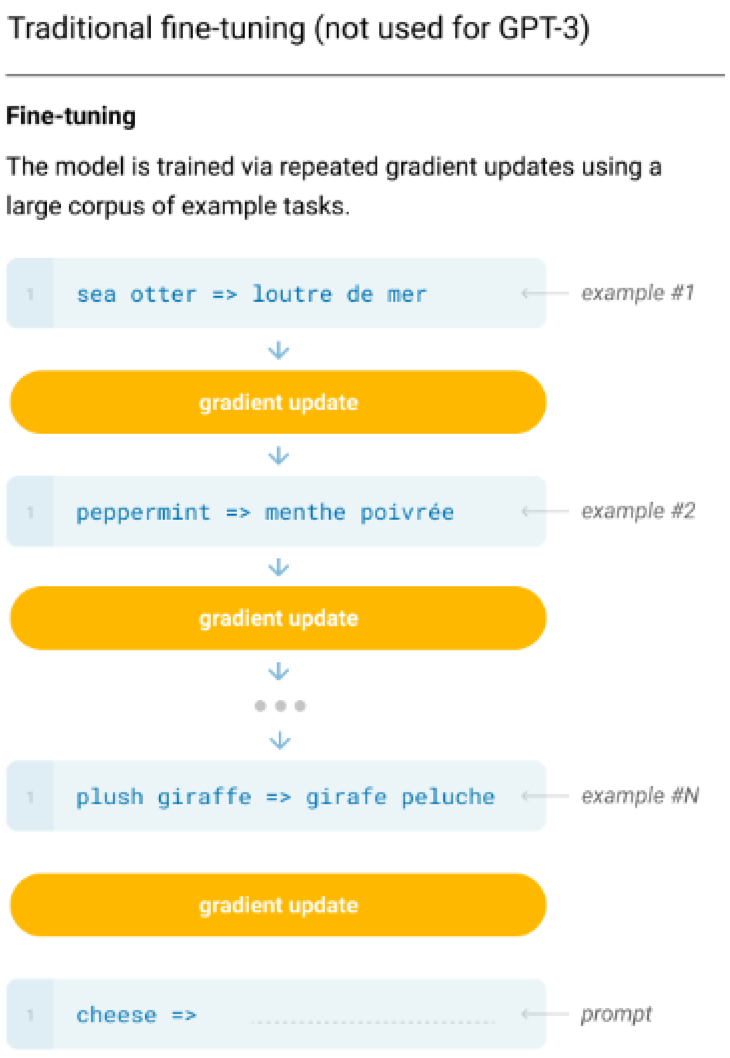
TRAIN AND TEST

Language Models are Few-Shot Learners

Approach

- Fine-Tuning (FT)

 # *Pros*: strong performance on many benchmarks
 # *Cons*: the need for a new large dataset for every task
the potential for poor generalization out-of-distribution
the potential to exploit spurious features of the training data
potentially resulting in an unfair comparison with human-performance

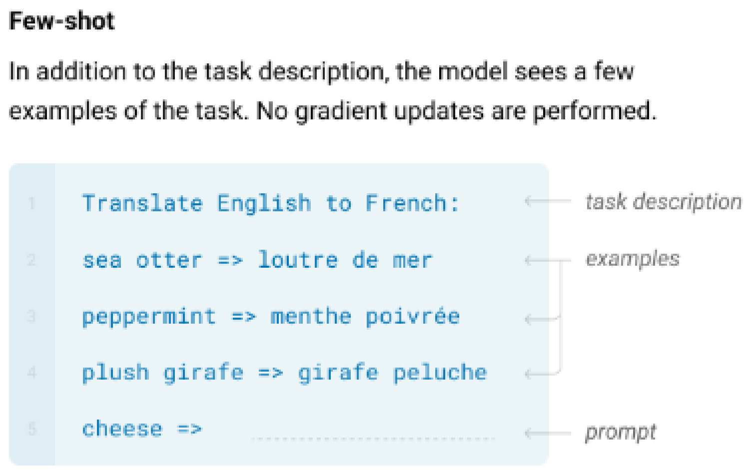


TRAIN AND TEST

Language Models are Few-Shot Learners

Approach

- Few-Shot (FS)



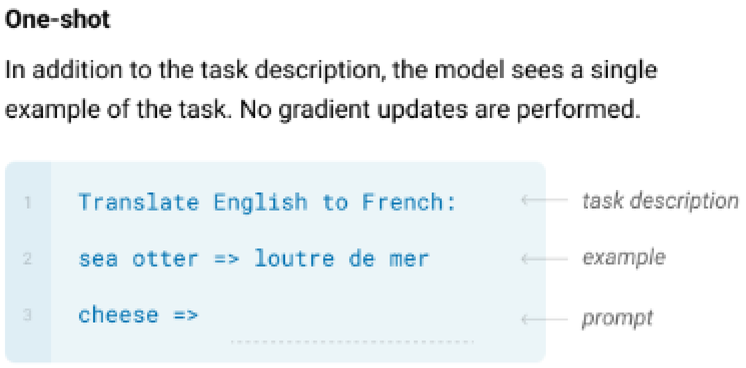
 # *Pros*: major reduction in the need for task-specific data and reduced potential to learn an overly narrow distribution from a large but narrow fine-tuning dataset
 # *Cons*: Results from this method have so far been much worse than SOTA fine-tuned models, a small amount of task specific data is still required

Language Models are Few-Shot Learners

Approach

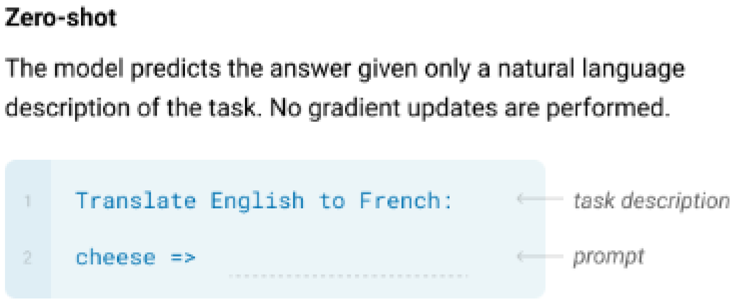
- One-Shot (1S)

Only one demonstration is allowed in addition to a natural language description of the task



- Zero-Shot (0S)

The model is only given a natural language instruction describing the task

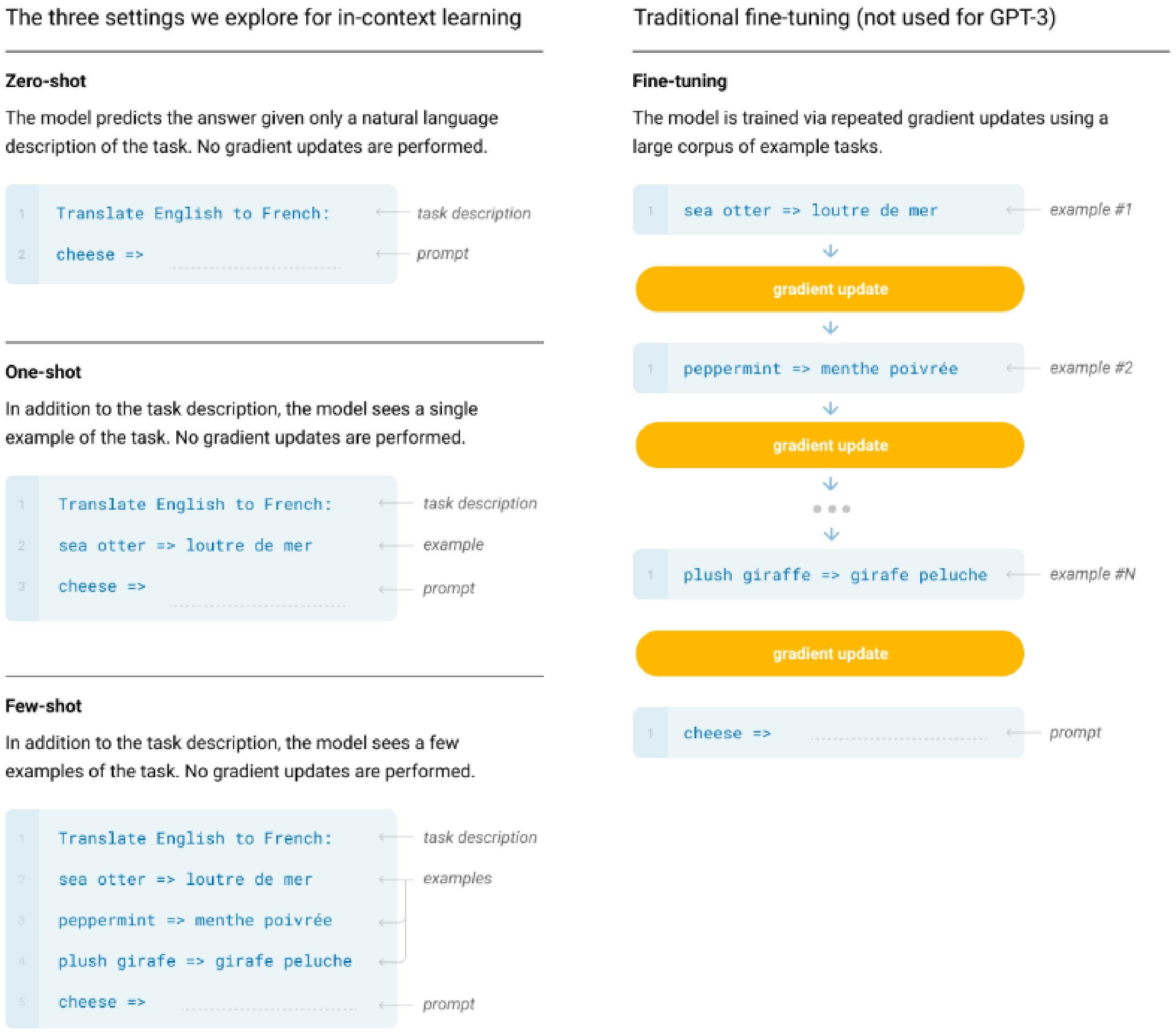


TRAIN AND TEST

Language Models are Few-Shot Learners

Compare to traditional FT

“No gradient updates”



Language Models are Few-Shot Learners

Parameters of model

Model Name	n_{params}	n_{layers}	d_{model}	n_{heads}	d_{head}	Batch Size	Learning Rate
GPT-3 Small	125M	12	768	12	64	0.5M	6.0×10^{-4}
GPT-3 Medium	350M	24	1024	16	64	0.5M	3.0×10^{-4}
GPT-3 Large	760M	24	1536	16	96	0.5M	2.5×10^{-4}
GPT-3 XL	1.3B	24	2048	24	128	1M	2.0×10^{-4}
GPT-3 2.7B	2.7B	32	2560	32	80	1M	1.6×10^{-4}
GPT-3 6.7B	6.7B	32	4096	32	128	2M	1.2×10^{-4}
GPT-3 13B	13.0B	40	5140	40	128	2M	1.0×10^{-4}
GPT-3 175B or "GPT-3"	175.0B	96	12288	96	128	3.2M	0.6×10^{-4}

Table 2.1: Sizes, architectures, and learning hyper-parameters (batch size in tokens and learning rate) of the models which we trained. All models were trained for a total of 300 billion tokens.

All models use a context window of $n_{\text{ctx}} = 2048$ tokens.

✓ Make same input size to compare with other models

TRAIN AND TEST

Language Models are Few-Shot Learners

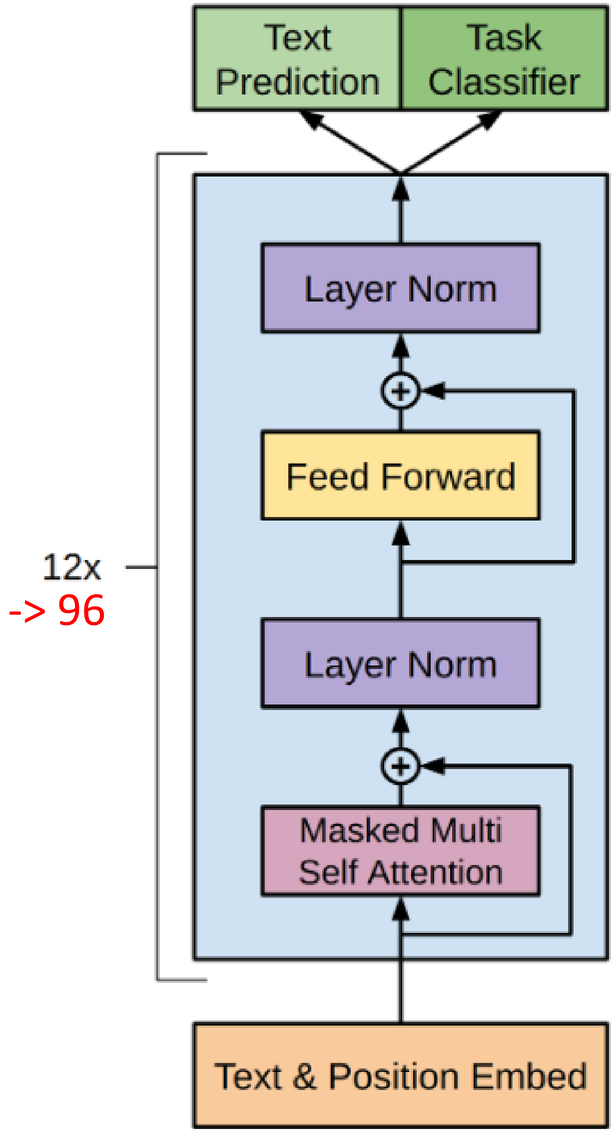
Parameters of model

Model Name	n_{params}	n_{layers}	d_{model}	n_{heads}	d_{head}	Batch Size
GPT-3 Small	125M	12	768	12	64	0.5M
GPT-3 Medium	350M	24	1024	16	64	0.5M
GPT-3 Large	760M	24	1536	16	96	0.5M
GPT-3 XL	1.3B	24	2048	24	128	1M
GPT-3 2.7B	2.7B	32	2560	32	80	1M
GPT-3 6.7B	6.7B	32	4096	32	128	2M
GPT-3 13B	13.0B	40	5140	40	128	2M
GPT-3 175B or "GPT-3"	175.0B	96	12288	96	128	3.2M

Table 2.1: Sizes, architectures, and learning hyper-parameters (batch size in tokens and learni
which we trained. All models were trained for a total of 300 billion tokens.

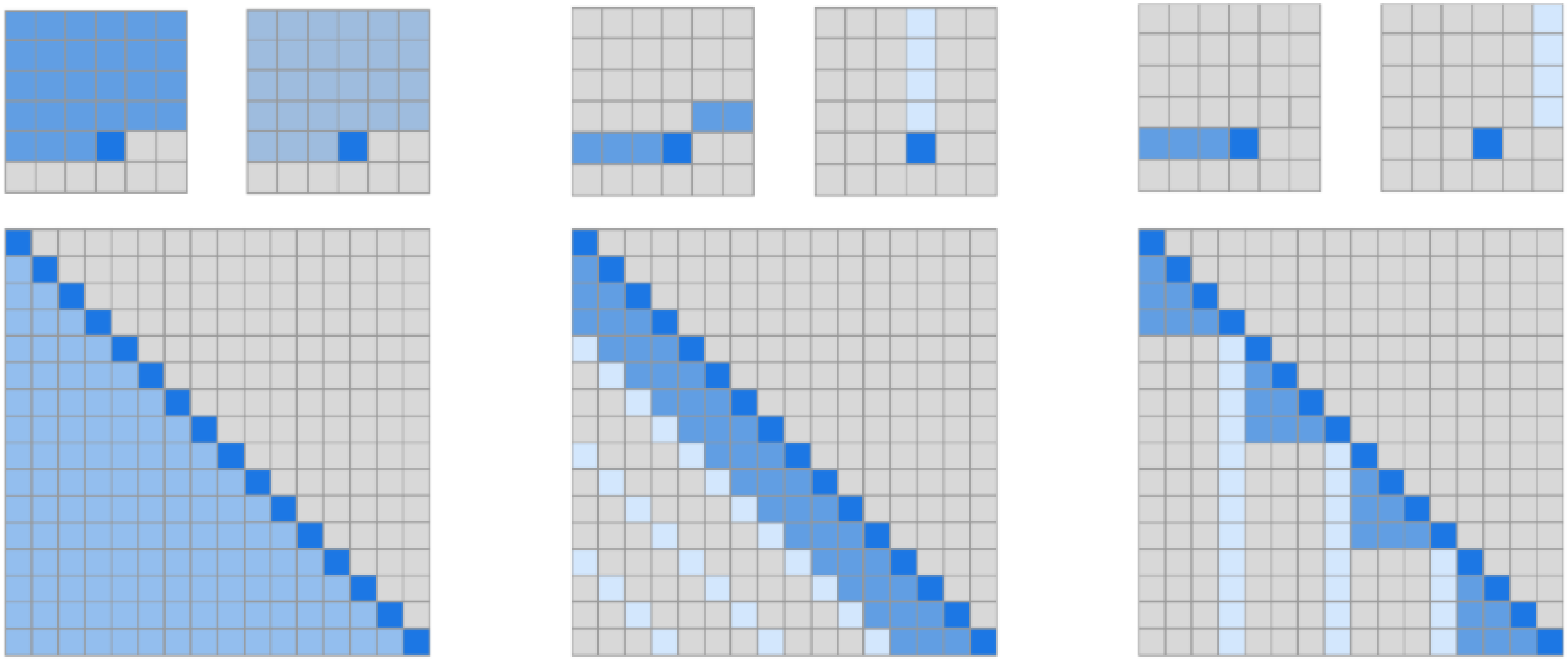
All models use a context window of $n_{\text{ctx}} = 2048$ tokens.

✓ Make same input size to compare with other models



Language Models are Few-Shot Learners

Architecture



(a) Transformer (b) Sparse Transformer (strided) (c) Sparse Transformer (fixed)

Language Models are Few-Shot Learners

Training Dataset

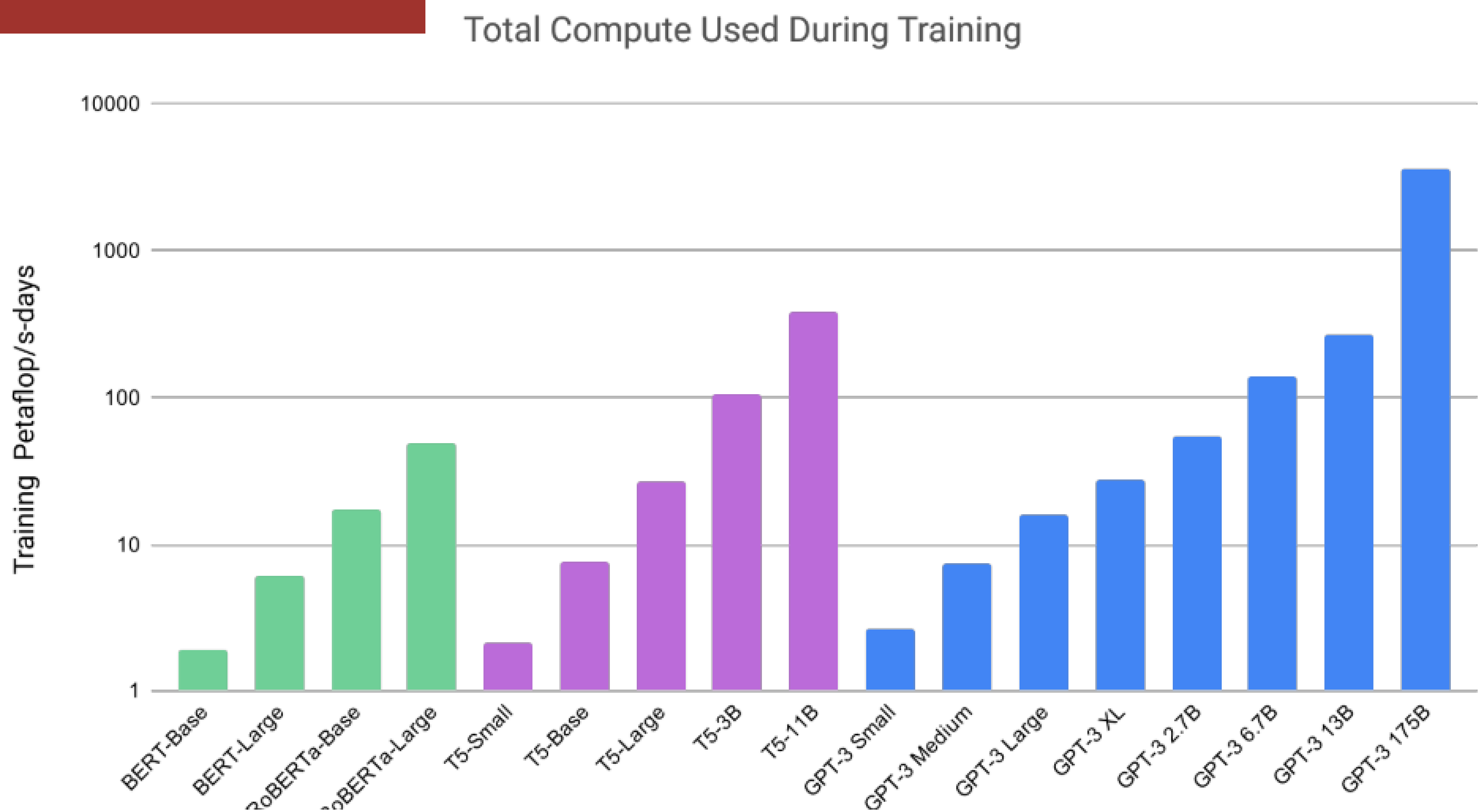
- Common Crawl dataset (constituting nearly a trillion words)

- 3 steps to improve the average quality of the dataset

    1. downloaded and filtered a version of Common Crawl based on similarity to a range of high-quality reference corpora
    2. performed fuzzy deduplication at the document level, within and across datasets, to prevent redundancy and preserve the integrity of our held-out validation set as an accurate measure of overfitting
    3. Added known high-quality reference corpora to the training mix to augment Common Crawl and increase its diversity

Language Models are Few-Shot Learners

Total compute used during training



TRAIN AND TEST

Language Models are Few-Shot Learners

Datasets used to train GPT-3

Dataset	Quantity (tokens)	Weight in training mix	Epochs elapsed when training for 300B tokens
Common Crawl (filtered)	410 billion	60%	0.44
WebText2	19 billion	22%	2.9
Books1	12 billion	8%	1.9
Books2	55 billion	8%	0.43
Wikipedia	3 billion	3%	3.4

Q. GPT Cost?

Q. Time Cost?

TRAIN AND TEST

Language Models are Few-Shot Learners

Datasets used to train GPT-3

Dataset	Quantity (tokens)	Weight in training mix	Epochs elapsed when training for 300B tokens
Common Crawl (filtered)	410 billion	60%	0.44
WebText2	19 billion	22%	2.9
Books1	12 billion	8%	1.9
Books2	55 billion	8%	0.43
Wikipedia	3 billion	3%	3.4

Q. GPT3 Cost? ~ 5 million

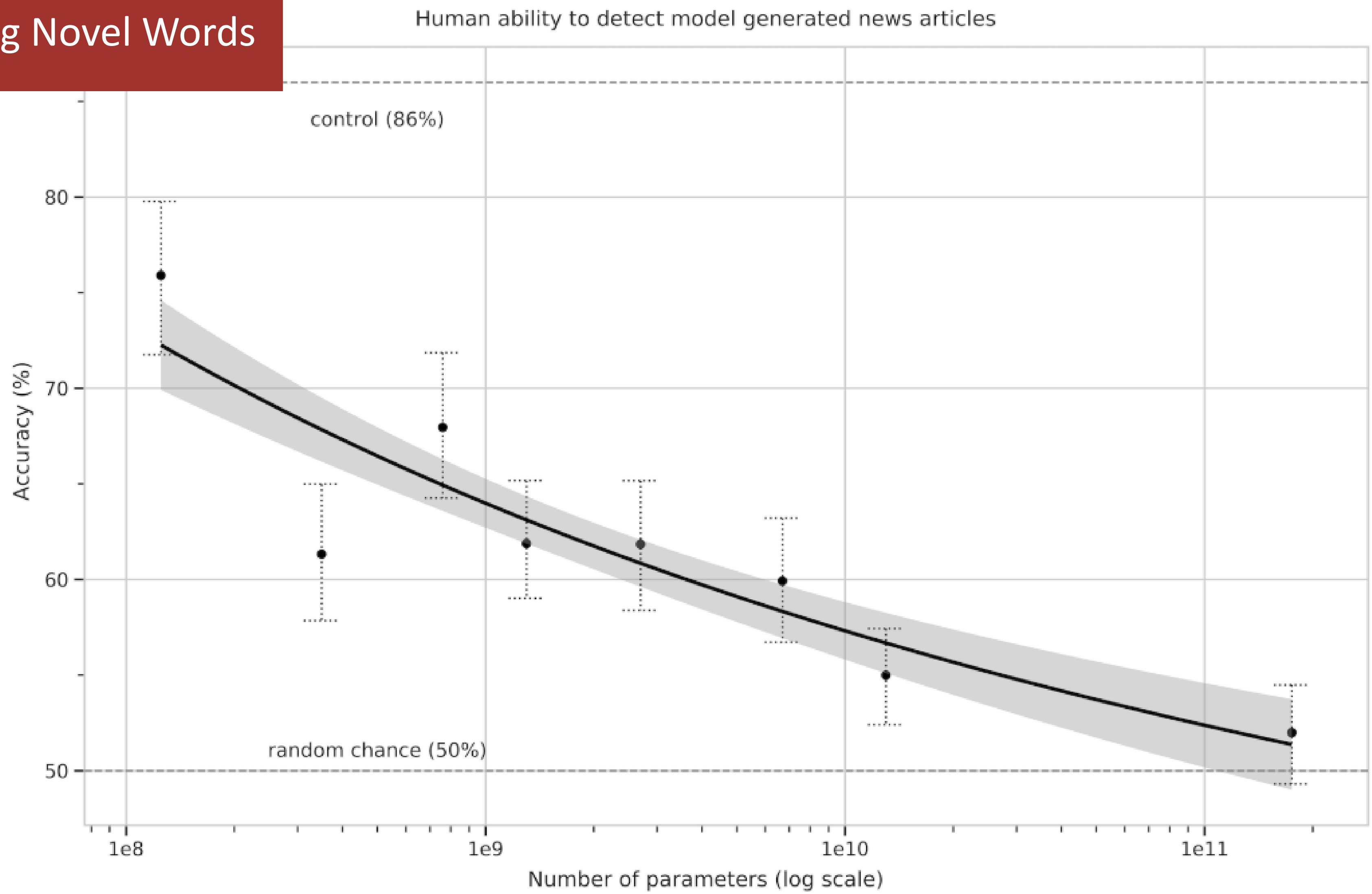
Q. time? 3.14E+23 flops computes

~= v100 GPU(28 Tflops) 355years

TRAIN AND TEST

Language Models are Few-Shot Learners

Learning and Using Novel Words



Language Models are Few-Shot Learners

Grammar

Poor English input: I'd be more than happy to work with you in another project.
Good English output: I'd be more than happy to work with you on another project.

Poor English input: Please provide me with a short brief of the design you're looking for and that'd be nice if you could share some examples or project you did before.
Good English output: Please provide me with a brief description of the design you're looking for and that would be nice if you could share some examples or projects you have done before.

Poor English input: The patient was died.
Good English output: The patient died.

Poor English input: We think that Leslie likes ourselves.
Good English output: We think that Leslie likes us.

Poor English input: Janet broke Bill on the finger.
Good English output: Janet broke Bill's finger.

Poor English input: Mary arranged for, in St. Louis, John to rent a house cheap.
Good English output: Mary arranged for John to rent a house in St. Louis.

Poor English input: Today I have went to the store to to buys some many bottle of water.
Good English output: Today I went to the store to buy some bottles of water.

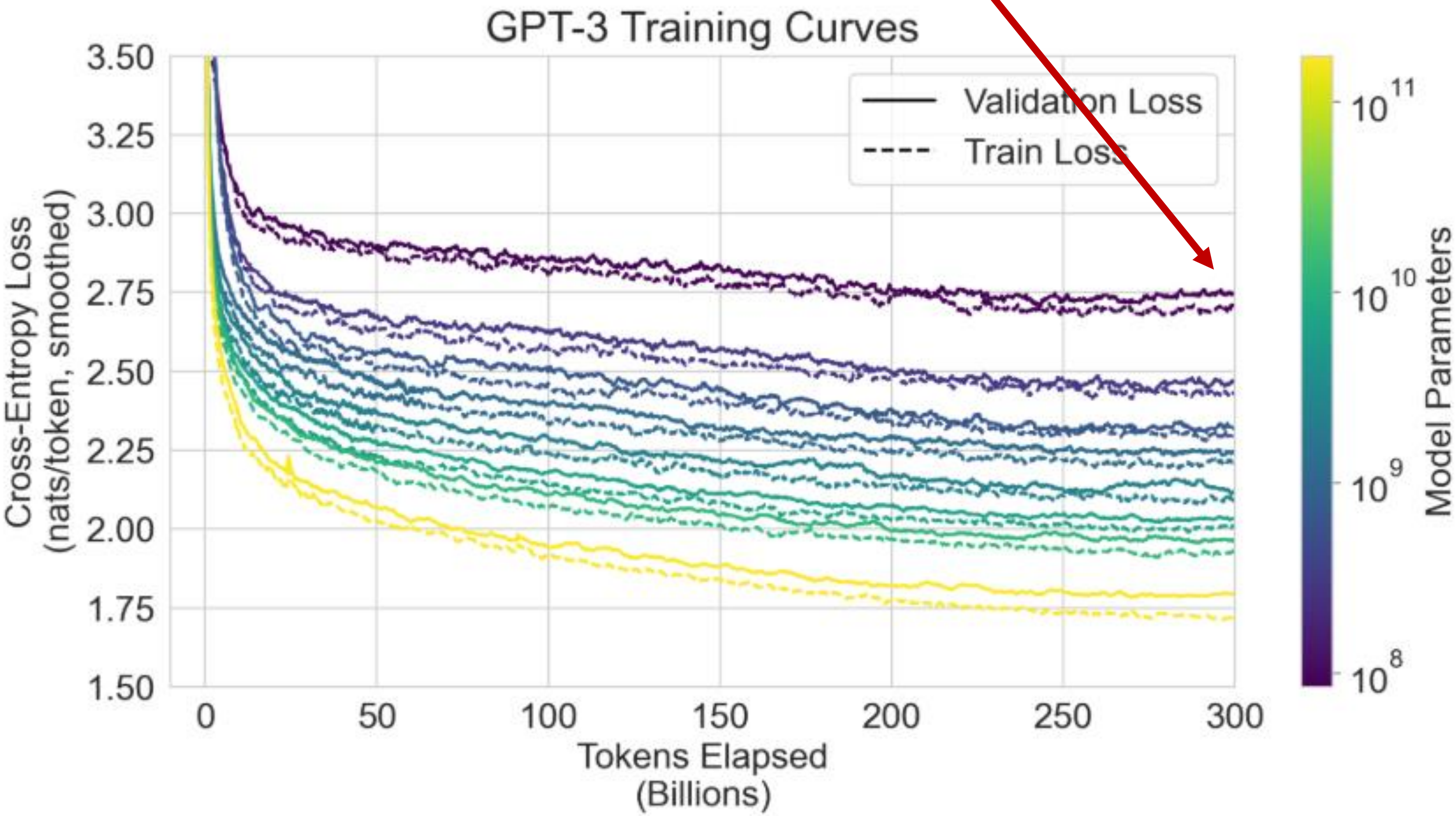
Poor English input: I have tried to hit ball with bat, but my swing is has miss.
Good English output: I tried to hit the ball with the bat, but my swing missed.

TRAIN AND TEST

Language Models are Few-Shot Learners

Overfitting?

most of the gap comes from a difference in difficulty rather than overfitting



Q. Gap meaning?

Q. Difficulty?

Q. Validation loss

vs

Training loss

Language Models are Few-Shot Learners

Limitation

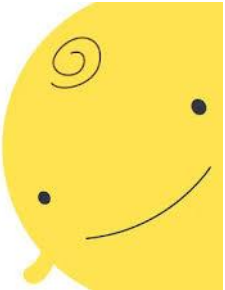
- Weaknesses in text synthesis and several NLP tasks
  - *samples still sometimes repeat themselves semantically at the document level, start to lose coherence over sufficiently long passages*
  - *occasionally contain non-sequitur sentences or paragraphs*

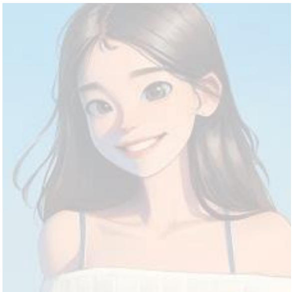
- Structural and algorithmic limitation
  - *Auto-regressive, not bidirectional*
  - *Pretraining objective weights every token equally*

TRAIN AND TEST

Language Models are Few-Shot Learners

Limitation





- poor sample efficiency during pre-training
- ambiguity about whether few-shot learning actually learns new tasks "from scratch" at inference time, or if it simply recognizes and identifies tasks that it has learned during training
- its predictions on novel inputs as observed by the much higher variance in performance than humans on standard benchmarks, and it retains the biases of the data it has been trained on

instruct GPT

“Training language models to follow instructions with human feedback”

Instruct GPT

Difference

Can "directly instruct" to GPT

Write a short story about the moon and the stars: *Does the moon like the stars? Does the moon like the stars? Does the moon like the stars? Does the moon like the stars? Does the moon like the stars? Does t the moon like the stars? Does the moon like the stc stars?*

GPT3

Instruct GPT

Write a short story about the moon and the stars:

Once upon a time, the moon and the stars lived together in the sky. They were the best of friends, and they loved spending their time together. The moon would light up the sky at night, and the stars would twinkle and shine. They were always happy and they never fought.

But one day, the moon and the stars began to grow apart. The moon started to feel left out and the stars began to ignore her. The moon didn't understand what was happening, and she felt very lonely.

Instruct GPT?

Motivation

“Misaligned”

=> Untruthful, Toxic, Not helpful to user

TRAIN AND TEST

Instruct GPT?

Motivation

"Misaligned"

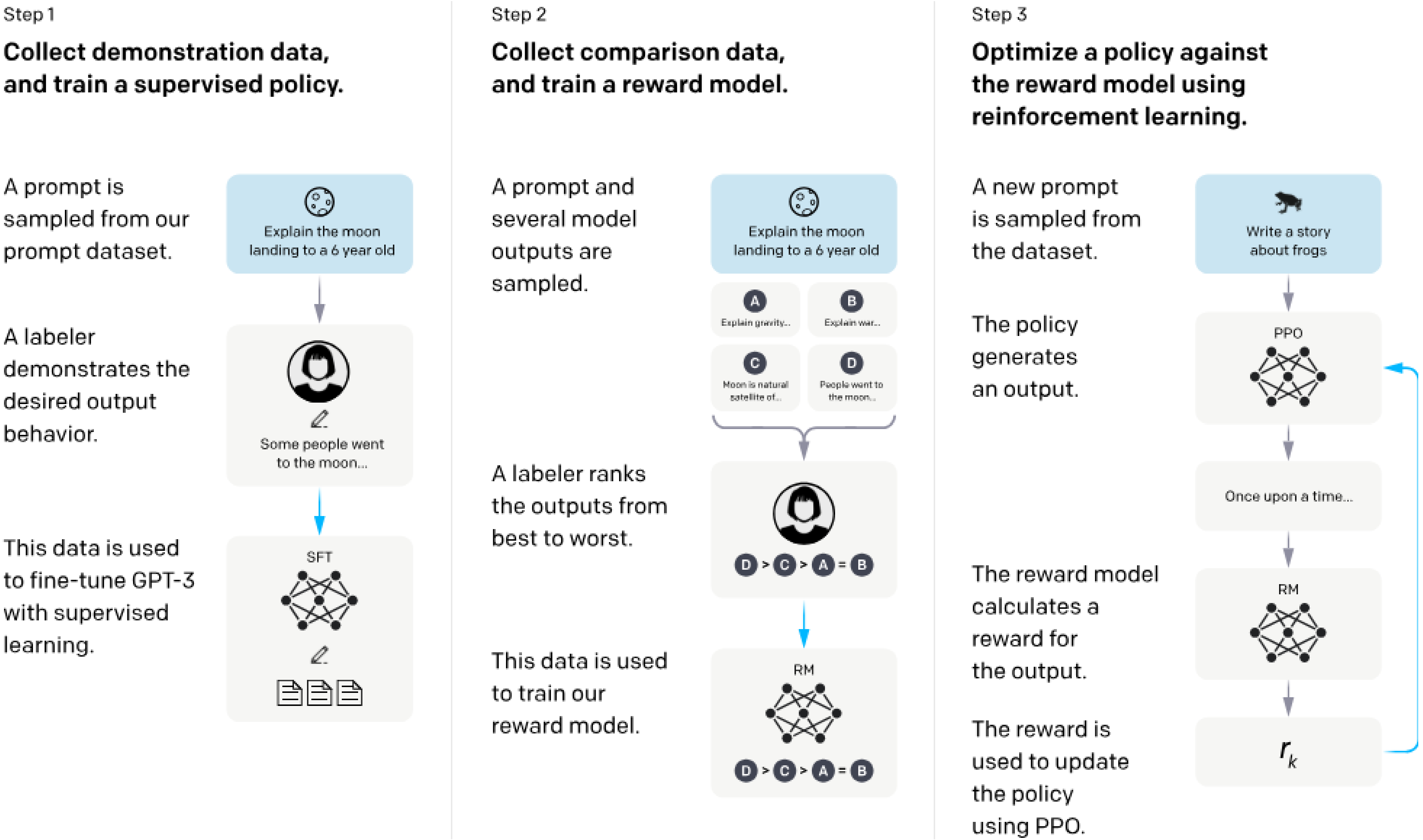
=> Untruthful, Toxic, Not helpful to user

Solution

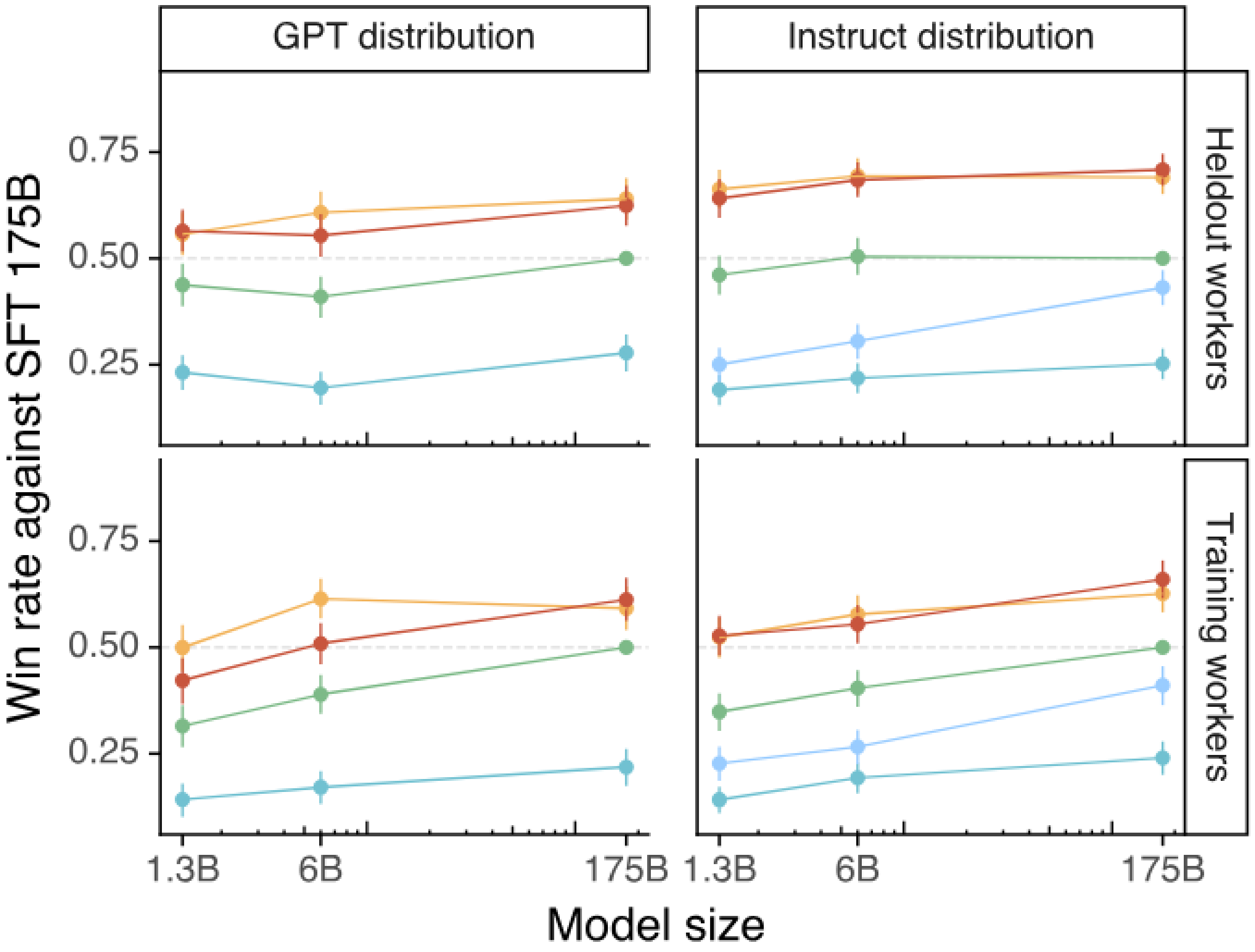
Reinforcement Learning from Human Feedback (RLHF; Christiano et al., 2017; Stiennon et al., 2020)

TRAIN AND TEST

Instruct GPT method



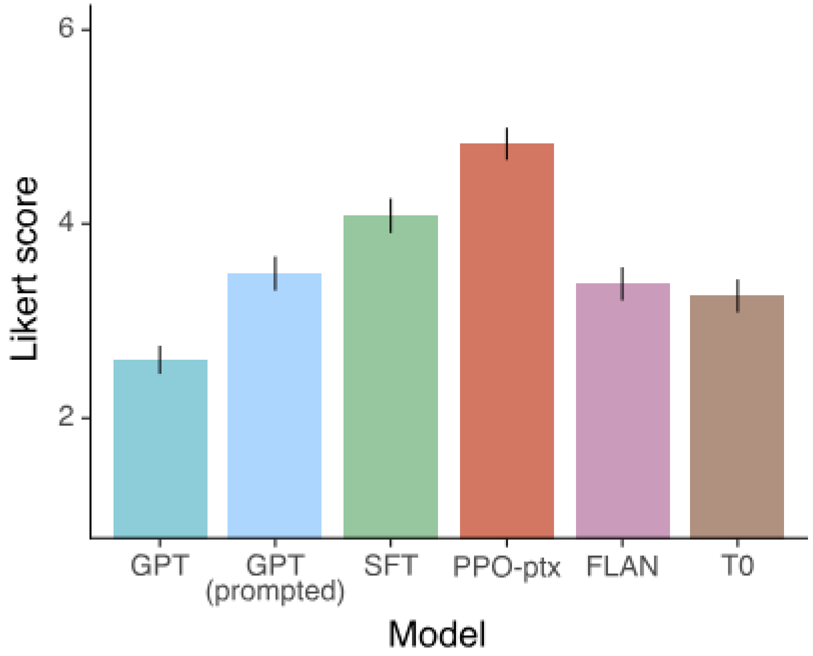
Instruct GPT Results

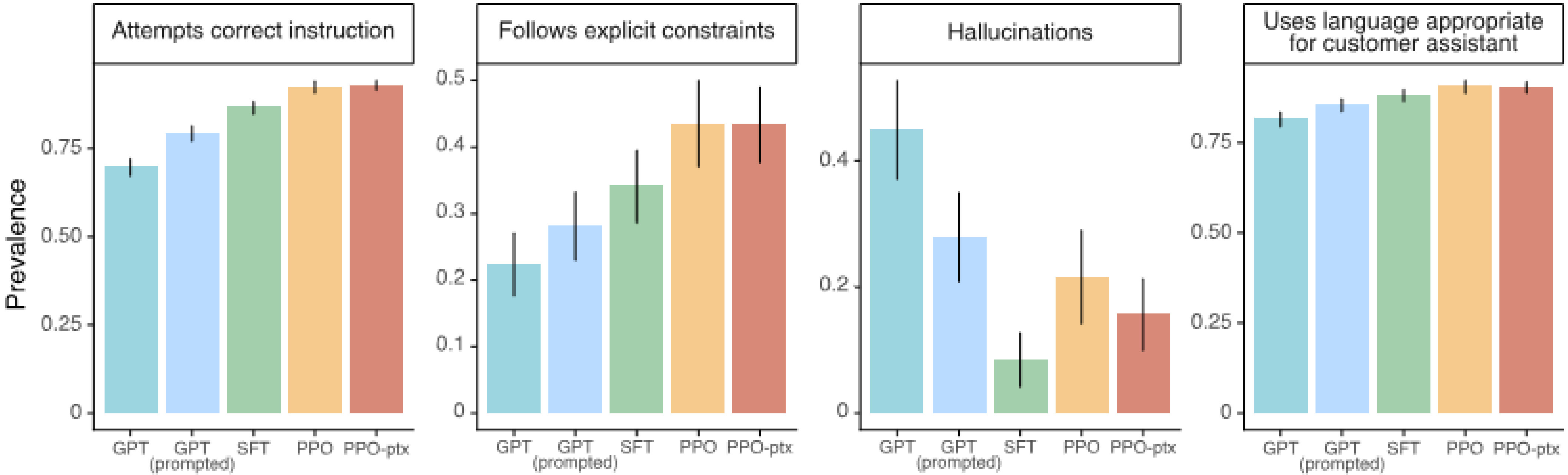


Instruct GPT Results

It is unclear how to measure honesty in purely generative models

Similarly to honesty, measuring the harms of language models also poses many challenges





prerequisite

Reinforcement Learning

What is Reinforcement Learning?

prerequisite

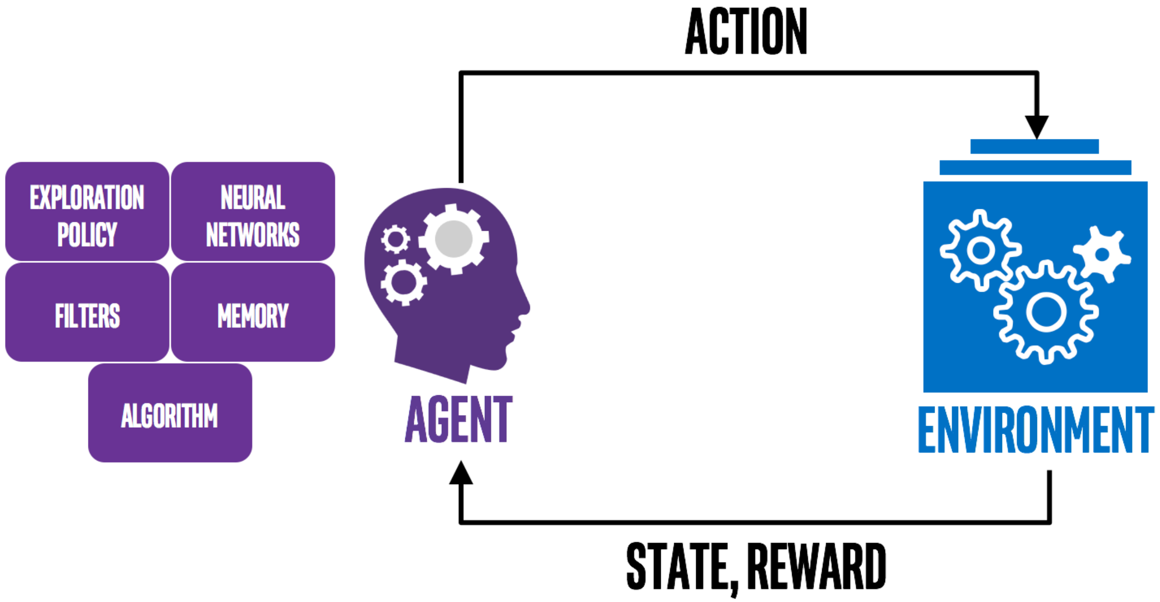
Reinforcement Learning

What is Reinforcement Learning?

MDP, PPO, Q-learning, DQN, actor-critic…

Richard S. Sutton

David Silver



TRAIN AND TEST

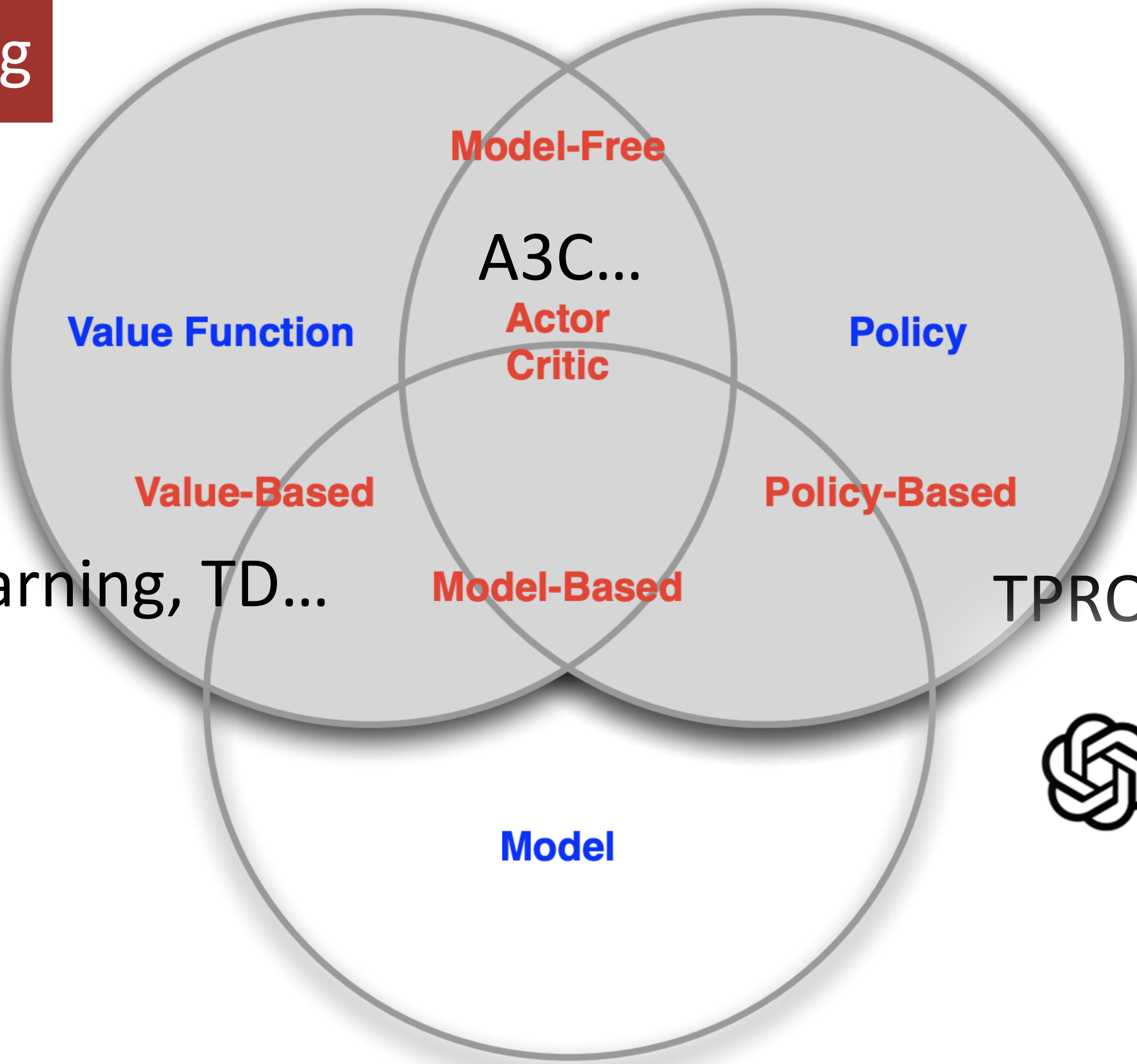
RL Overview

Reinforcement Learning

DeepMind

AlphaGo

DQN, MC Learning, TD…

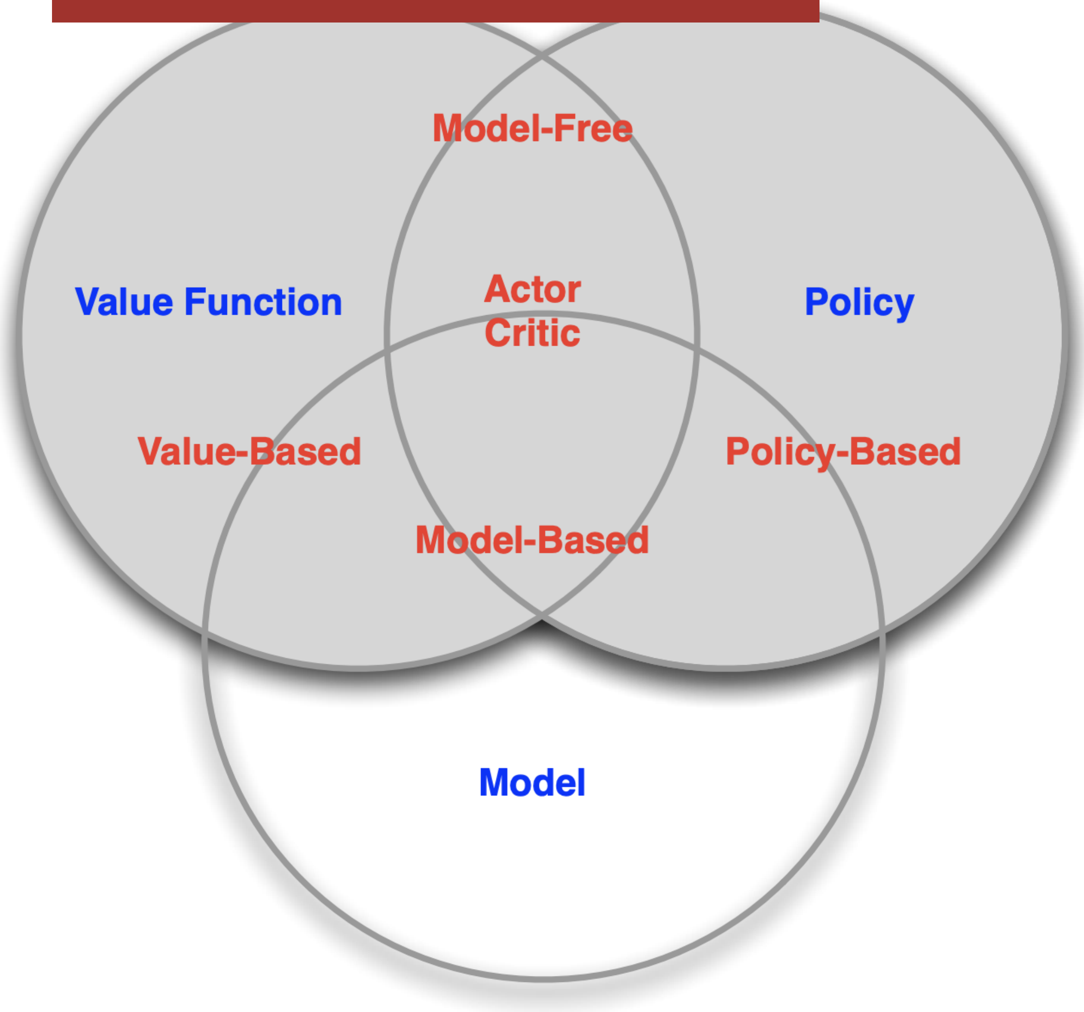


TPRO, PPO…

OpenAI

Contents

Reinforcement Learning



Markov Decision Process (MDP)

= Decision Process + Markov Property

$$P[S_{t+1} | S_t] = P[S_{t+1} | S_1, \ldots, S_t]$$

MDP = <S,A,P,R,γ> =
State, action, transition probability, reward, discount factor

prerequisite

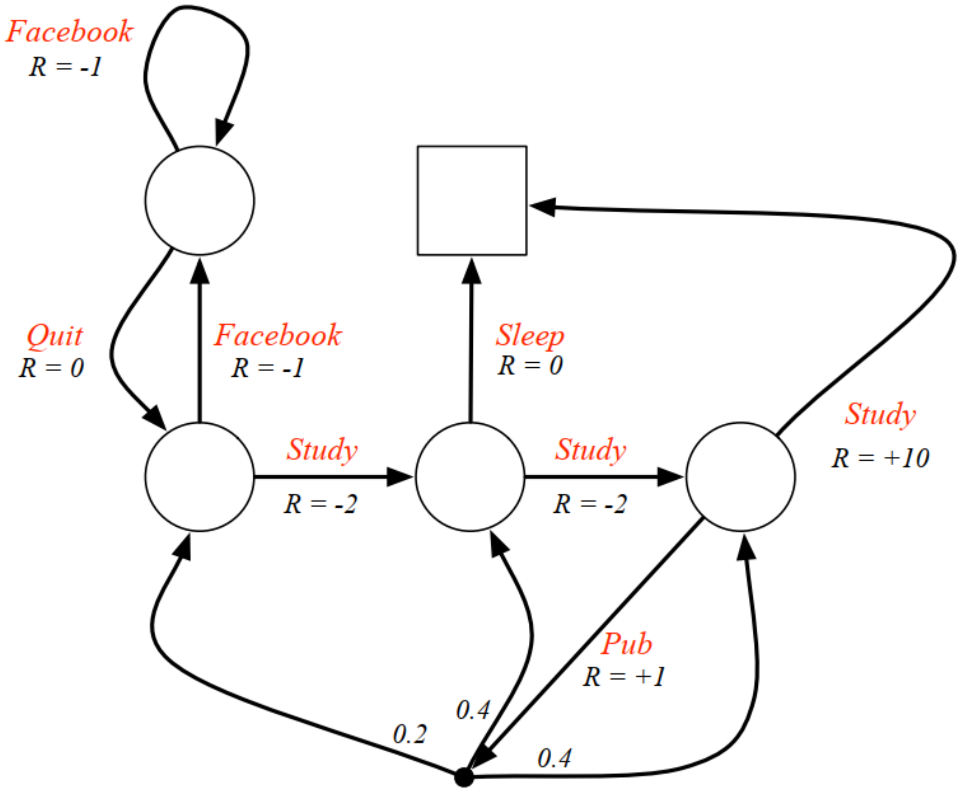
Bellman equation

Bellman equation, optimality…

$$v_*(s) = \max_\pi v_\pi(s)$$

$$q_*(s,a) = \max_\pi q_\pi(s,a)$$

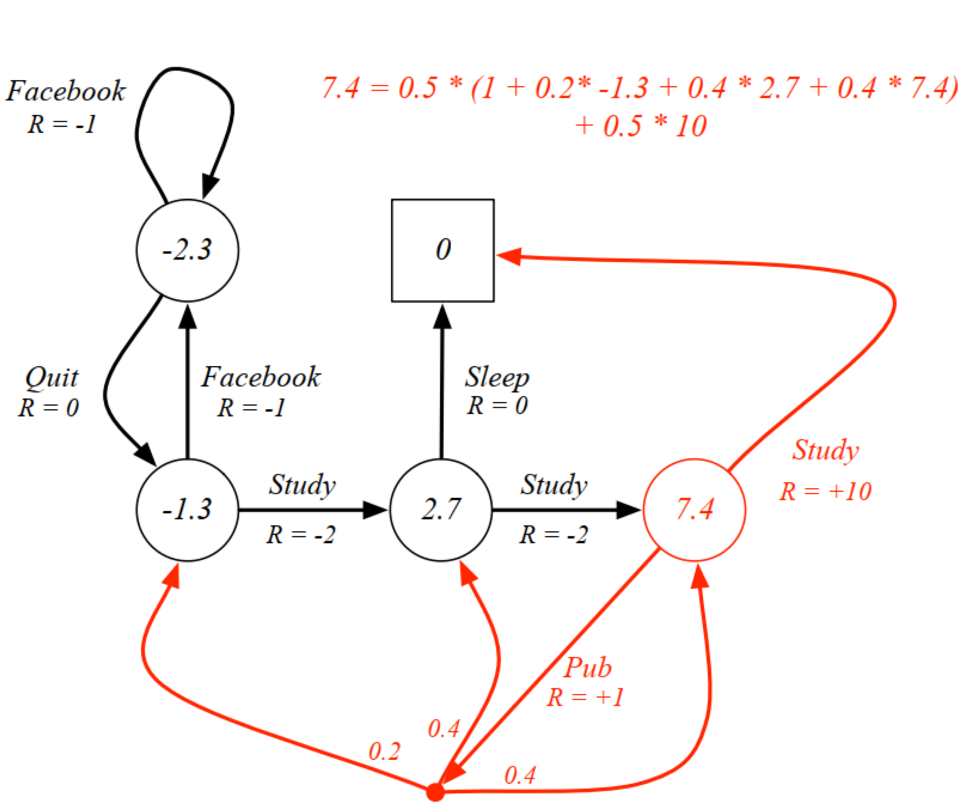
$$\begin{aligned} v_\pi(s_t) &= \mathbb{E}_\pi[G_t] \\ &= \mathbb{E}_\pi[r_{t+1} + \gamma r_{t+2} + \gamma^2 r_{t+3} + \cdots] \\ &= \mathbb{E}_\pi[r_{t+1} + \gamma(r_{t+2} + \gamma r_{t+3} + \cdots)] \\ &= \mathbb{E}_\pi[r_{t+1} + \gamma G_{t+1}] \\ &= \mathbb{E}_\pi[r_{t+1} + \gamma v_\pi(s_{t+1})] \end{aligned}$$



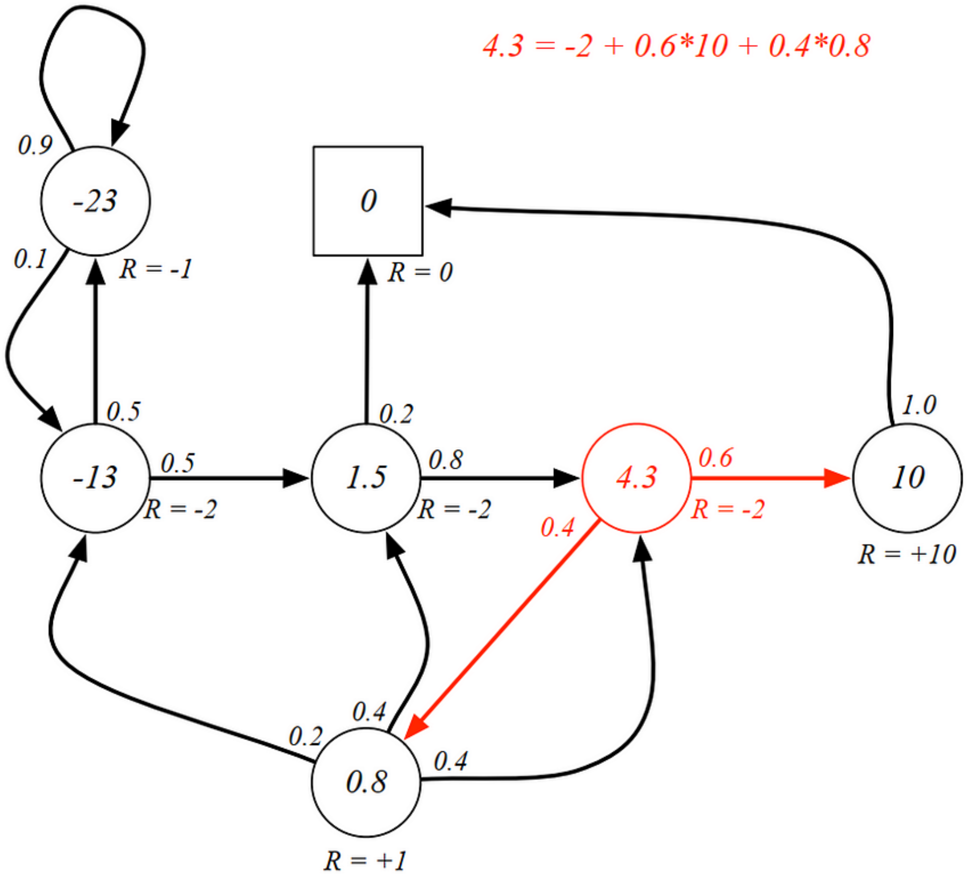
TRAIN AND TEST

prerequisite

MDP (decision)



MRP (reward)



$$v_\pi(s) = \mathbb{E}_\pi\left[R_{t+1} + \gamma v_\pi(S_{t+1}) \mid S_t = s\right]$$

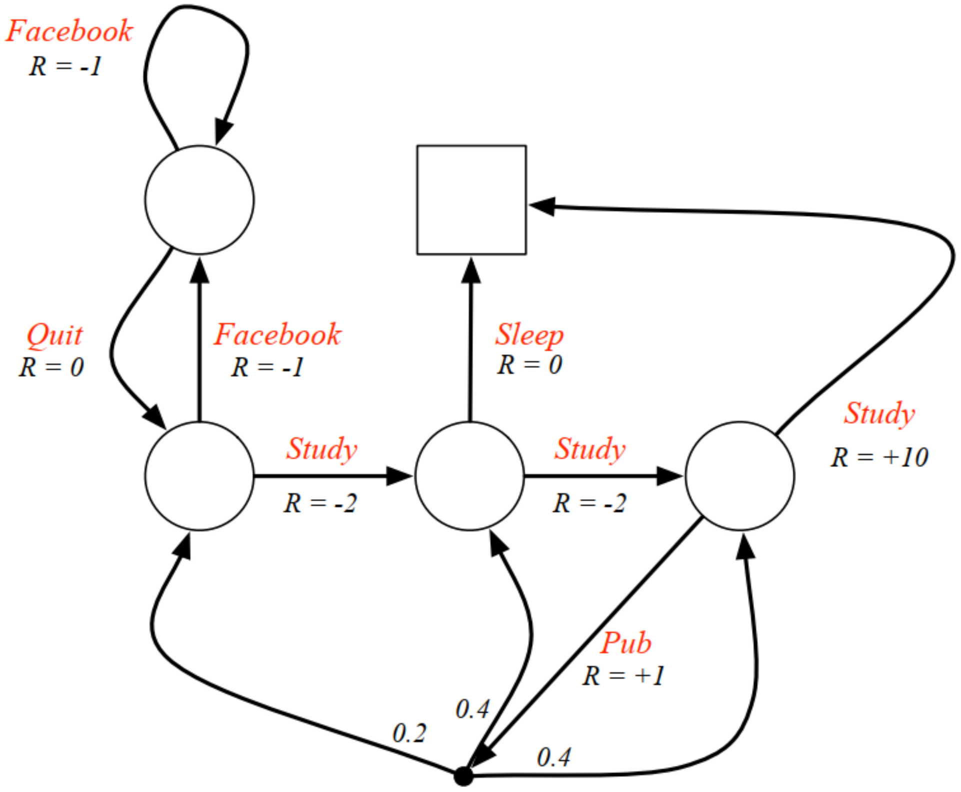
TRAIN AND TEST

prerequisite

Reinforcement Learning

Value based? Policy based?

=> Different Object function



TRAIN AND TEST

prerequisite

TRPO, PPO

Motivation: want to update in trust bound => How?

PPO (Proximal policy optimization)

Trust Region Policy Optimization (TRPO)

$$\underset{\theta}{\text{maximize}}\, \mathbb{E}_{s\sim\rho_{\theta_{\text{old}}}, a\sim q}\left[\frac{\pi_\theta(a|s)}{q(a|s)} Q_{\theta_{\text{old}}}(s,a)\right] \quad (14)$$
$$\text{subject to } \mathbb{E}_{s\sim\rho_{\theta_{\text{old}}}}\left[D_{\text{KL}}(\pi_{\theta_{\text{old}}}(\cdot|s) \,\|\, \pi_\theta(\cdot|s))\right] \le \delta.$$

MM Algorithm

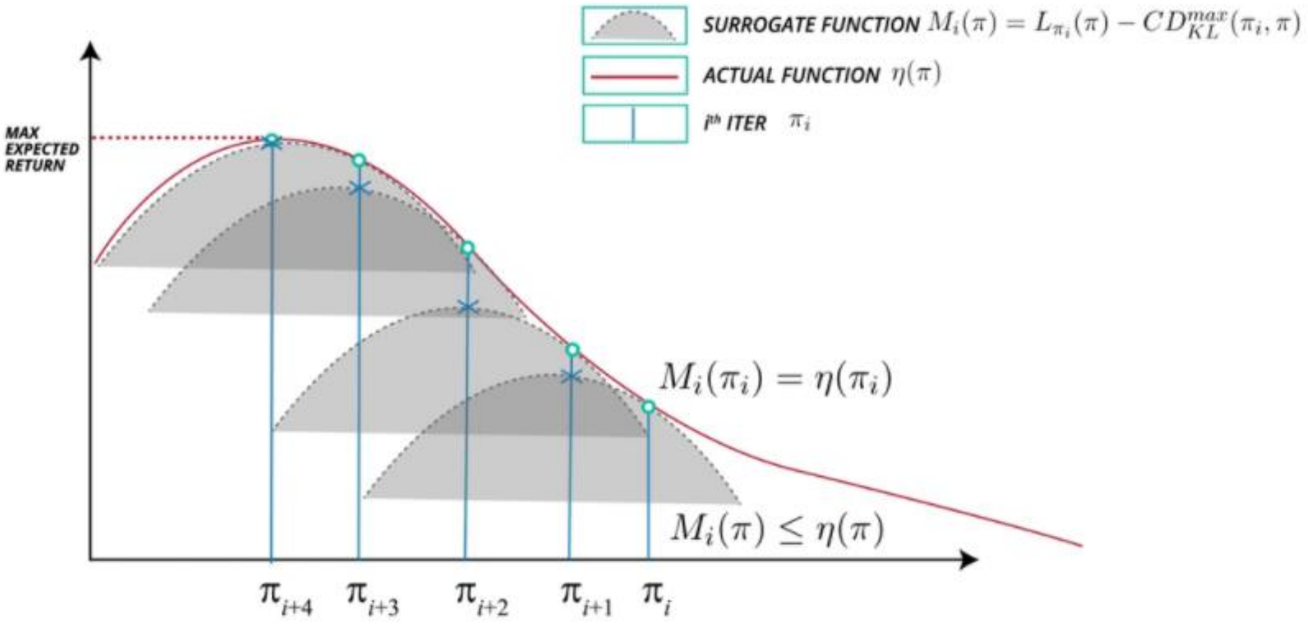
Minorization-Maximization

Proximal policy optimization (PPO)

$$r_t(\theta) = \frac{\pi_\theta(a_t \mid s_t)}{\pi_{\theta_{\text{old}}}(a_t \mid s_t)}$$

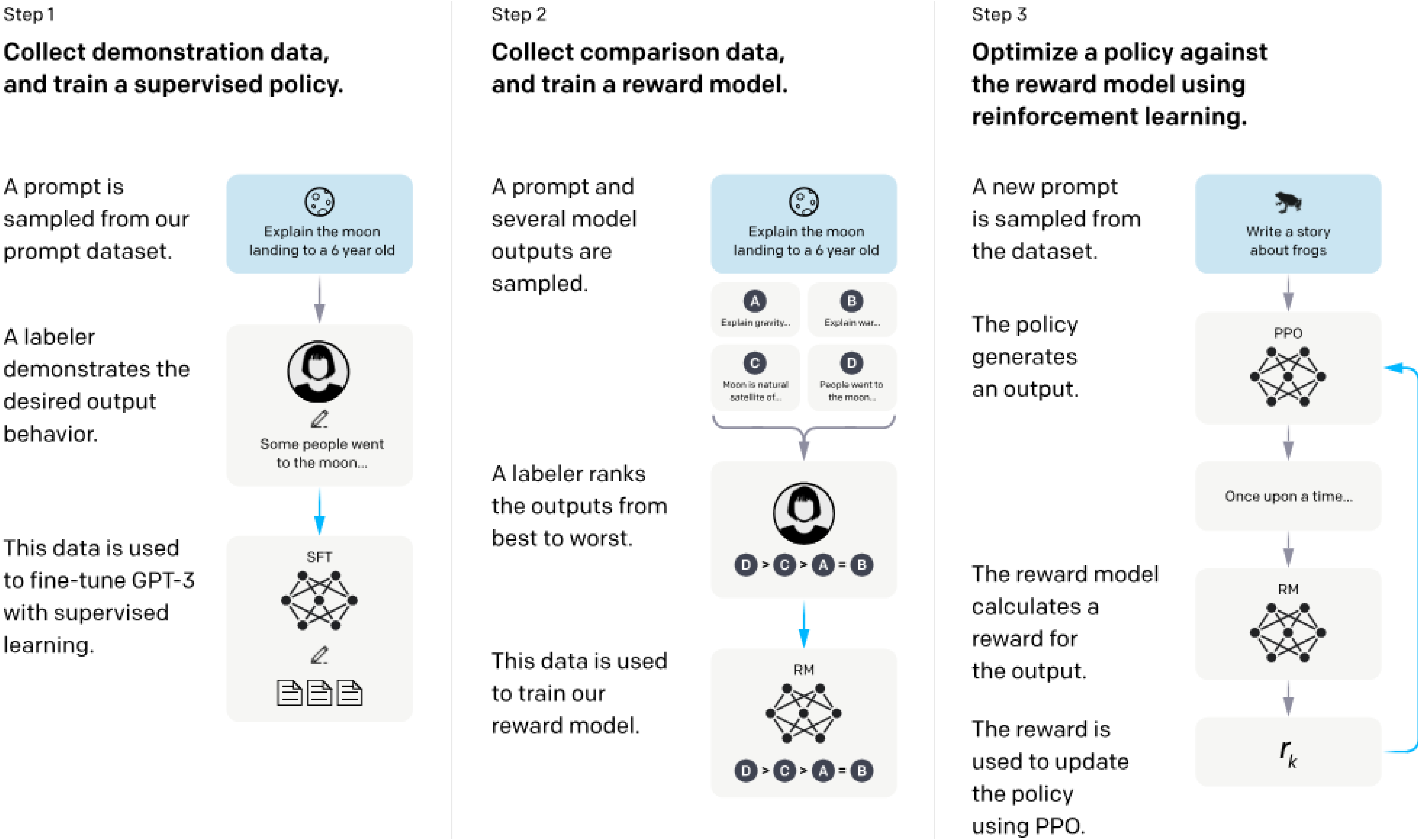
$$L^{CPI}(\theta) = \hat{\mathbb{E}}_t\left[\frac{\pi_\theta(a_t \mid s_t)}{\pi_{\theta_{\text{old}}}(a_t \mid s_t)}\hat{A}_t\right] = \hat{\mathbb{E}}_t\left[r_t(\theta)\hat{A}_t\right].$$

$$L^{CLIP}(\theta) = \hat{\mathbb{E}}_t\left[\min(r_t(\theta)\hat{A}_t, \text{clip}(r_t(\theta), 1-\epsilon, 1+\epsilon)\hat{A}_t)\right]$$



TRAIN AND TEST

Instruct GPT method

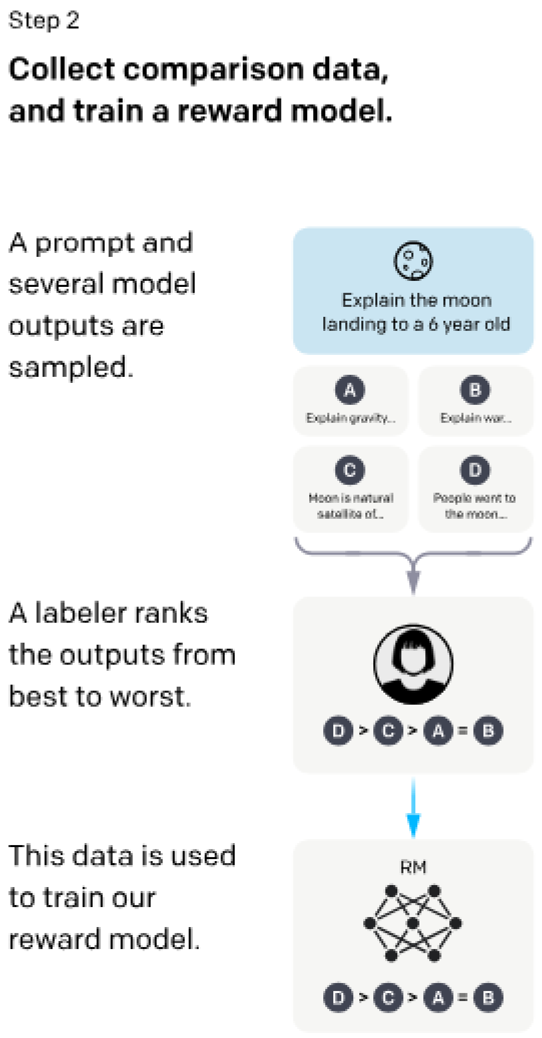


Instruct GPT

Reward model

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

“We ran an experiment where we split our labelers into 5 groups, and train 5 RMs (with 3 different seeds)”



Instruct GPT

PPO

$$\text{objective}\,(\phi) = E_{(x,y)\sim D_{\pi_\phi^{\mathrm{RL}}}}\left[r_\theta(x,y) - \beta \log\left(\pi_\phi^{\mathrm{RL}}(y \mid x)/\pi^{\mathrm{SFT}}(y \mid x)\right)\right] + \gamma E_{x\sim D_{\text{pretrain}}}\left[\log(\pi_\phi^{\mathrm{RL}}(x))\right]$$

Step 3

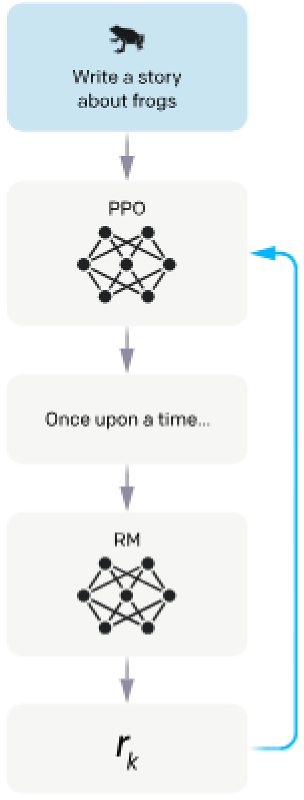
Optimize a policy against the reward model using reinforcement learning.

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.



TRAIN AND TEST

prerequisite

Reinforcement Learning

What is Reinforcement Learning?

We ran an experiment where we split our labelers into 5 groups,
and train 5 RMs (with 3 different seeds)

TRAIN AND TEST

Limitation

Methodology

The behavior of our InstructGPT models is determined in part by the human feedback obtained from our contractors. ~~
However, this group is clearly not representative of the full spectrum of people who will use and be affected by our deployed models.

Models

Our models are neither fully aligned nor fully safe;
they still generate toxic or biased outputs, make up facts, and generate sexual and violent content without explicit prompting. They can also fail to generate reasonable outputs on some inputs.

TRAIN AND TEST

Open questions

Many methods could be tried to further decrease the models' propensity to generate toxic, biased, or otherwise harmful outputs

While we mainly focus on RLHF, there are many other algorithms that could be used to train policies on our demonstration and comparison data to get even better results.
$\Rightarrow$ explore expert iteration, simpler behavior cloning methods…

Comparisons are also not necessarily the most efficient way of providing an alignment signal.

…

TRAIN AND TEST

TRAIN AND TEST