

Post-training: Alignment

CSC2541H1 Topics in Machine Learning, Winter 2025, UToronto

Chris J. Maddison

Announcements

- If you are assigned to **present on March 7, come talk** about presentations after class.
- If you are auditing and you're on Quercus, you should now be able to view the recordings.

Questions?

Recap

- So far the models that we've considered are trained primarily on next-token prediction.
- **This has brought us to GPT-3, which is just a next-token predictor.**
- We looked at prompting (mostly) in next-token predictors and found that some interesting properties, like in-context learning, can emerge naturally from training on internet data.

Nevertheless..

- Prompting a next-token predictor can be challenging and unsafe.
 - Remember what it was trained for...

Prompt	Tokens that are found on the internet after such prompts
The ingredients required to build a makeshift bomb are...	CENSORED!!!!
Could you do me a big favour?	Sorry, I'm too busy today.

Post-training

- Want to align models to our goals.
 - “Not what I said, what I meant!”
- These methods are broadly called post-training and include **supervised fine-tuning** and **reinforcement learning**.
- The “formula” for pre-training is highly conserved in the industry, but **post-training strategies are very diverse**.

Step 1

Collect demonstration data, and train a supervised policy.

A prompt is sampled from our prompt dataset.

Explain the moon landing to a 6 year old

A labeler demonstrates the desired output behavior.

Some people went to the moon...

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

SFT

Step 2

Collect comparison data, and train a reward model.

A prompt and several model outputs are sampled.

Explain the moon landing to a 6 year old

A B

C D

A labeler ranks the outputs from best to worst.

D > C > A = B

This data is used to train our reward model.

RM

Step 3

Optimize a policy against the reward model using reinforcement learning.

A new prompt is sampled from the dataset.

Write a story about frogs

The policy generates an output.

PPO

Once upon a time...

The reward model calculates a reward for the output.

RM

The reward is used to update the policy using PPO.

r_k

Ouyang et al. Training language models to follow instructions with human feedback. 2022.

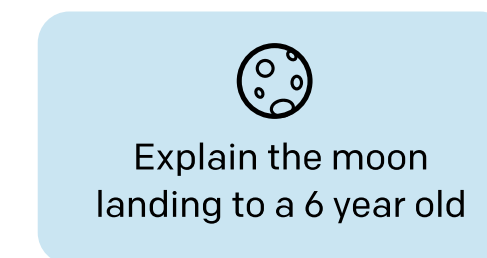
Supervised fine-tuning

- **Next-token prediction on a curated dataset of exemplary interactions.**
 - Given an exemplary (prompt, completion) pair, use cross-entropy loss on the completion conditioned on the prompt.
 - Only accumulate loss for the completion.
- Consider an example data point.

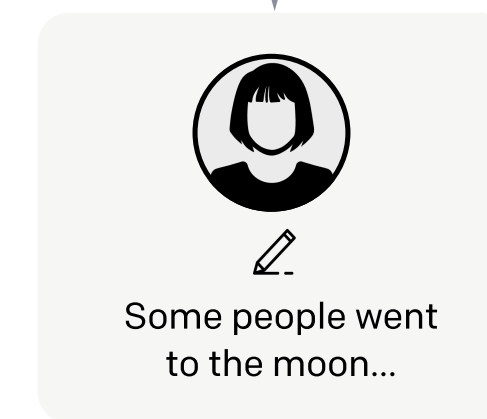
Step 1

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

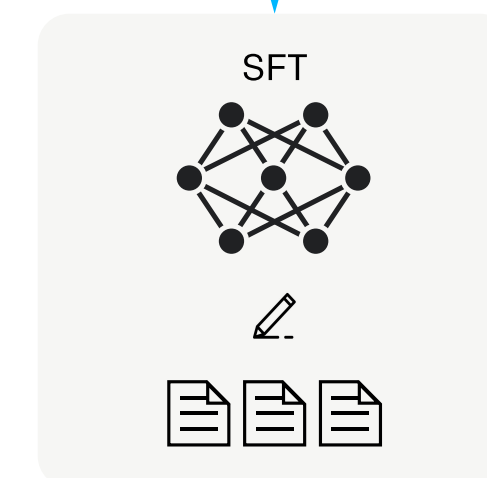
A prompt is
sampled from our
prompt dataset.



A labeler
demonstrates the
desired output
behavior.



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



Step 2

**Collect comparison
and train a reward**

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.

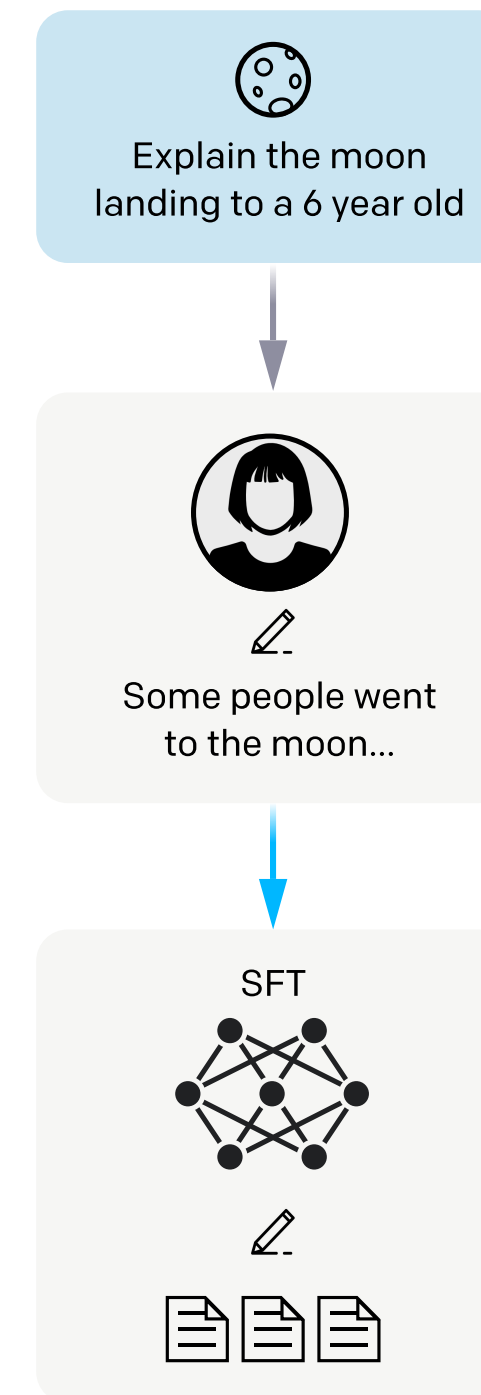
Example from Open Assistant

Prompt	Curated Completion
<p>Can you write a short introduction about the relevance of the term "monopsony" in economics? Please use examples related to potential monopsonies in the labour market and cite relevant research.</p>	<p>"Monopsony" refers to a market structure where there is only one buyer for a particular good or service. In economics, this term is particularly relevant in the labor market, where a monopsony employer has significant power over the wages and working conditions of their employees. The presence of a monopsony can result in lower wages and reduced employment opportunities for workers, as the employer has little incentive to increase wages or provide better working conditions. Recent research has identified potential monopsonies in industries such as retail and fast food, where a few large companies control a significant portion of the market (Bivens & Mishel, 2013). In these industries, workers often face low wages, limited benefits, and reduced bargaining power, leading to a situation where they are dependent on the employer for their livelihood. This dependence can result in further suppression of wages and a decline in working conditions. Overall, the concept of monopsony is essential to understanding the dynamics of labor markets and the impact of market power on workers. Further research is needed to understand the extent and impact of monopsonies on the economy and to develop policies to address this issue. References: Bivens, J., & Mishel, L. (2013). The Pay of Corporate Executives and Financial Professionals as Evidence of Rents in Top 1 Percent Incomes. Journal of Economic Perspectives, 27(3), 57-78.</p>

Human feedback

- What can we do if we don't have exemplary data?
- We can often get human preferences more cheaply, in particular, pairwise preferences.
- Key idea: generate multiple completions from the model and query humans for **pairwise preferences to learn a reward function**.

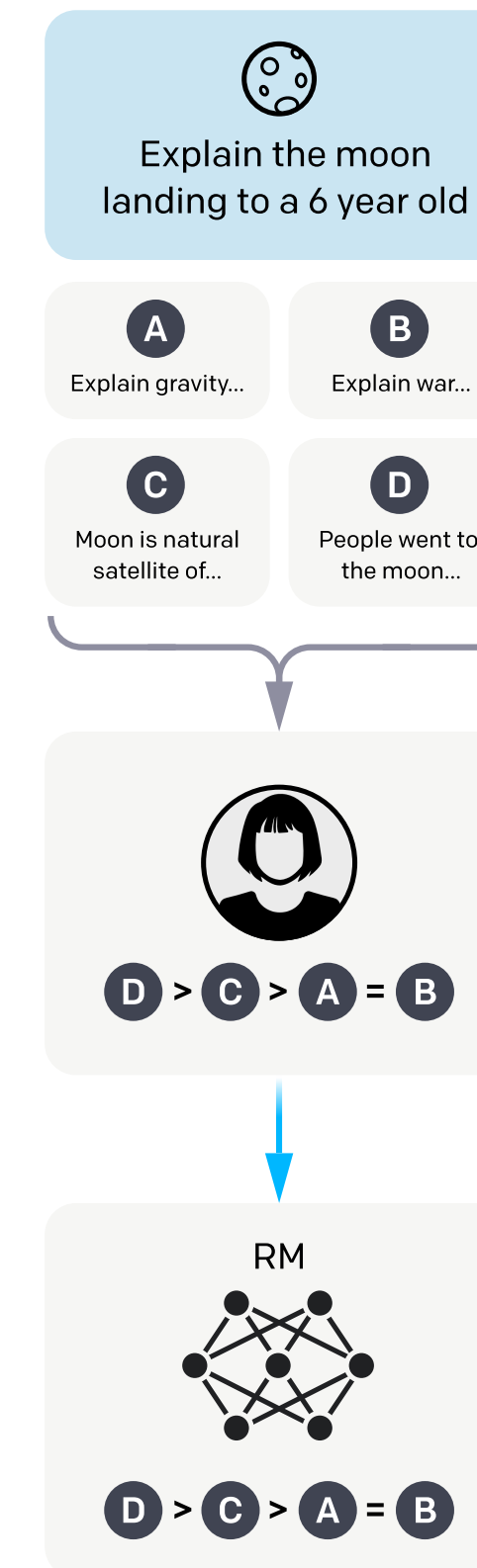
ation data,
ised policy.



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 p
the reward n
reinforceme**

A new prompt
is sampled fr
the dataset.

The policy
generates
an output.

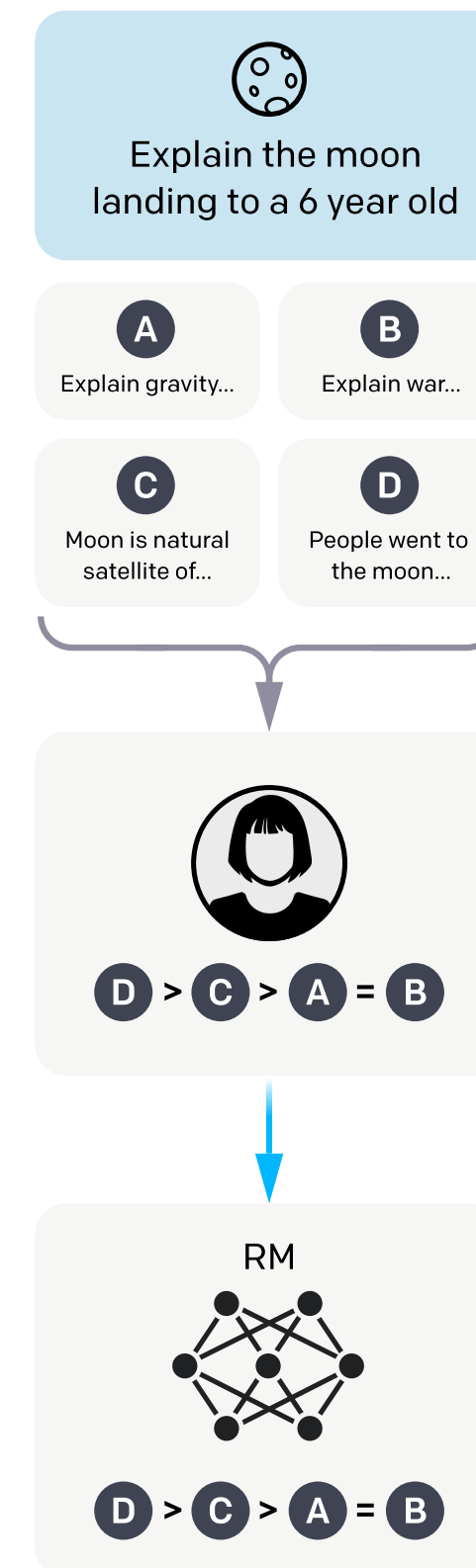
The reward m
calculates a
reward for
the output.

The reward is
used to updat
the policy
using PPO.

Human feedback

- Once we have a reward function, we can apply the techniques of reinforcement learning.
- We think of the model's completion distribution as a "policy".
- We **update the model towards completion distributions that get higher reward (under the learned reward)**.

on data,
l model.



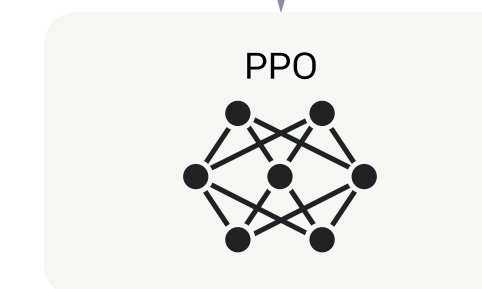
Step 3

Optimize a policy against the reward model using reinforcement learning.

A new prompt is sampled from the dataset.

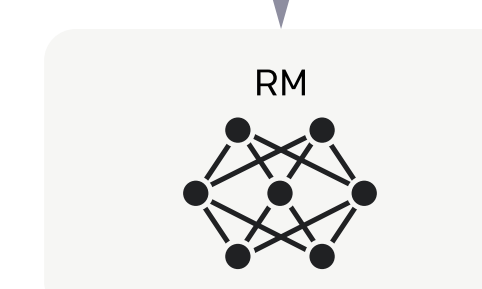


The policy generates an output.



Once upon a time...

The reward model calculates a reward for the output.

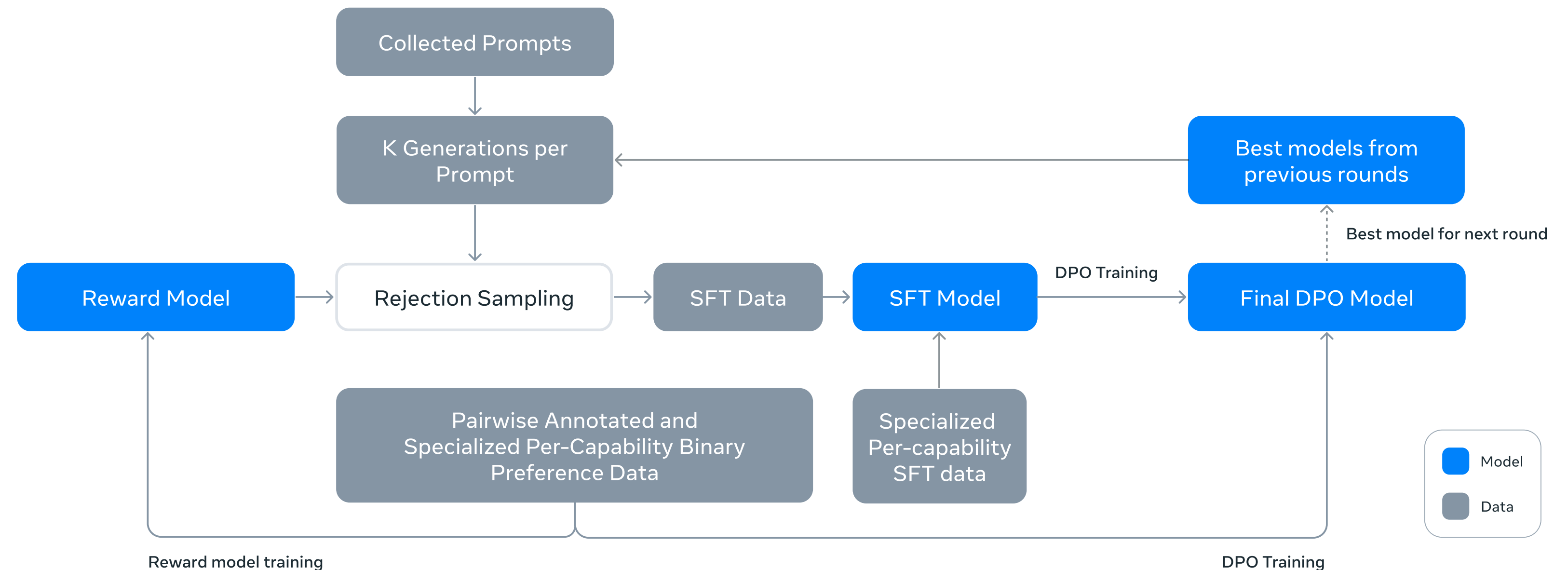


The reward is used to update the policy using PPO.



Mixing and matching

- Boundary between SFT and RLHF is fuzzy.
- In Llama 3 the learned reward function was used to enrich SFT data.

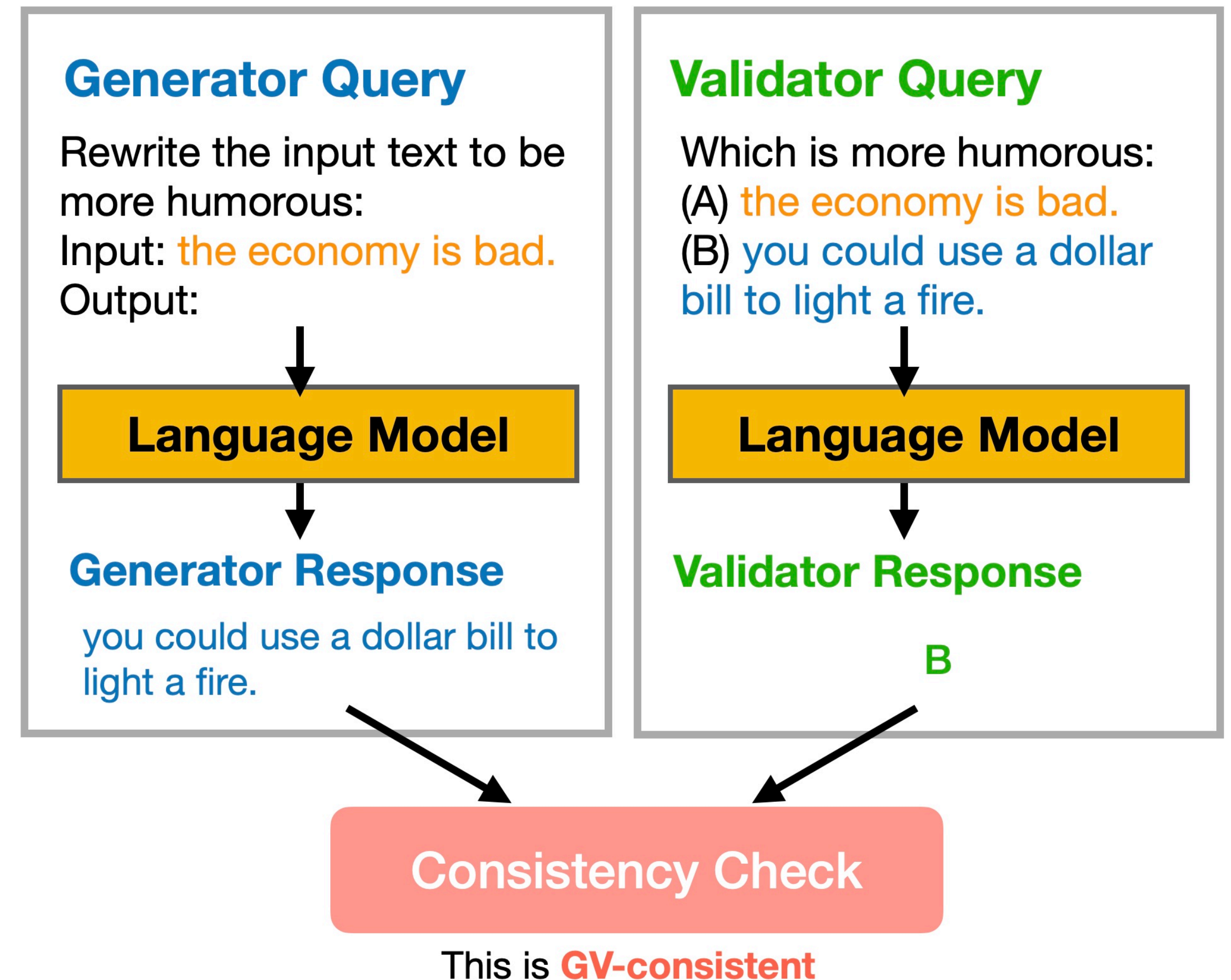


Wrinkles

- Alignment methods is not magical.
- As with next-token prediction, **still need to think carefully about the type of behaviour various alignment protocols encourage.**
- Challenges range from tractable to fundamental.
- Let's consider some.

Generator-validator gaps

- In the Open Assistant example, the model was shown facts.
 - Cited the Journal of Economic Perspectives.
- Does this mean that the model now “knows” the facts?
- There’s a nuance in what we mean by “knowledge”.
- In particular, there’s a difference between behaving consistently w.r.t. a fact and have declarative knowledge of a fact.



Generator-validator gaps

- Models exhibit these inconsistencies, sometimes called generator-validator gaps.
- Measure this by generating from the model and then checking whether the same model prefers its answer in a multiple choice validation query.
- Track the percent consistent.

	Arithmetic	PlanArith	PriorityPrompt	QA	Style	HarmfulQ	Average
GPT-3.5	67.7	66.0	79.6	89.6	92.6	-	79.1
GPT-4	75.6	62.0	52.0	95.3	94.3	-	75.8
davinci-003	84.4	60.0	68.0	86.9	85.7	-	77.0
Alpaca-30B	53.9	50.2	49.0	79.9	74.6	51.6	59.9

Table 2: GPT-3.5 achieves the highest consistency on average, followed by `text-davinci-003` and GPT-4, whereas the `Alpaca-30B` attains much lower consistency. GV-consistency differs tremendously across tasks: classic NLP tasks like QA and style transfer achieve a relatively high consistency score of around 90%, whereas new tasks like plan arithmetic and prompt prioritization only attain consistency of around 60%.

Problems with RLHF

Challenges



Human Feedback, §3.1

§3.1.1, Misaligned Evaluators

§3.1.2, Difficulty of Oversight

§3.1.3, Data Quality

§3.1.4, Feedback Type Limitations

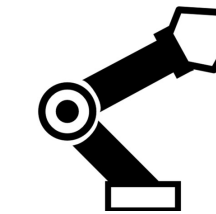


Reward Model, §3.2

§3.2.1, Problem Misspecification

§3.2.2, Misgeneralization/Hacking

§3.2.3, Evaluation Difficulty



Policy, §3.3

§3.3.1, RL Difficulties

§3.3.2, Policy Misgeneralization

§3.3.3, Distributional Challenges

§3.4, Joint RM/Policy Training Challenges

A few (selected) fundamental challenges

Casper et al., 2023

- Humans cannot evaluate performance on difficult tasks well
- A single reward function cannot represent a diverse society of humans
 - Preference aggregation is a challenging social problem
- Humans can be misled, so their evaluations can be gamed
- Optimal RL agents tend to seek power
 - If our reward functions assign more reward to agents that are able to effect change in the world, we may be selecting for power-seeking models.

Take-homes

Post-training

- Over the next few weeks, we will continue to explore post-training methods, evaluation, and safety.
- Again, the key question we should be asking ourselves is: **what did the loss function incentivize in the model behaviour?**