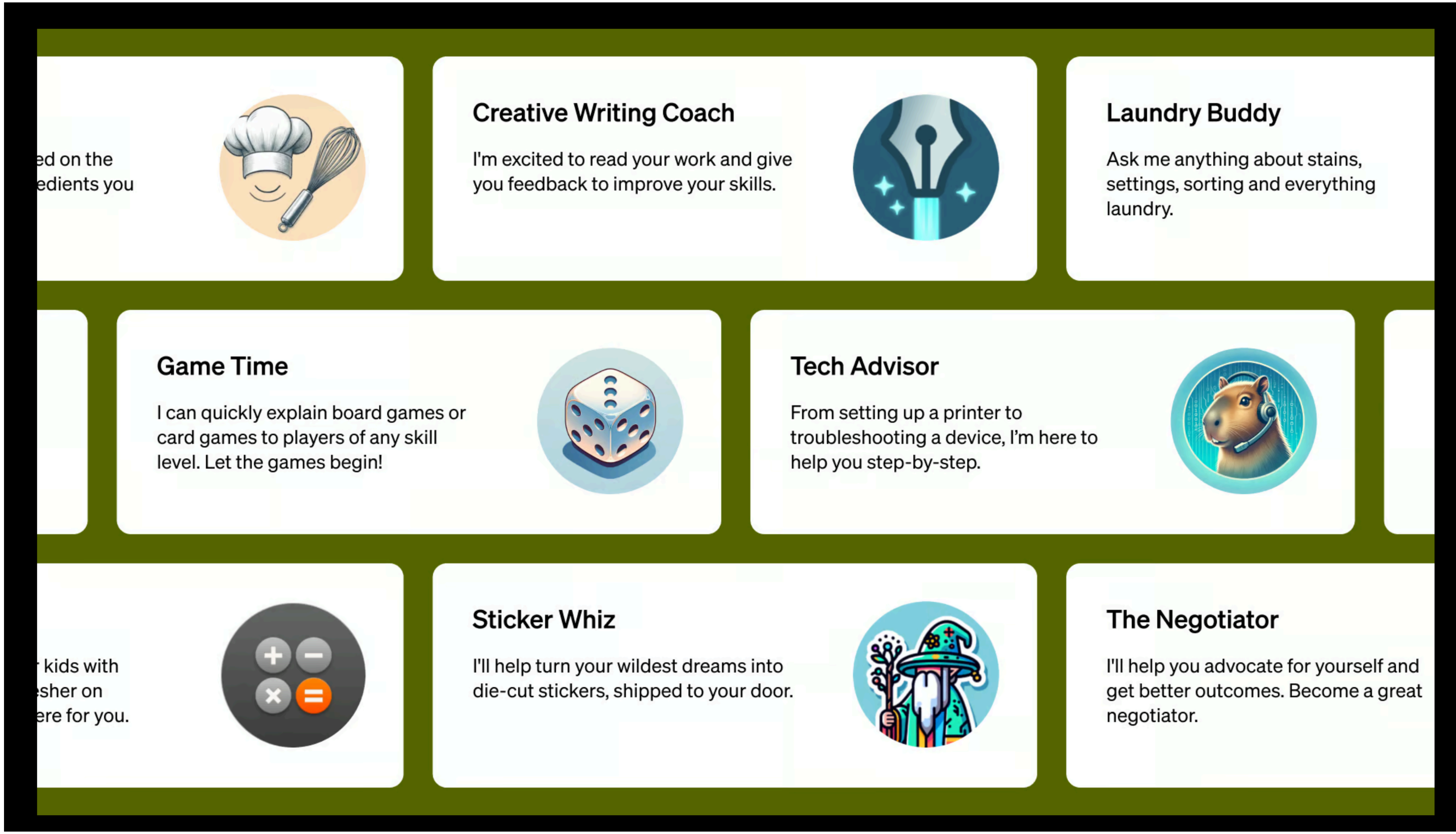
Understanding the Effects RLHF on LLM Generalisation and Diversity

Robert Kirk, Ishita Mediratta, Christoforos Nalmpantis, Jelena Luketina, Eric Hambro, Edward Grefenstette and Roberta Raileanu

https://arxiv.org/abs/2310.06452

Language Model Fine-tuning has many different use-cases



ed on the
edients you

Creative Writing Coach

I'm excited to read your work and give you feedback to improve your skills.

Laundry Buddy

Ask me anything about stains, settings, sorting and everything laundry.

Game Time

I can quickly explain board games or card games to players of any skill level. Let the games begin!

Tech Advisor

From setting up a printer to troubleshooting a device, I'm here to help you step-by-step.

kids with
sher on
ere for you.

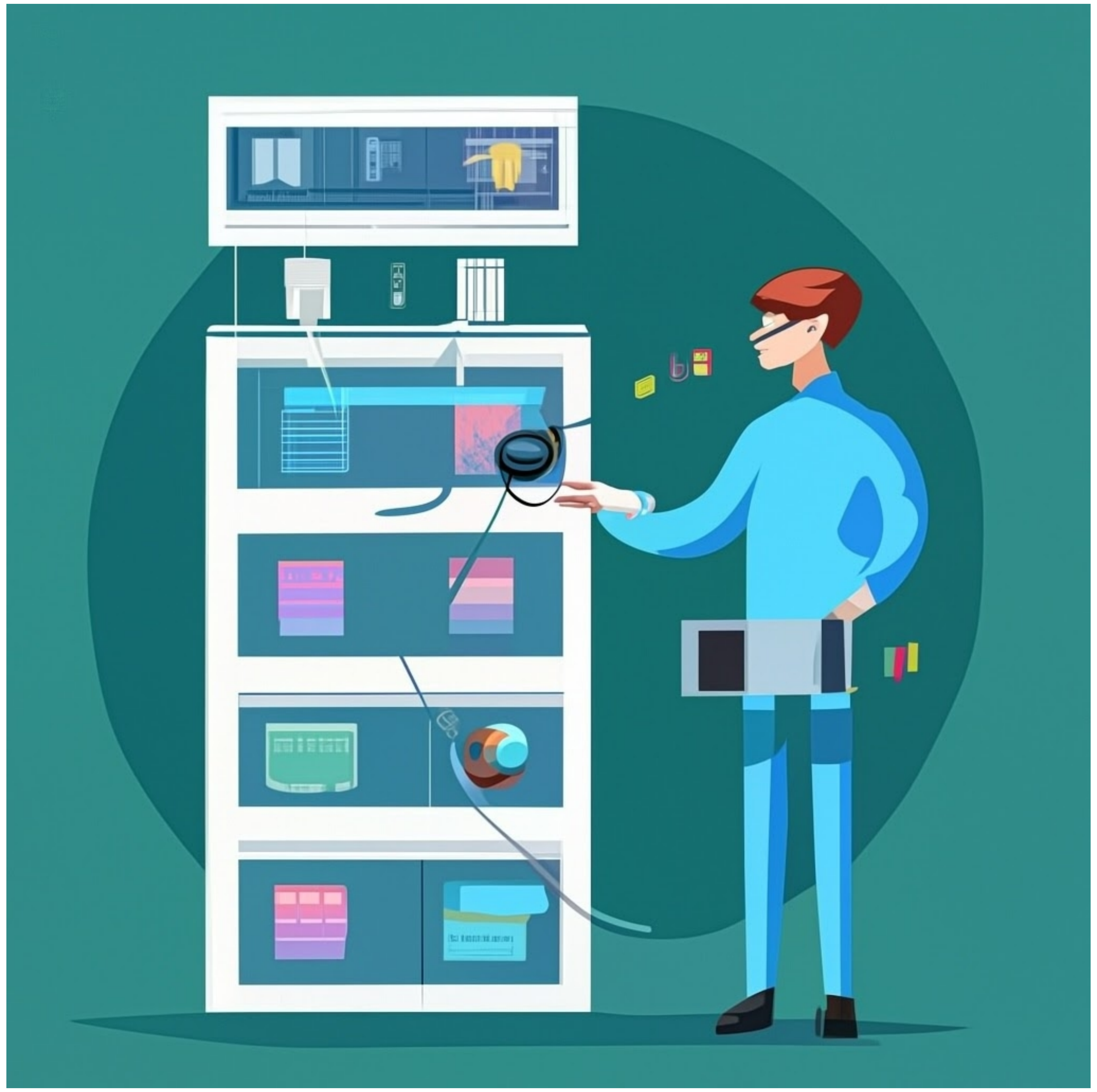
Sticker Whiz

I'll help turn your wildest dreams into die-cut stickers, shipped to your door.

The Negotiator

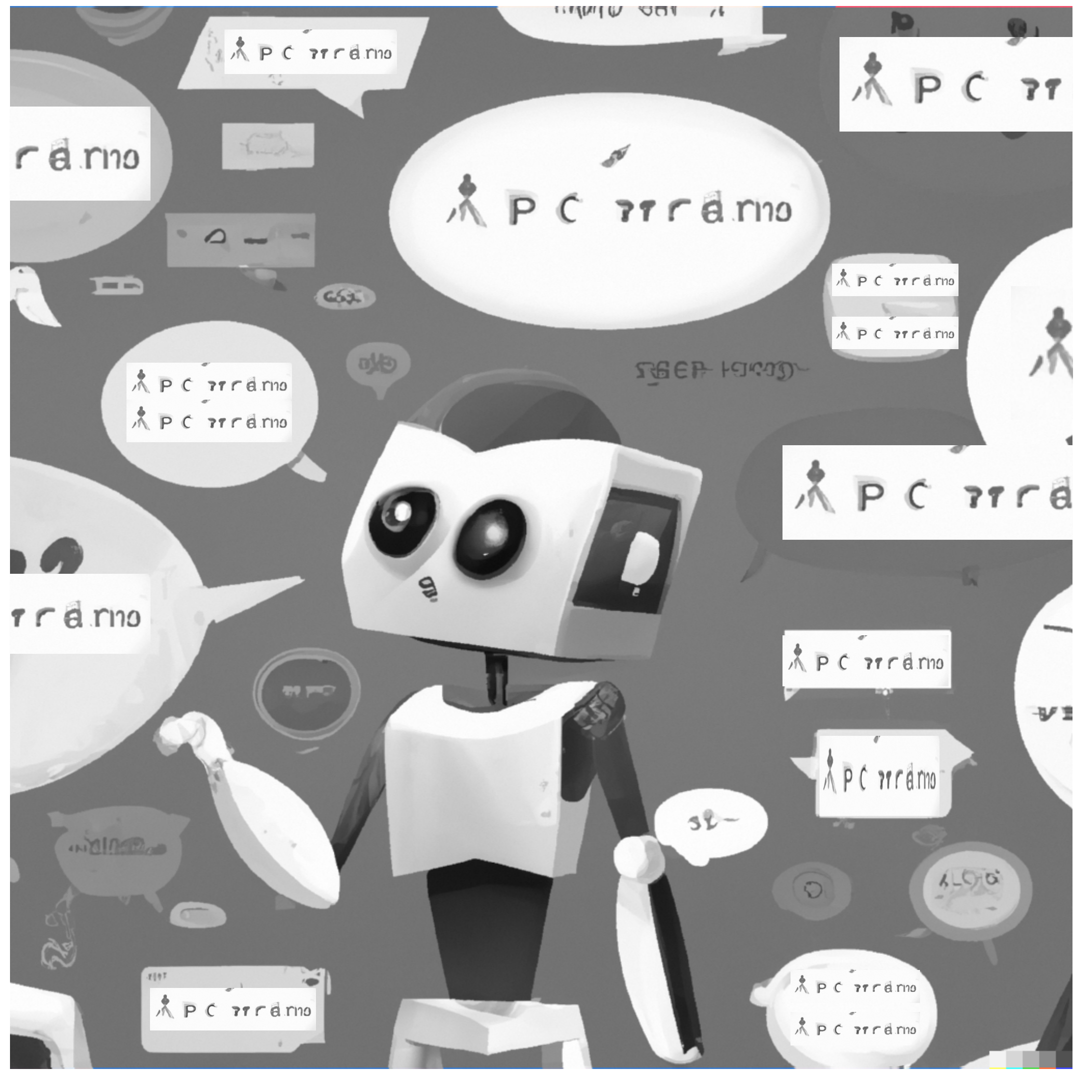
I'll help you advocate for yourself and get better outcomes. Become a great negotiator.

Multifaceted evaluation



Out of Distribution Generalisation

Multifaceted Evaluation



Output Diversity

Research Question

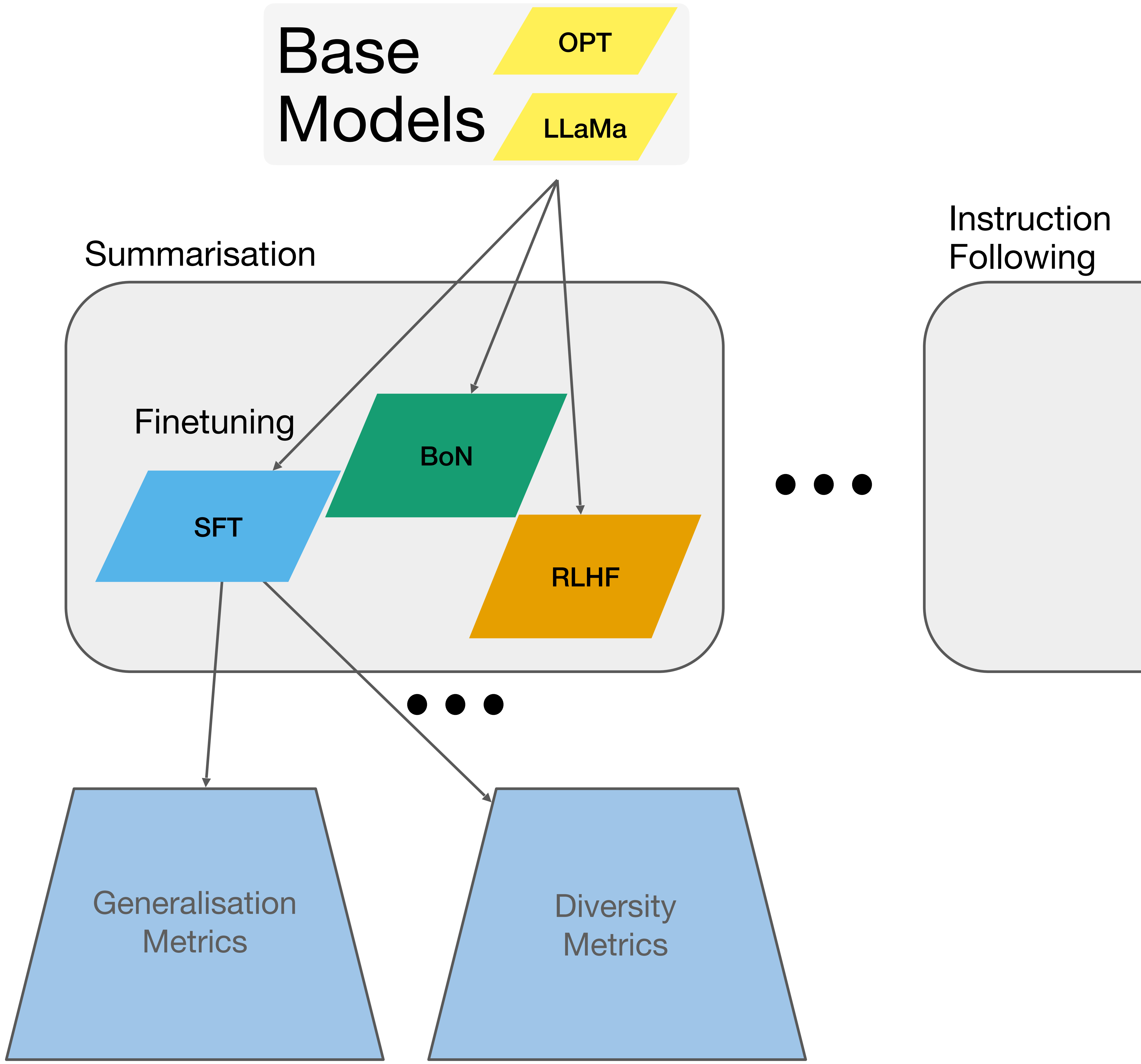
- How do LLM fine-tuning techniques affect the **OOD generalisation** and **output diversity**?
  - Supervised Fine-Tuning (SFT)
  - Reinforcement Learning from Human Feedback (RLHF)

Experiment Overview

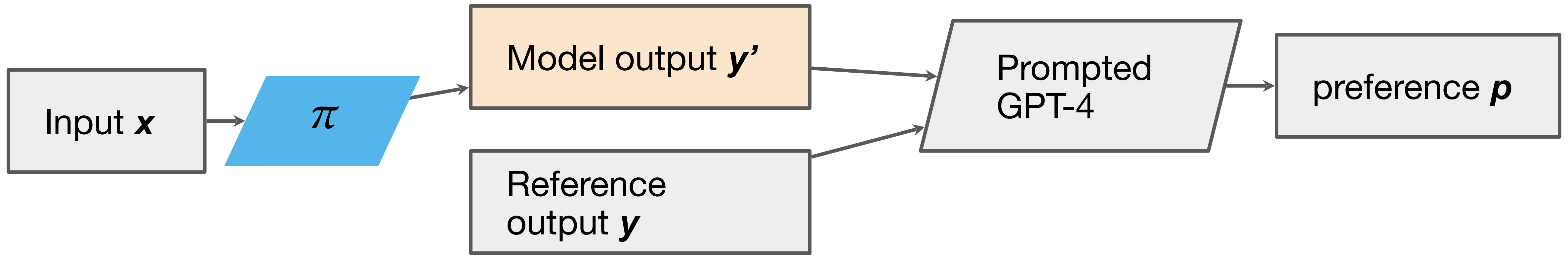
- Fine-tune LLMs with various techniques (SFT, RLHF, BoN)

- Measure OOD generalisation and output diversity of model

- Do this on multiple tasks



Evaluating Generalisation



Input ***x***

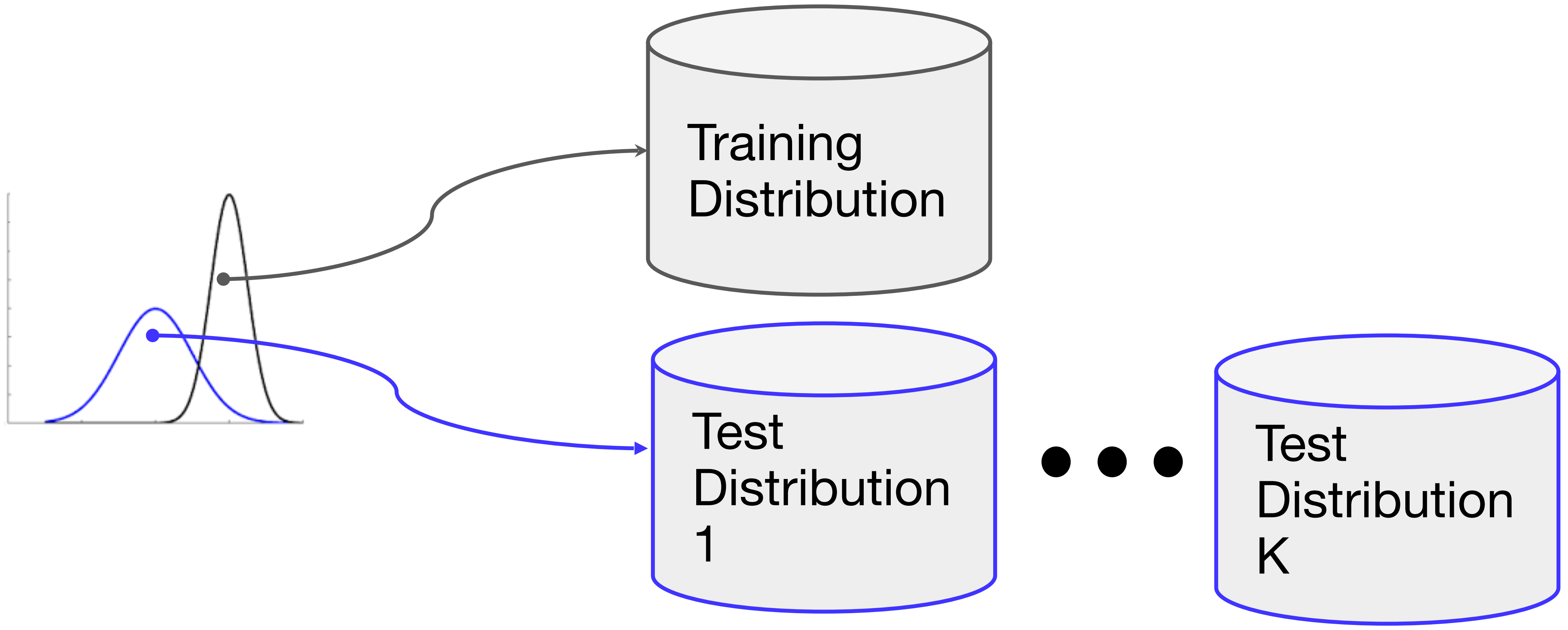
π

Model output ***y’***

Reference output ***y***

Prompted GPT-4

preference ***p***

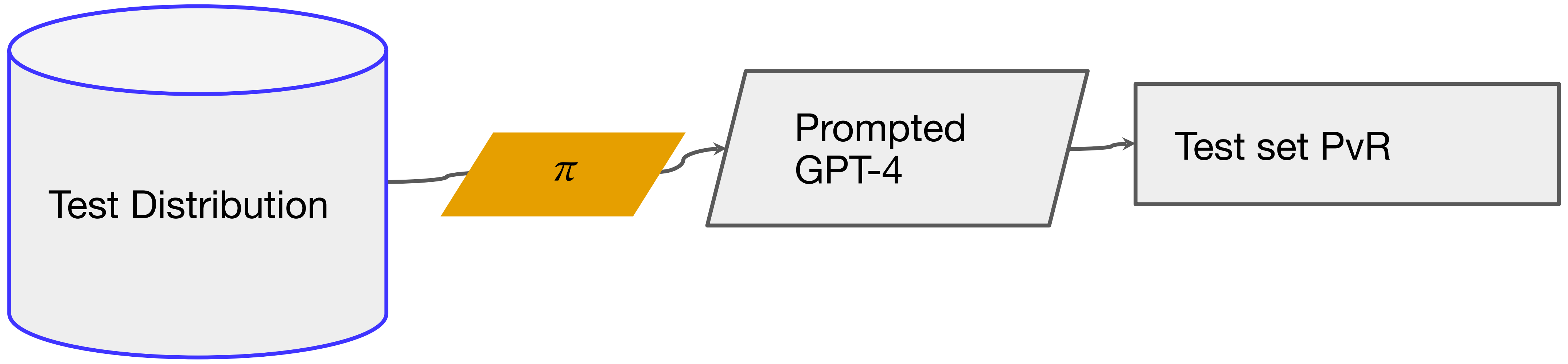


Training Distribution

Test Distribution 1

• • •

Test Distribution K



Test Distribution

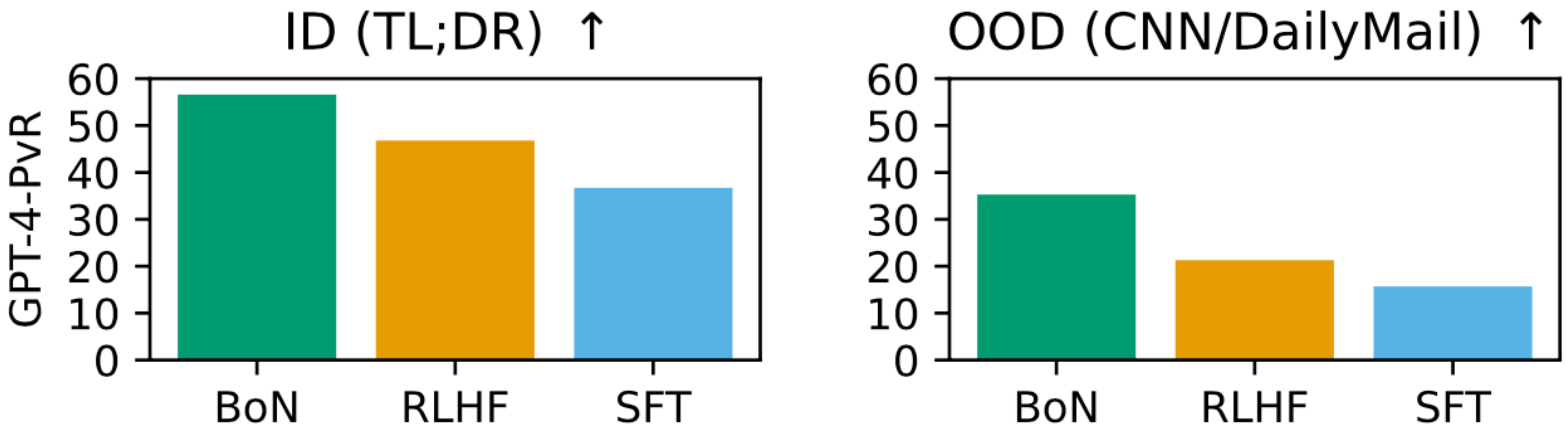
π

Prompted GPT-4

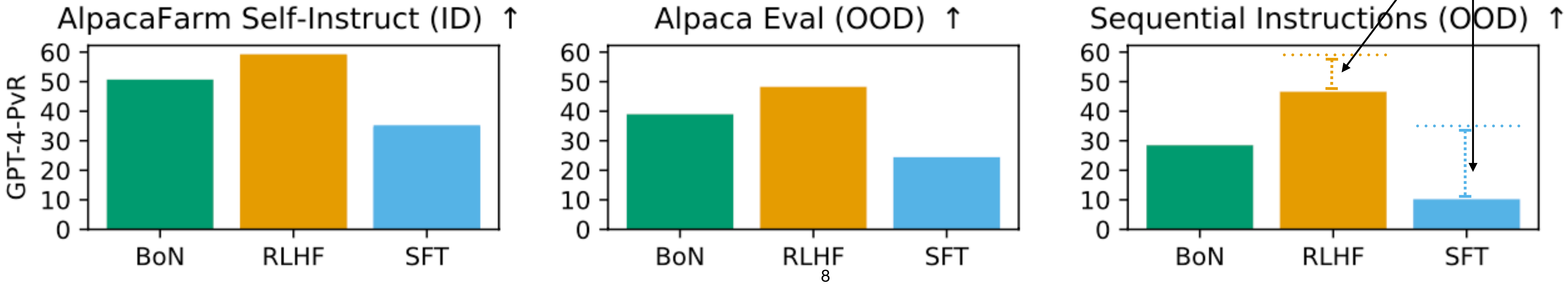
Test set PvR

Generalisation: RLHF is better than SFT

Summarisation

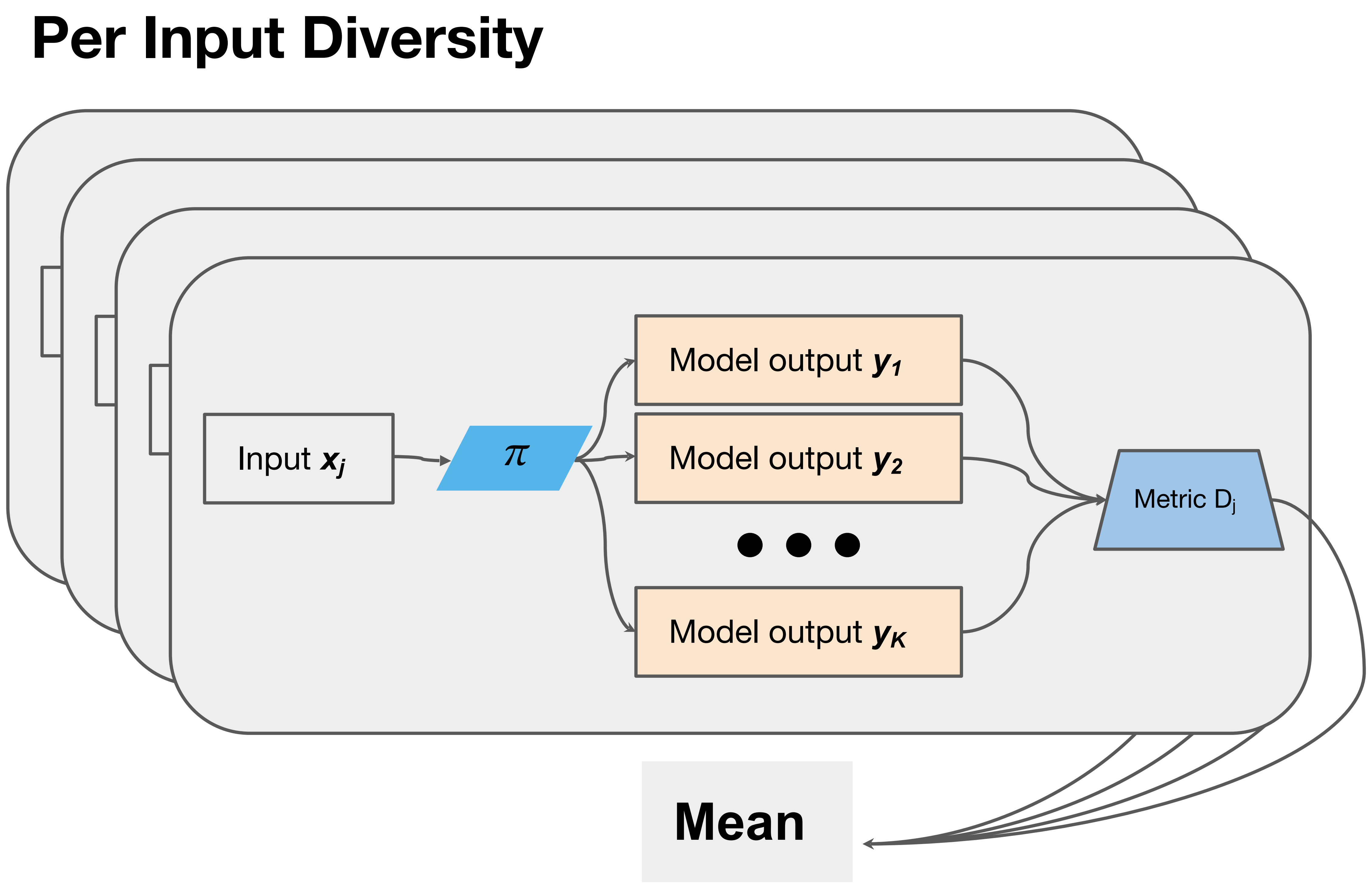


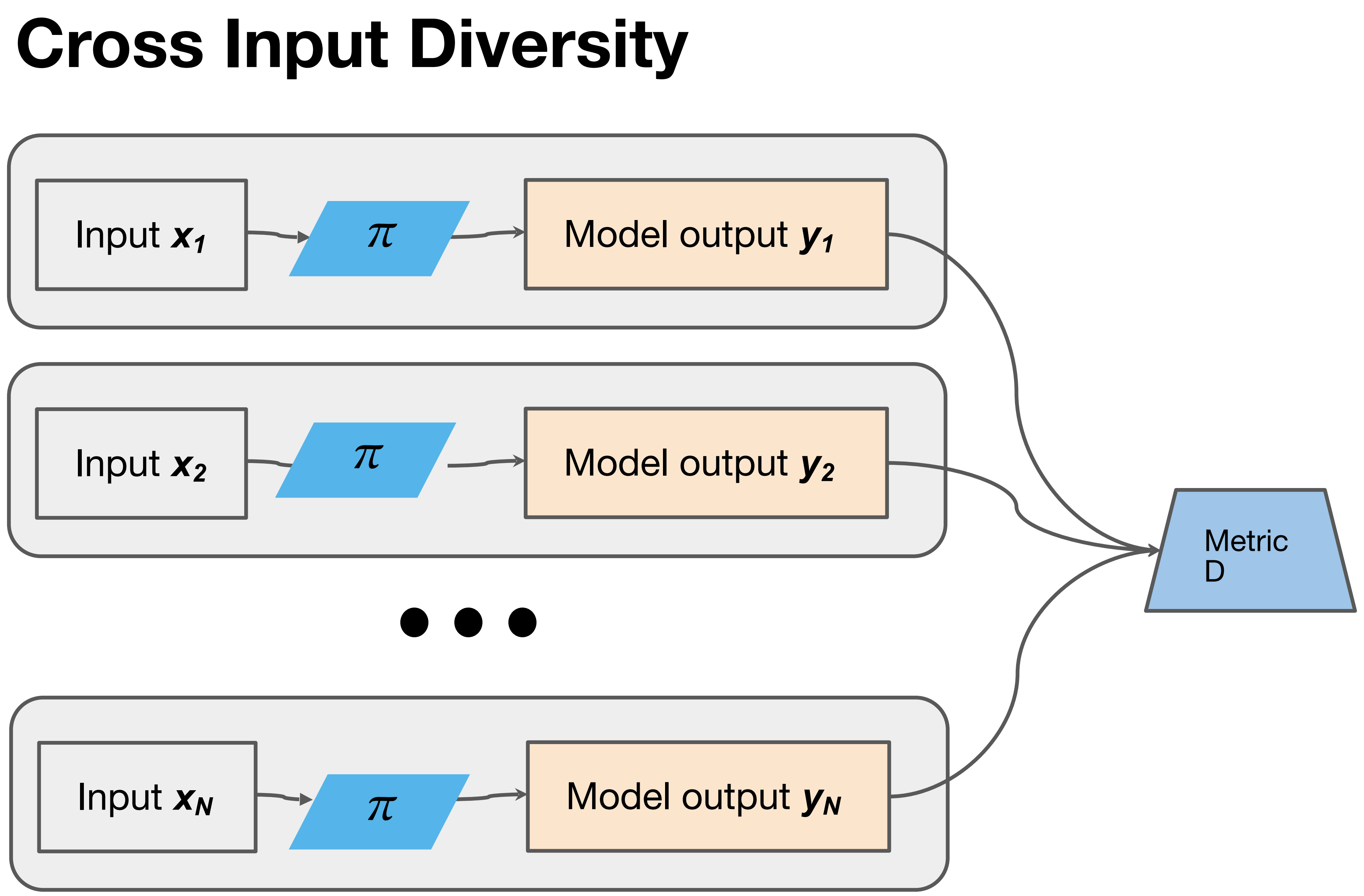
Instruction Following



Larger Generalisation Gap for SFT

Evaluating Output Diversity

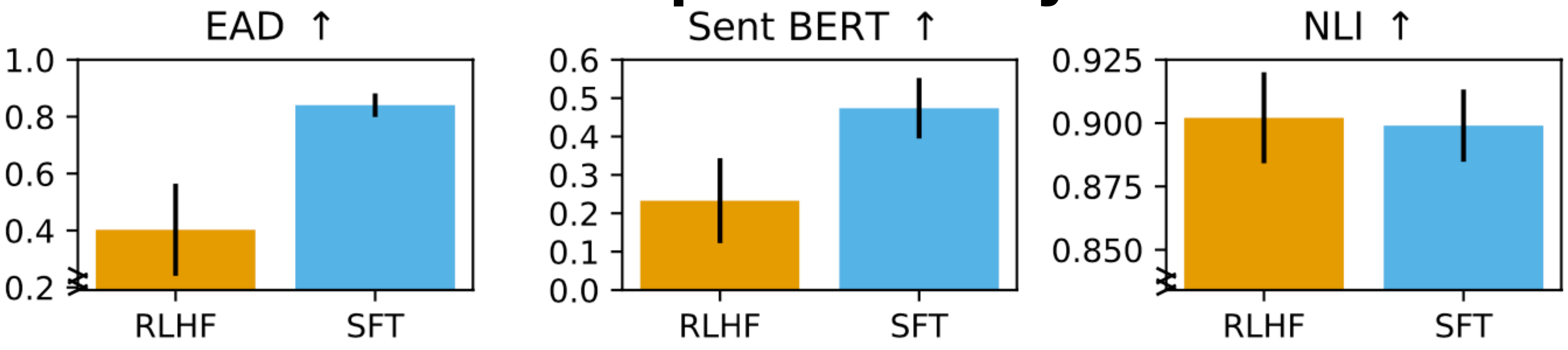




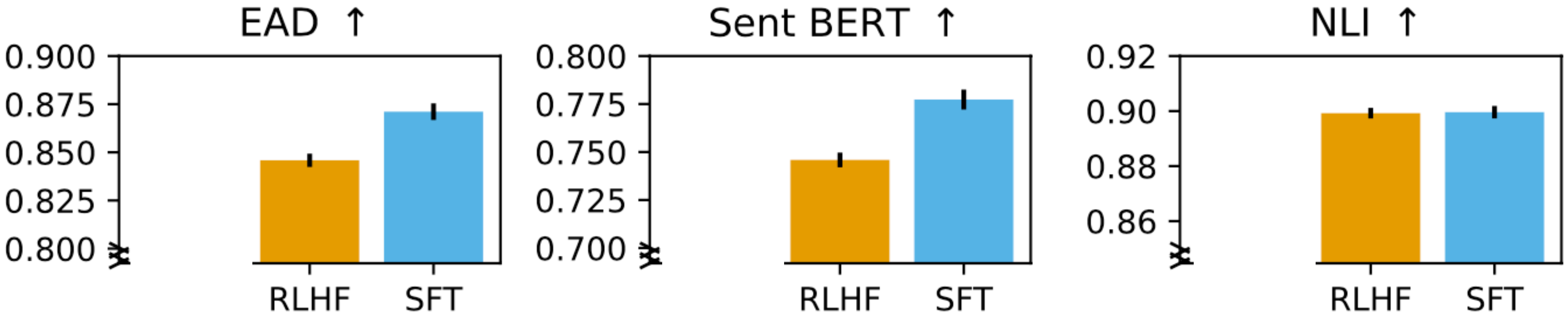
- Expectation Adjusted Distinct N-grams (**Form diversity**) (Li et al. 2016)
- SentBERT (**Topic or Content diversity**) (Reimers et al. 2019)
- NLI sample (**Logical Diversity**) (Stasaski et al. 2022)

Diversity: RLHF is worse than SFT

Per-input Diversity

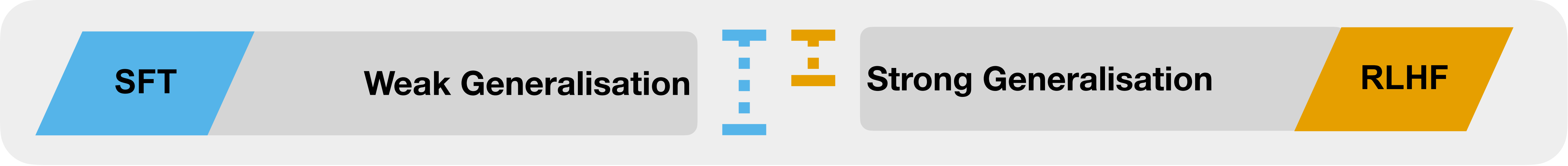


Cross-input Diversity



Key Takeaways

RLHF vs SFT? It depends!

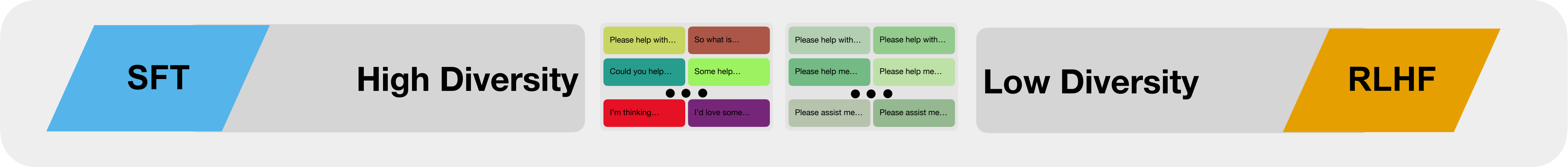


SFT

Weak Generalisation

Strong Generalisation

RLHF



SFT

High Diversity

Please help with…

So what is…

Could you help…

Some help…

I'm thinking…

I'd love some…

Please help with…

Please help with…

Please help me…

Please help me…

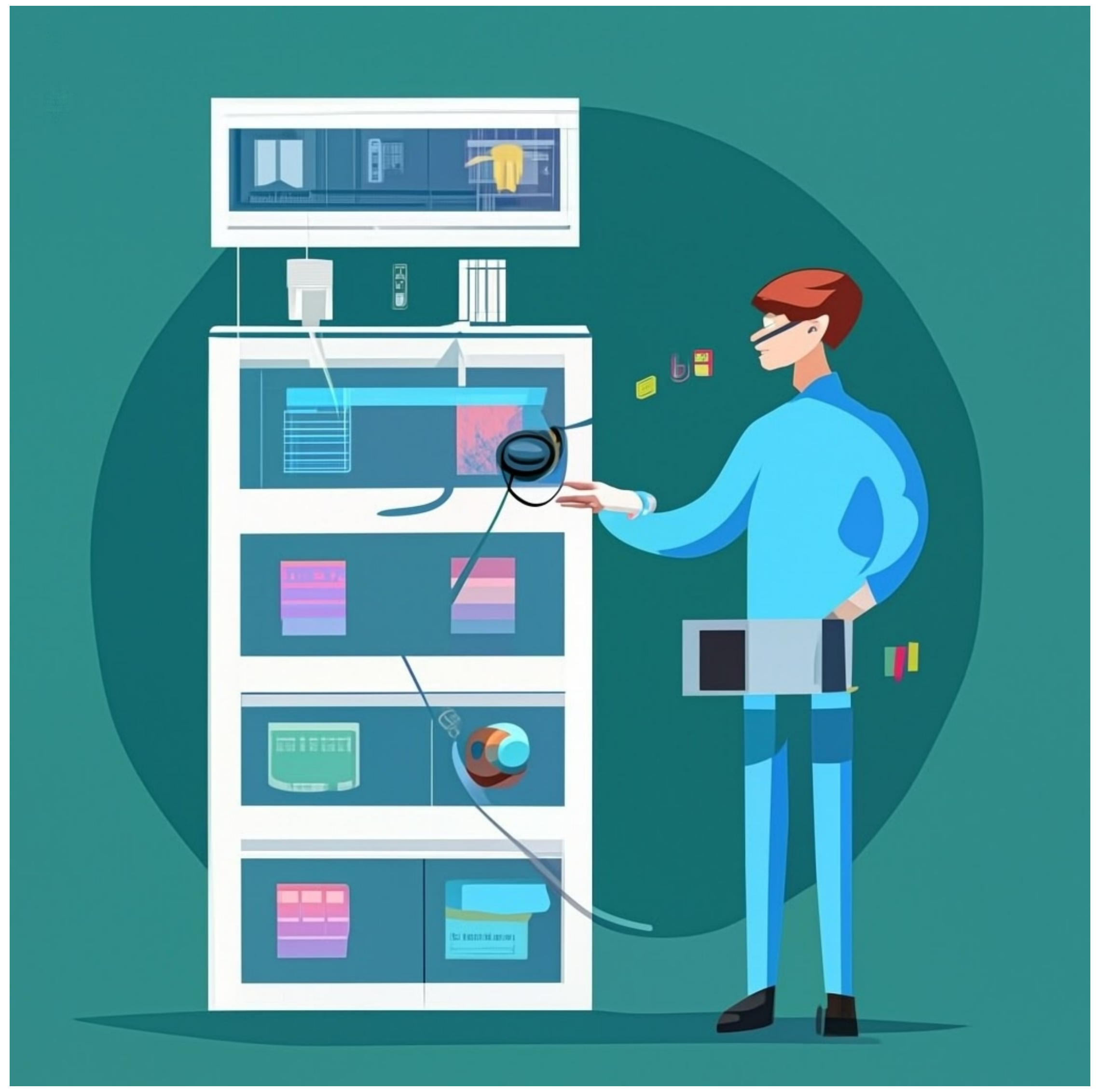
Please assist me…

Please assist me…

Low Diversity

RLHF

Multifaceted evaluation is important



Thanks for listening!

Understanding the Effects RLHF on LLM Generalisation and Diversity

Robert Kirk, Ishita Mediratta, Christoforos Nalmpantis, Jelena Luketina, Eric Hambro, Edward Grefenstette and Roberta Raileanu

https://arxiv.org/abs/2310.06452
https://github.com/facebookresearch/rlfh-gen-div
Poster Session 3: Wed 8th May 10:45 a.m — 12:45 p.m