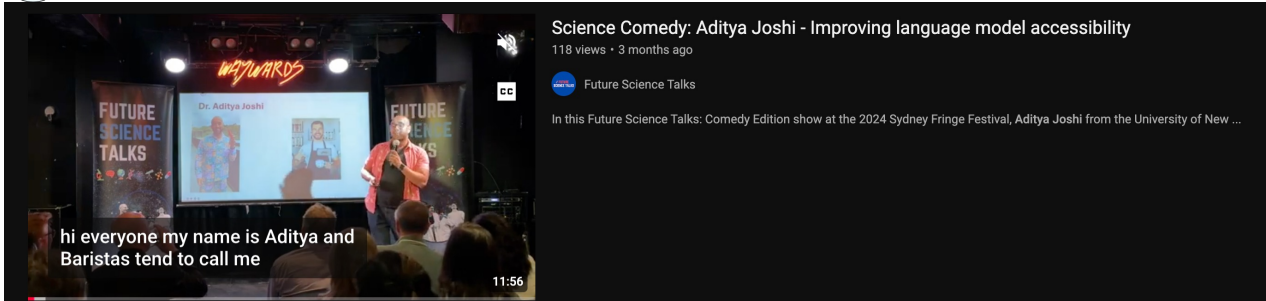


**UNSW**  
SYDNEY



Science Comedy: Aditya Joshi - Improving language model accessibility  
118 views • 3 months ago  
Future Science Talks

In this Future Science Talks: Comedy Edition show at the 2024 Sydney Fringe Festival, Aditya Joshi from the University of New ...

hi everyone my name is Aditya and Baristas tend to call me

11:56

# Week 9b: COMP6713

17 April 2025

Note to self: **TEACH IT RIGHT BUT USE MINIMUM WORDS ON SLIDES**

1



**UNSW**  
SYDNEY



actions



2



UNSW  
SYDNEY







rewarding




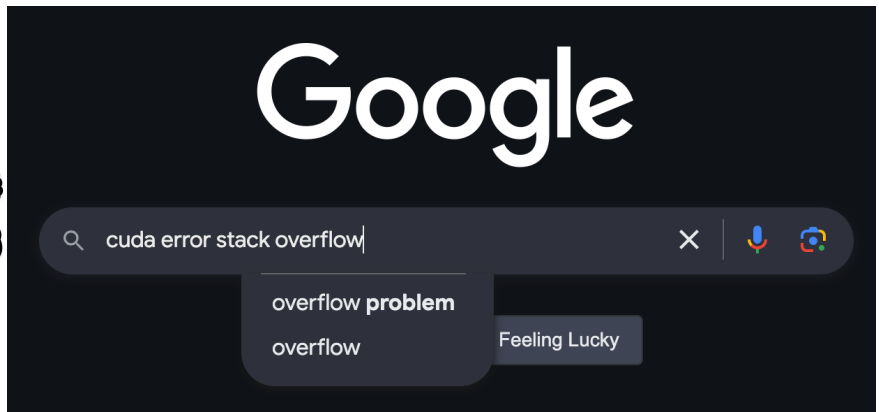
3



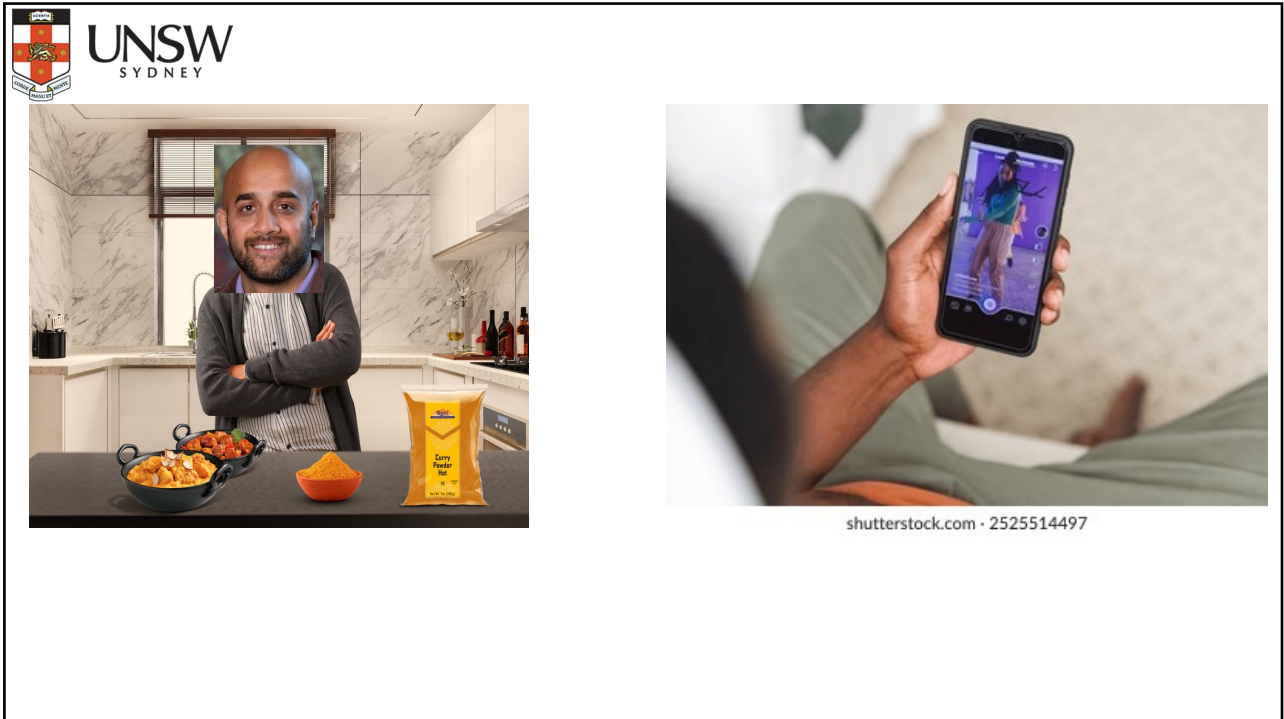
UNSW  
SYDNEY



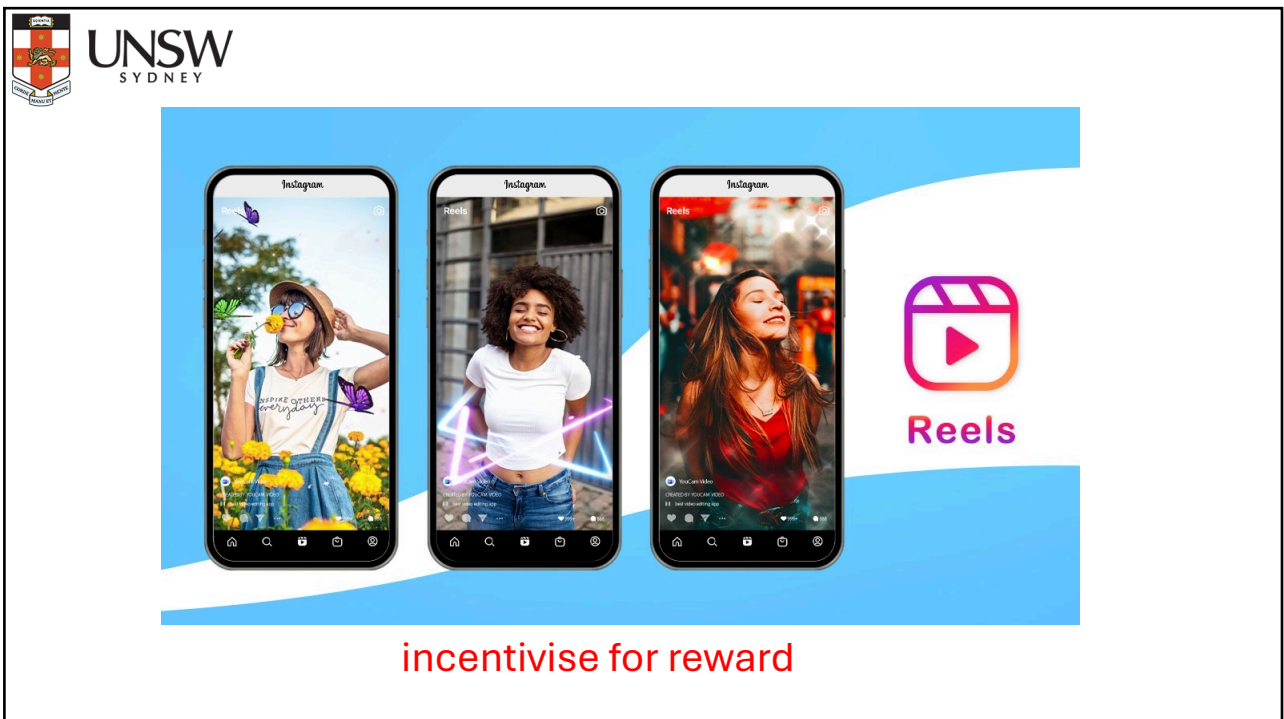




4



5



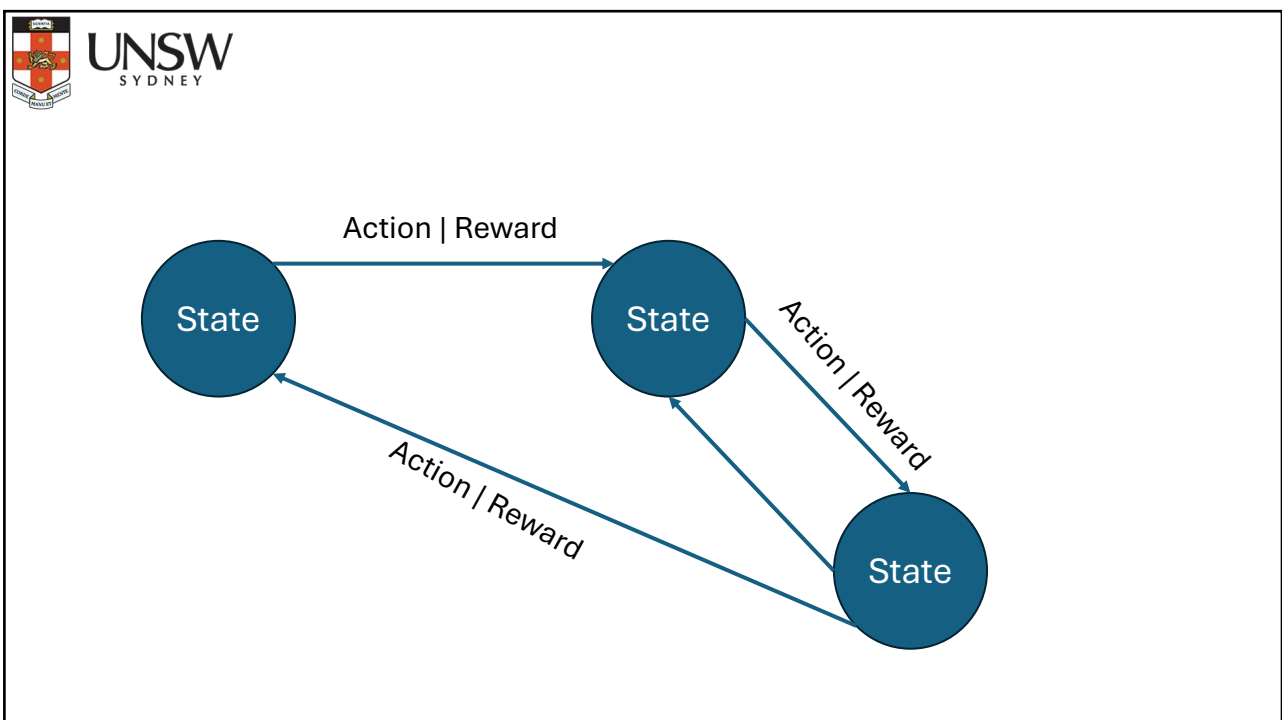
6

UNSW SYDNEY

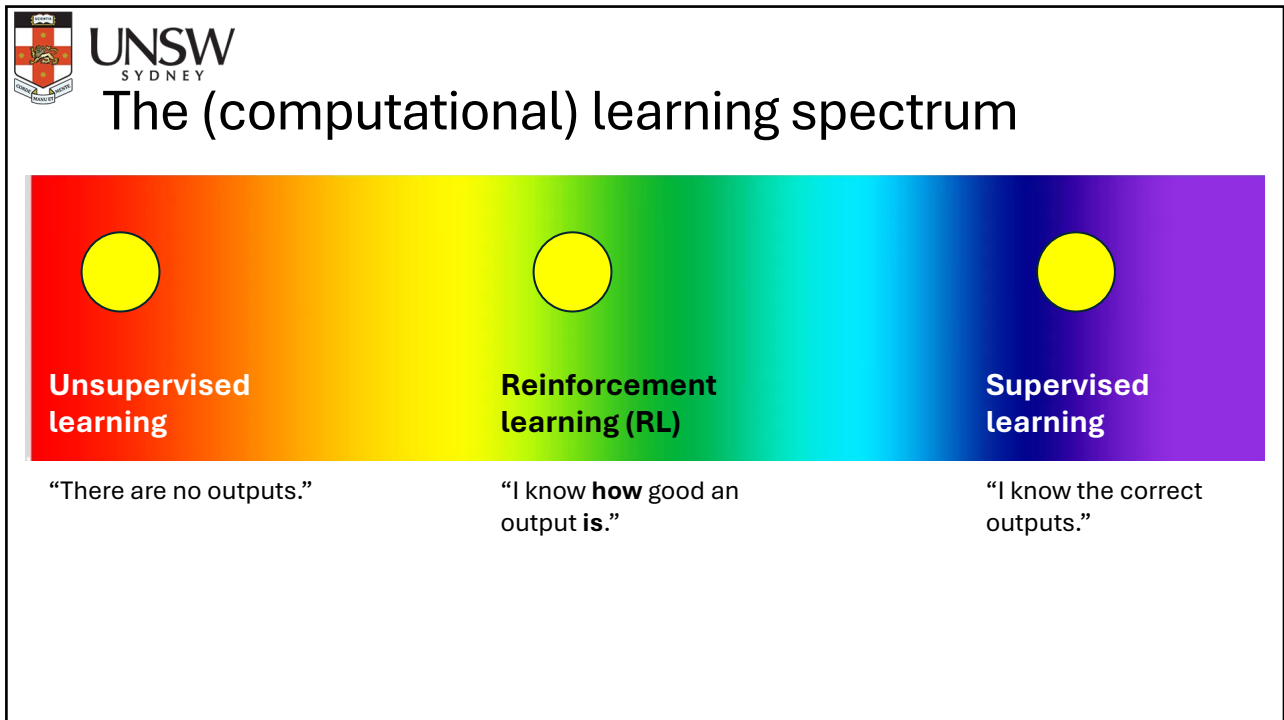


.... what the hell is happening, bro?!

7



8



9


Luca learns from books AND his parents' smile

The slide features a photograph of a classroom with rows of desks and chairs. To the right of the photo is a stylized icon of a house with a person inside, a smiling face above, and a sad face with a radio wave above, symbolizing learning from both books and family. At the bottom, there are three flags: the Australian flag, a stylized person icon, and the Italian flag.

Luca Leonardo Moretti


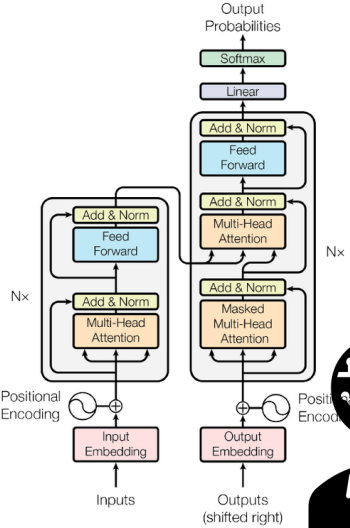

10

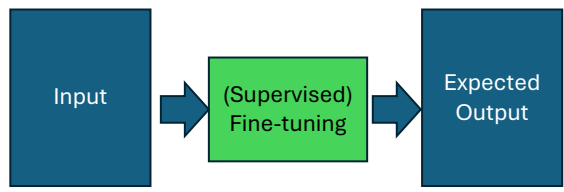




Write a summary of this paper:  
<https://aclanthology.org/2025.coling-main.471.pdf>


The paper titled "Can We Afford The Perfect Prompt? Balancing Cost and Accuracy with the Economical Prompting Index" introduces the Economical Prompting Index (EPI), a novel metric that combines accuracy scores with token consumption. This index is adjusted by a user-specified cost concern level to reflect different resource constraints. The study evaluates six advanced prompting techniques, including Chain-of-Thought and Self-Consistency, across ten widely-used language models and four diverse datasets. The findings suggest that simpler techniques like Chain-of-Thought often surpass more complex methods like Self-Consistency in terms of cost-effectiveness, especially in resource-constrained scenarios




Luca Leonardo Moretti (LLM)

11




LucaLM learns from books.

Can LucaLM also learn from rewards?

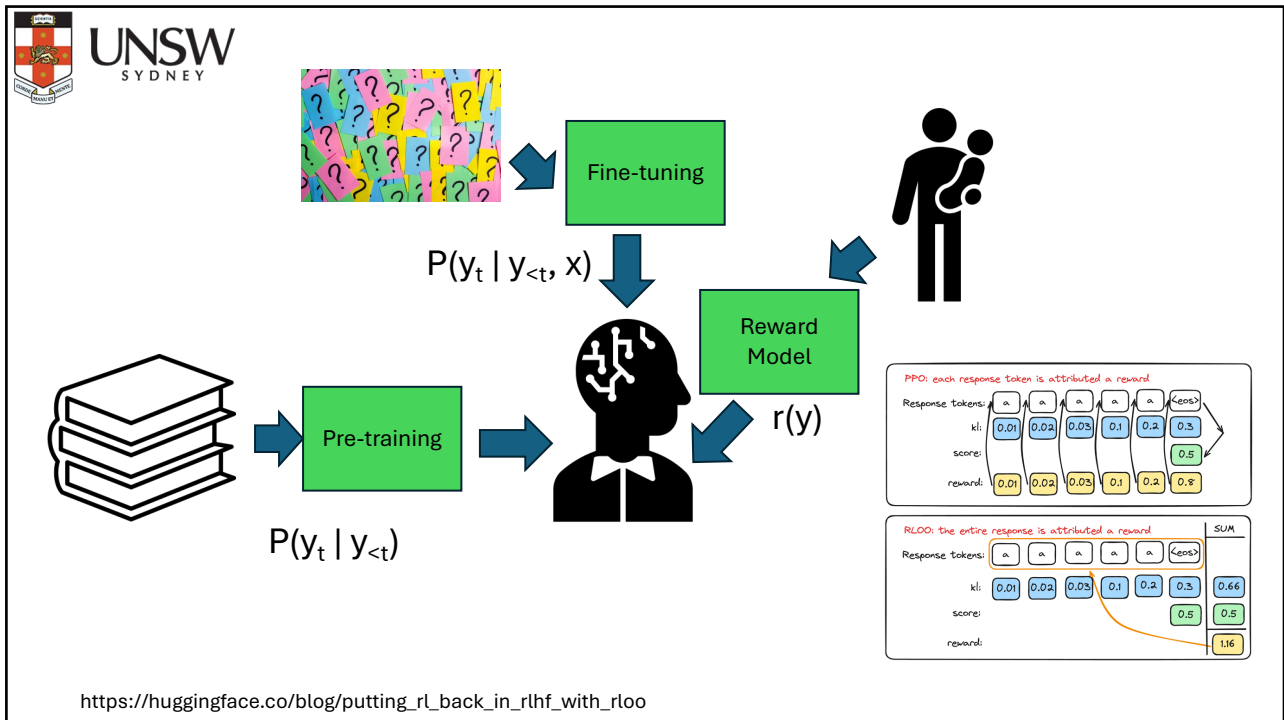


Luca Leonardo Moretti

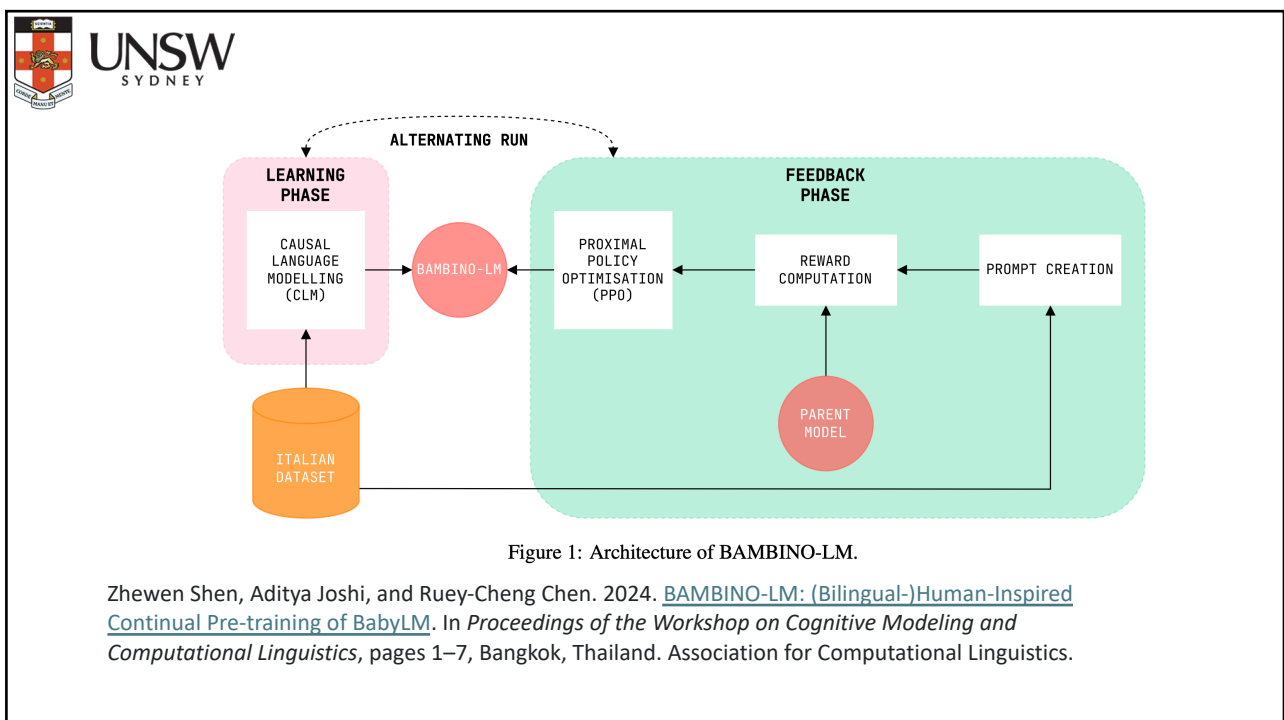


Luca Leonardo Moretti (LLM)

12



13



14



UNSW  
SYDNEY

## LLMs + RL

- LLM Fine-tuning
  - Correctness
  - Learn to produce the CORRECT output
- RL
  - Behaviours
  - Learn to produce GOOD outputs and minimise BAD outputs

15



UNSW  
SYDNEY

## Behaviours

### Abstract

While large-scale unsupervised language models (LMs) learn broad world knowledge and some reasoning skills, achieving precise control of their behavior is difficult due to the completely unsupervised nature of their training. Existing methods for gaining such steerability collect human labels of the relative quality of model generations and fine-tune the unsupervised LM to align with these preferences, often with reinforcement learning from human feedback (RLHF). However, RLHF is a complex and often unstable procedure, first fitting a reward model that reflects the human preferences, and then fine-tuning the large unsupervised LM using reinforcement learning to maximize this estimated reward without drifting too far from the original model. In this paper we introduce a new parameterization of the reward model in RLHF that enables extraction of the corresponding optimal policy in closed form, allowing us to solve the standard RLHF problem with only a simple classification loss. The resulting algorithm, which we call *Direct Preference Optimization* (DPO), is stable, performant, and computationally lightweight, eliminating the need for sampling from the LM during fine-tuning or performing significant hyperparameter tuning. Our experiments show that DPO can fine-tune LMs to align with human preferences as well as or better than existing methods. Notably, fine-tuning with DPO exceeds PPO-based RLHF in ability to control sentiment of generations, and matches or improves response quality in summarization and single-turn dialogue while being substantially simpler to implement and train.

### Summary 1:

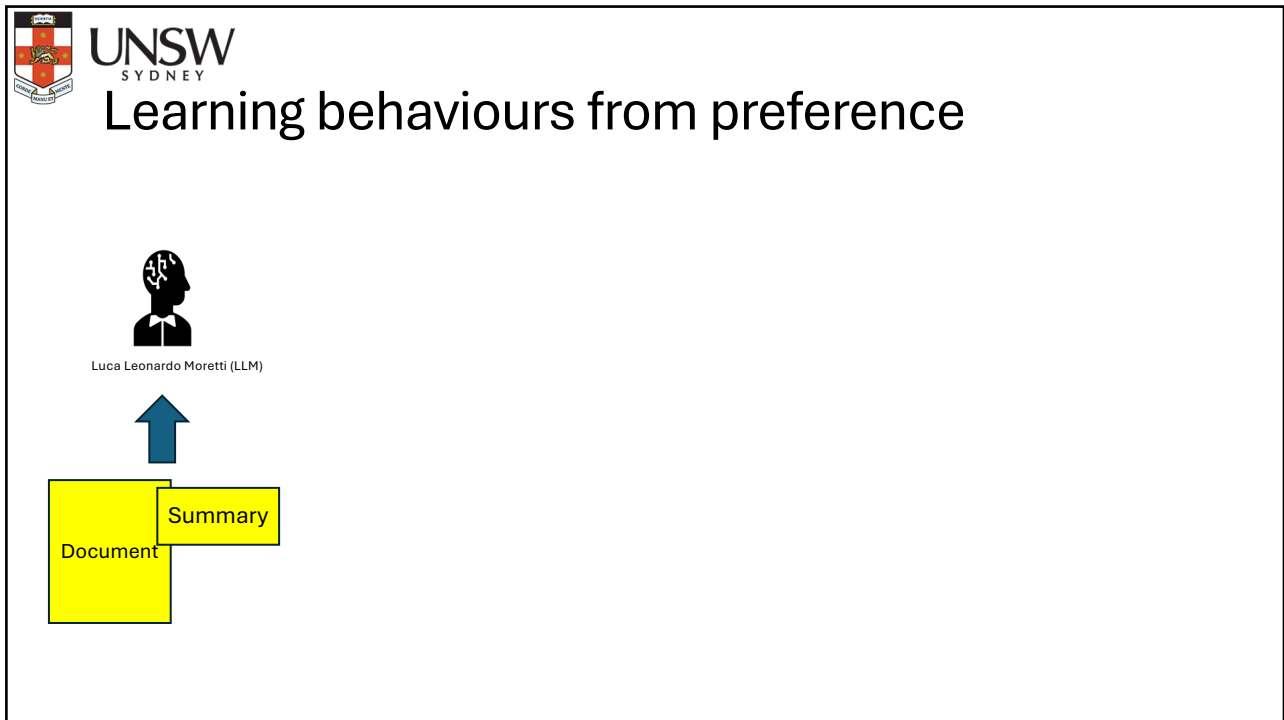
Direct Preference Optimization enhances language models' ability to emulate behaviours using human preferences. It follows a two step process, akin to RLHF.

### Summary 2:

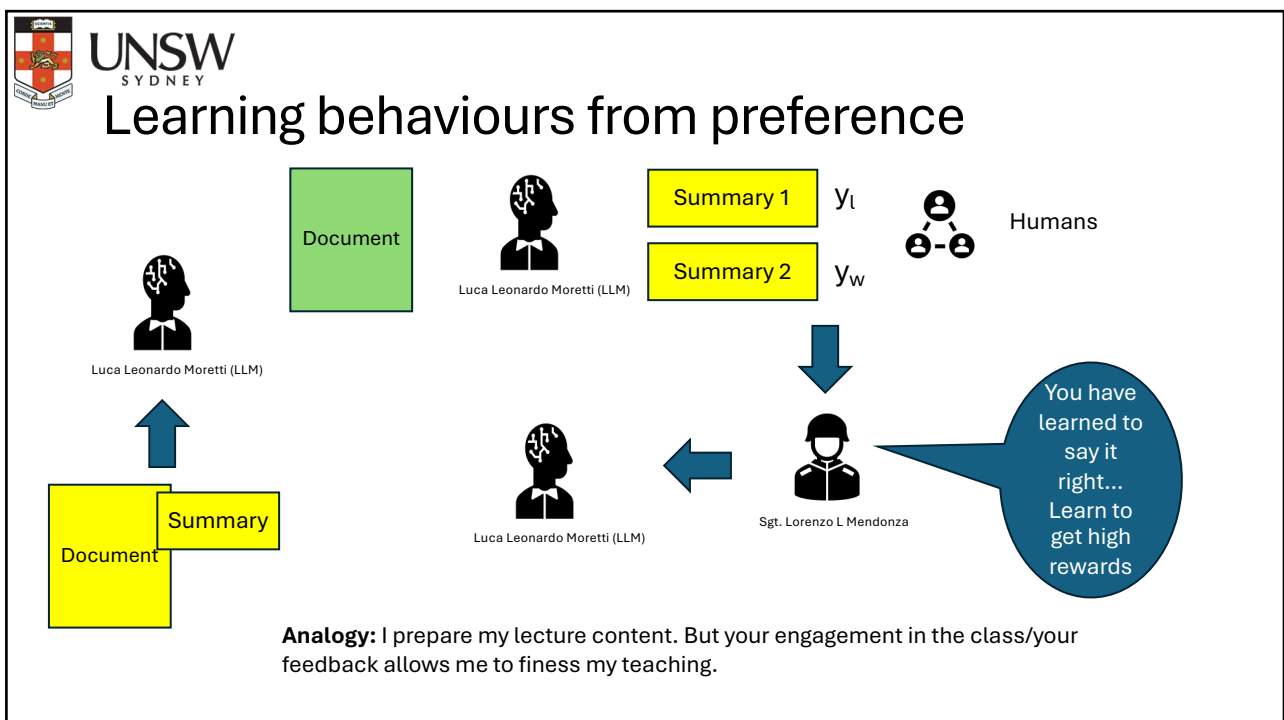
The paper presents Direct Preference Optimization which is a lightweight approach to achieve control of their behaviour.

16

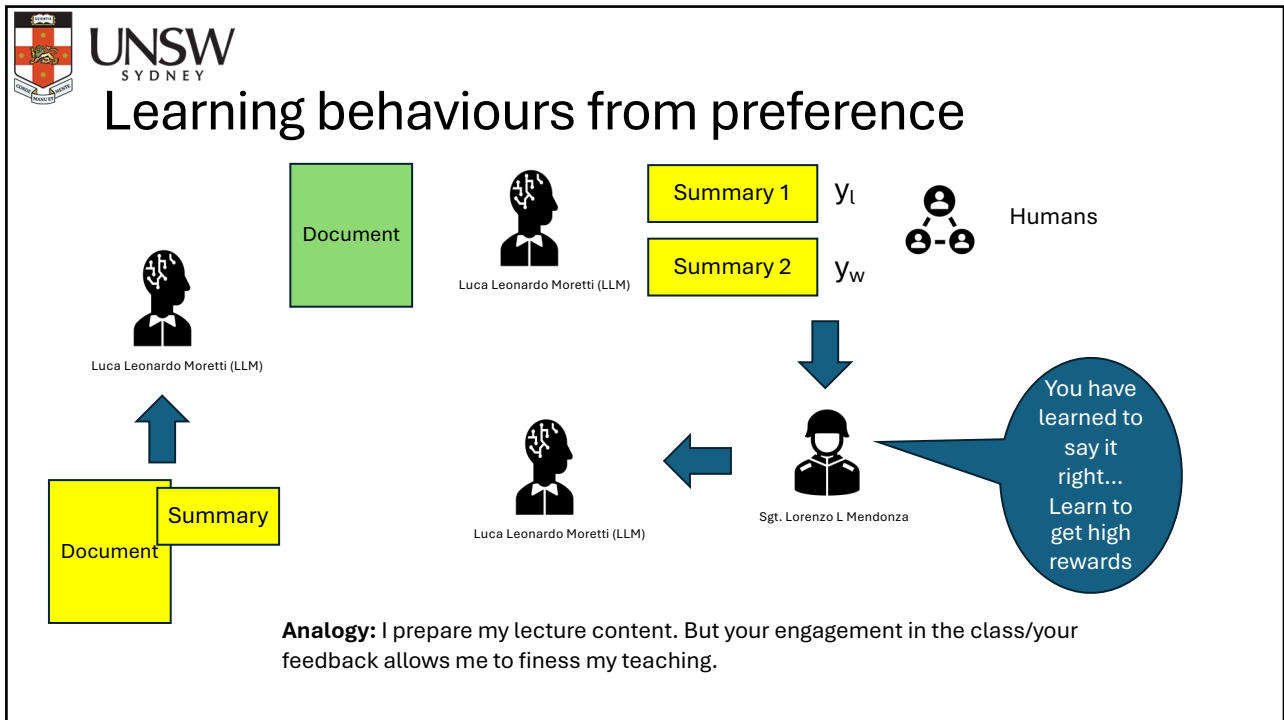




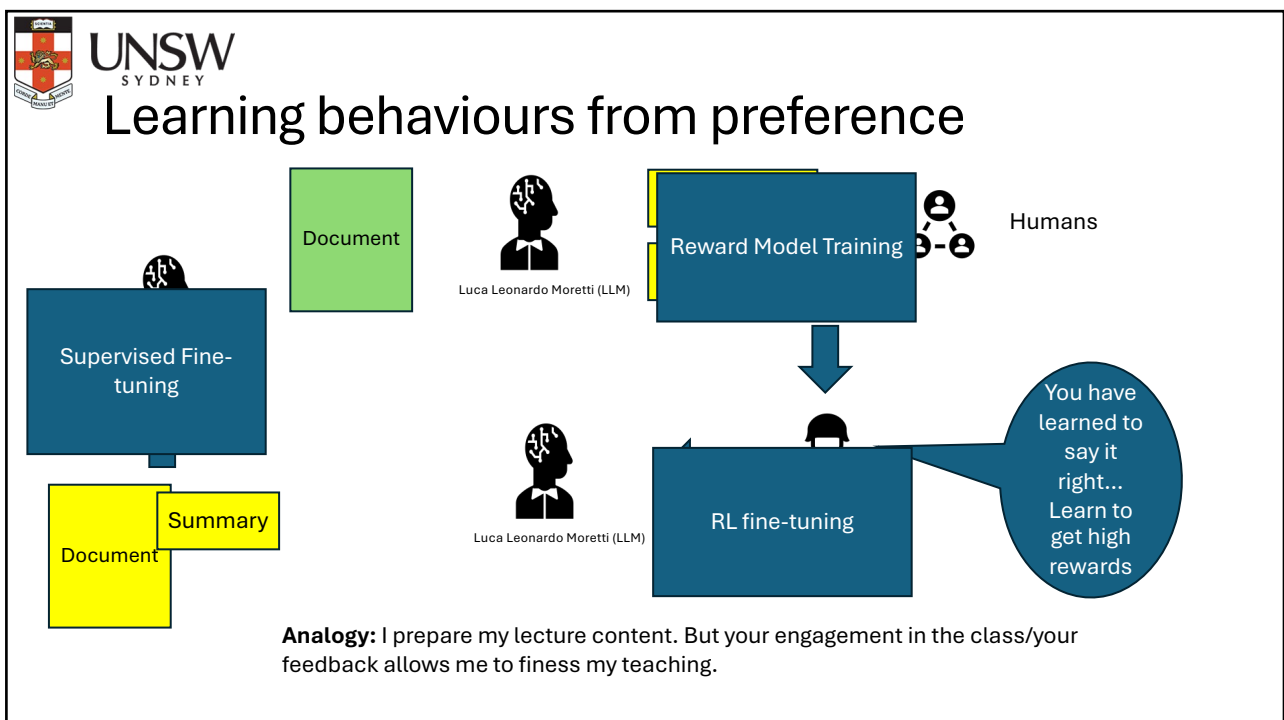
17



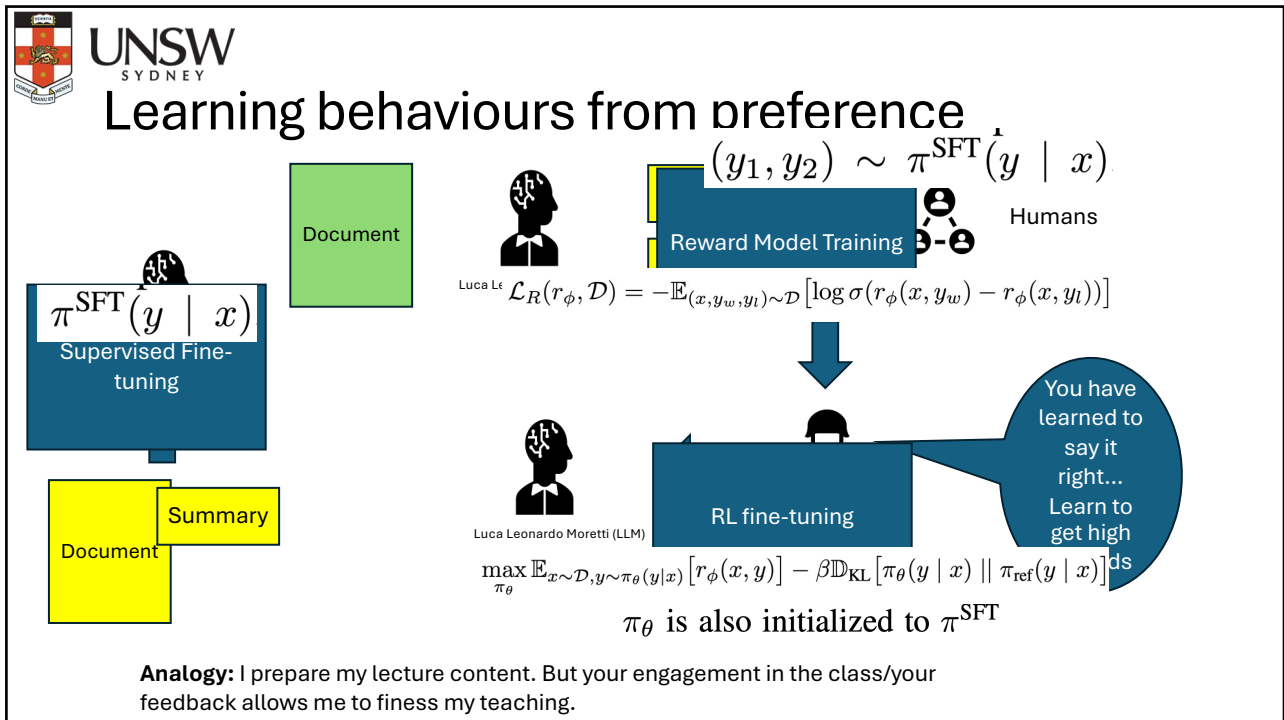
18



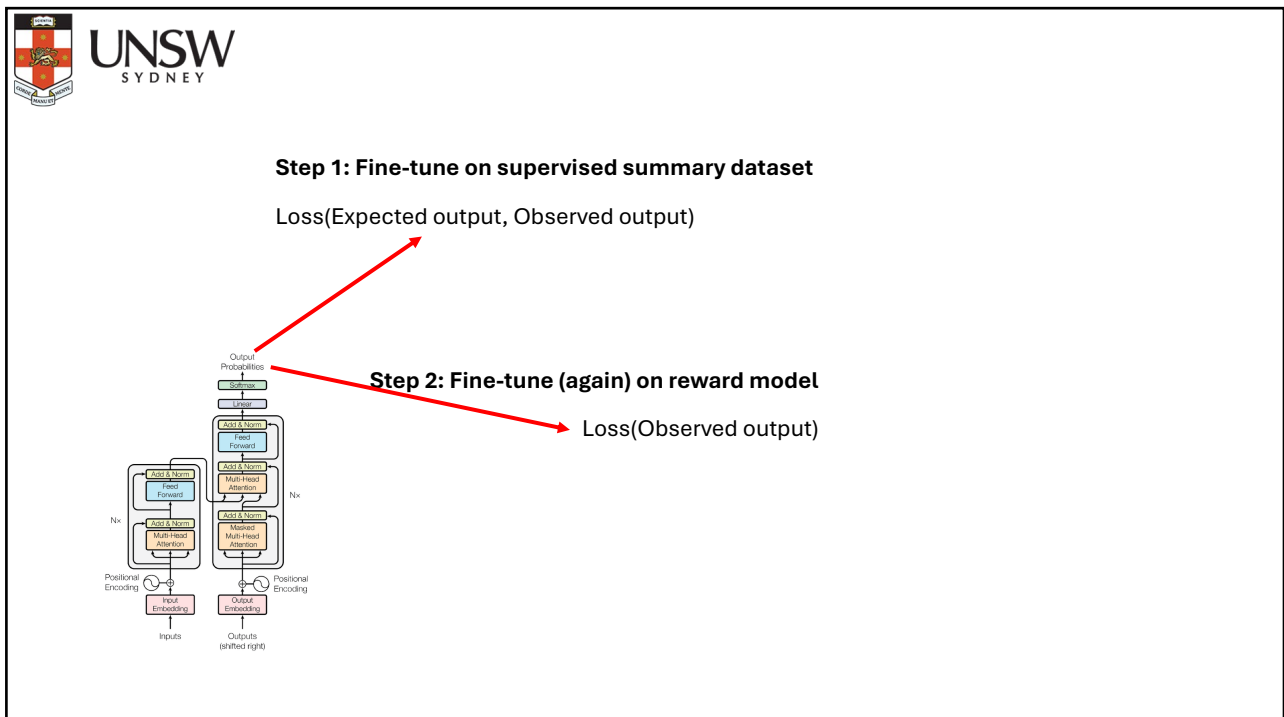
19



20



21



22



# UNSW SYDNEY

## Code Example

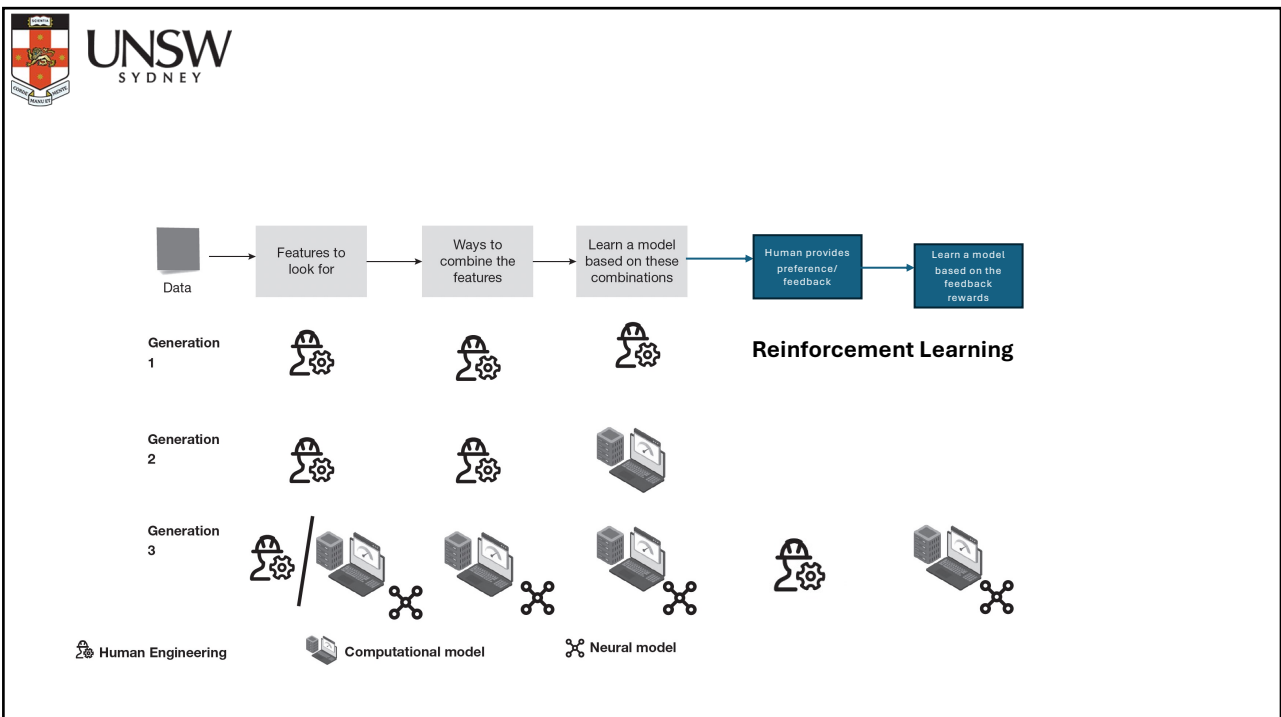
```
# 4. Initialize PPOTrainer
ppo_trainer = PPOTrainer(
    model=model,                # The pre-trained model (GPT-2 in this case)
    ref_model=model,            # Reference model (could be the same as model)
    tokenizer=tokenizer,        # Tokenizer for text input
    dataset=tokenized_dataset, # Dataset to fine-tune on
    config=ppo_config,         # PPO training configuration
    reward_fn=reward_fn        # The reward function used for training
)

# 5. Training loop
for batch in tokenized_dataset:
    # Tokenize prompts (input)
    input_ids = batch["input_ids"]
    # Generate responses (outputs) from the model
    outputs = model.generate(input_ids, max_length=100)
    # Decode generated responses and calculate reward
    responses = [tokenizer.decode(output, skip_special_tokens=True) for output in outputs]
    rewards = reward_fn(batch['text'], responses)
    # Perform PPO step
    ppo_trainer.step(batch["input_ids"], responses, rewards)
    print("Batch trained with PPO step")


# Save the fine-tuned model
model.save_pretrained("ppo_fine_tuned_gpt2")
tokenizer.save_pretrained("ppo_fine_tuned_gpt2")
```

<https://medium.com/@jimwang3589/what-is-rlhf-and-how-to-use-it-to-train-an-llm-part-4-1146228b74ef>



23




24



It's nearly the end of the course...  
time for us to compute our reward! 😊




**COMP6713 Natural  
Language Processing**


Student

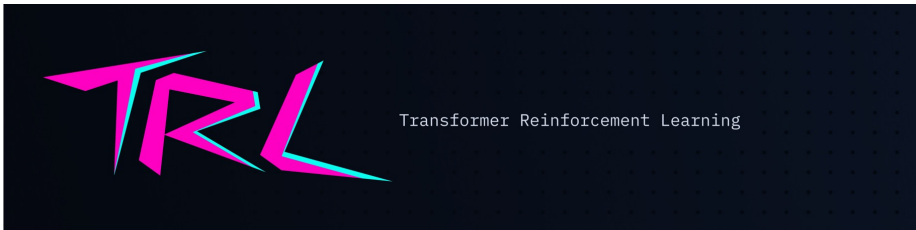
<https://go.blueja.io/Dlp4NkIfqUaTjH7kFAZAZQ>

---

 To access the evaluation, scan this QR code with your mobile phone.

25





**Library:** <https://huggingface.co/docs/trl/index>  
**Reddit Summarisation Dataset:** <https://huggingface.co/datasets/webis/tldr-17>

**Suggested Reading:**

Rafailov, Rafael, Archit Sharma, Eric Mitchell, Christopher D. Manning, Stefano Ermon, and Chelsea Finn. "Direct preference optimization: Your language model is secretly a reward model." *Advances in Neural Information Processing Systems* 36 (2023): 53728-53741.

Stiennon, Nisan, Long Ouyang, Jeffrey Wu, Daniel Ziegler, Ryan Lowe, Chelsea Voss, Alec Radford, Dario Amodei, and Paul F. Christiano. "Learning to summarize with human feedback." *Advances in neural information processing systems* 33 (2020): 3008-3021.

26