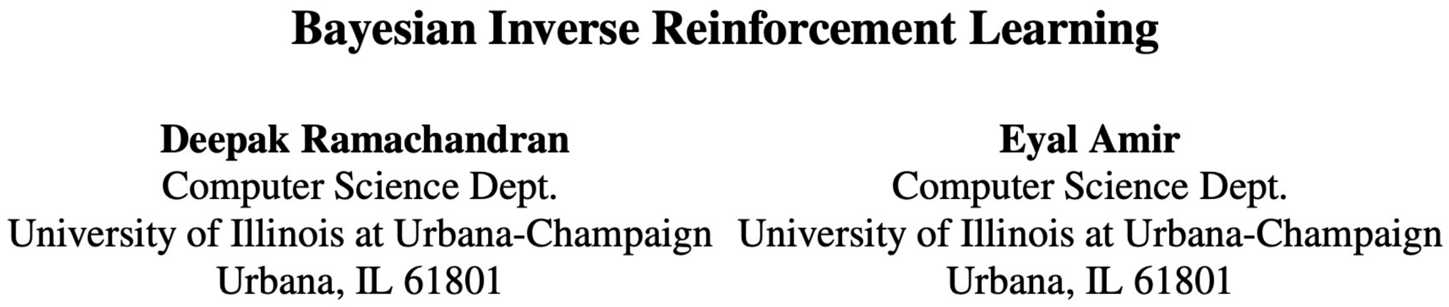
CSE 574 Planning and Learning Methods in AI

Ransalu Senanayake

Week 9

Uncertainty in rewards

- Inverse Reinforcement Learning (IRL)
- Bayesian Inverse Reinforcement Learning (BIRL)



Bayesian Inverse Reinforcement Learning

Deepak Ramachandran
Computer Science Dept.
University of Illinois at Urbana-Champaign
Urbana, IL 61801

Eyal Amir
Computer Science Dept.
University of Illinois at Urbana-Champaign
Urbana, IL 61801

Considerations in Human-in-the-Loop

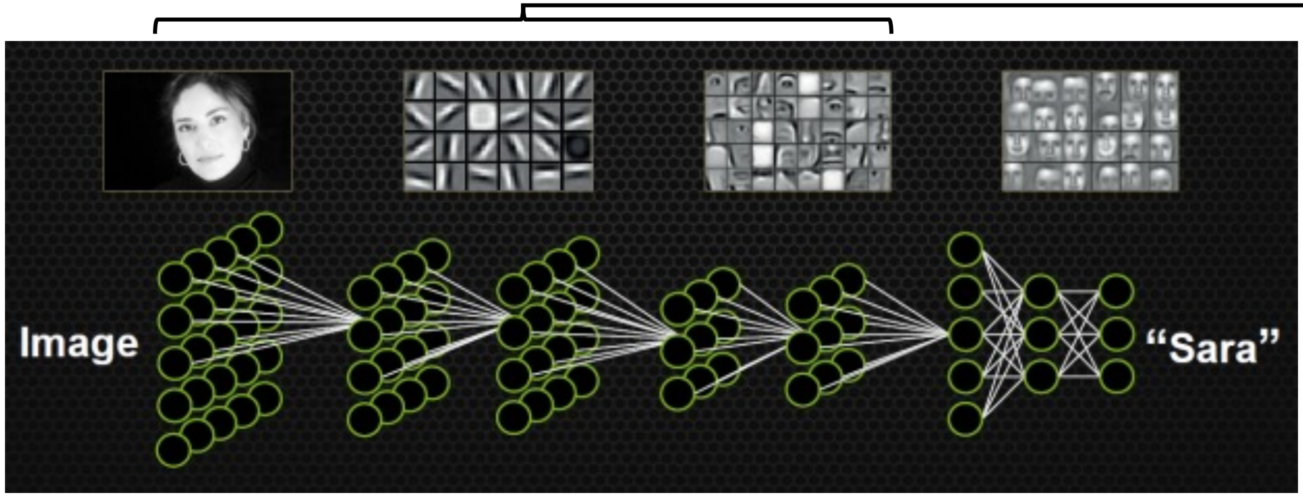
- Ways to get feedback
  - Use simpler feedback whenever possible (binary evaluations are easier than ranking or writing a paragraph but provide less information, making it longer to train)
  - Using the keyboard or touch screen can be easier than the mouse in most cases. Sometimes we have to drive a car/robot and show and that can raise safety concerns
  - Expert feedback can be expensive
- Effect of human bias and error
- Maintaining exploration and exploitation

Considerations in Human-in-the-Loop

- Self-supervised learning (SSL) for scalability
  1. Non-contrastive learning (e.g., by creating a new task)
     - Step 1: Create a new supervised learning task with labels based on the data you have so that the neural network learns basic features in the first few layers

       E.g. 1: Masking and predicting the word, next word prediction, etc. See BERT

       E.g. 2: Rotate the image and try to predict what the angle is
     - Step 2: Supervised fine-tune (the few last layers) with a few lables



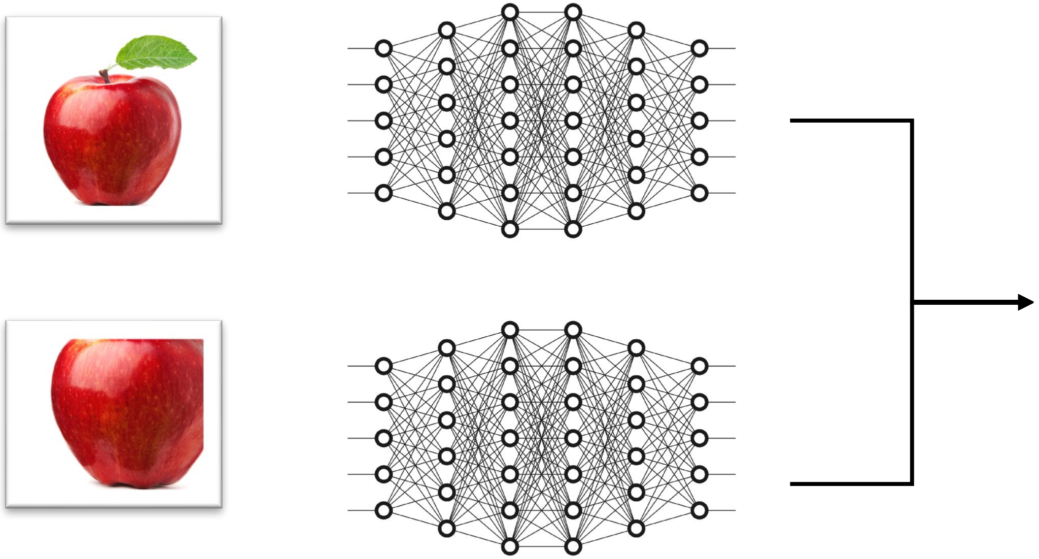
https://developer.nvidia.com/blog/accelerate-machine-learning-cudnn-deep-neural-network-library/

Considerations in Human-in-the-Loop

- Self-supervised learning (SSL) for scalability

  2. Contrastive learning

     E.g. Show positive and negative samples (e.g., full image and part of the same image would be a positive sample) and minimize the loss that maximizes the similarity between positive sample while minimizes the similarity between negative samples



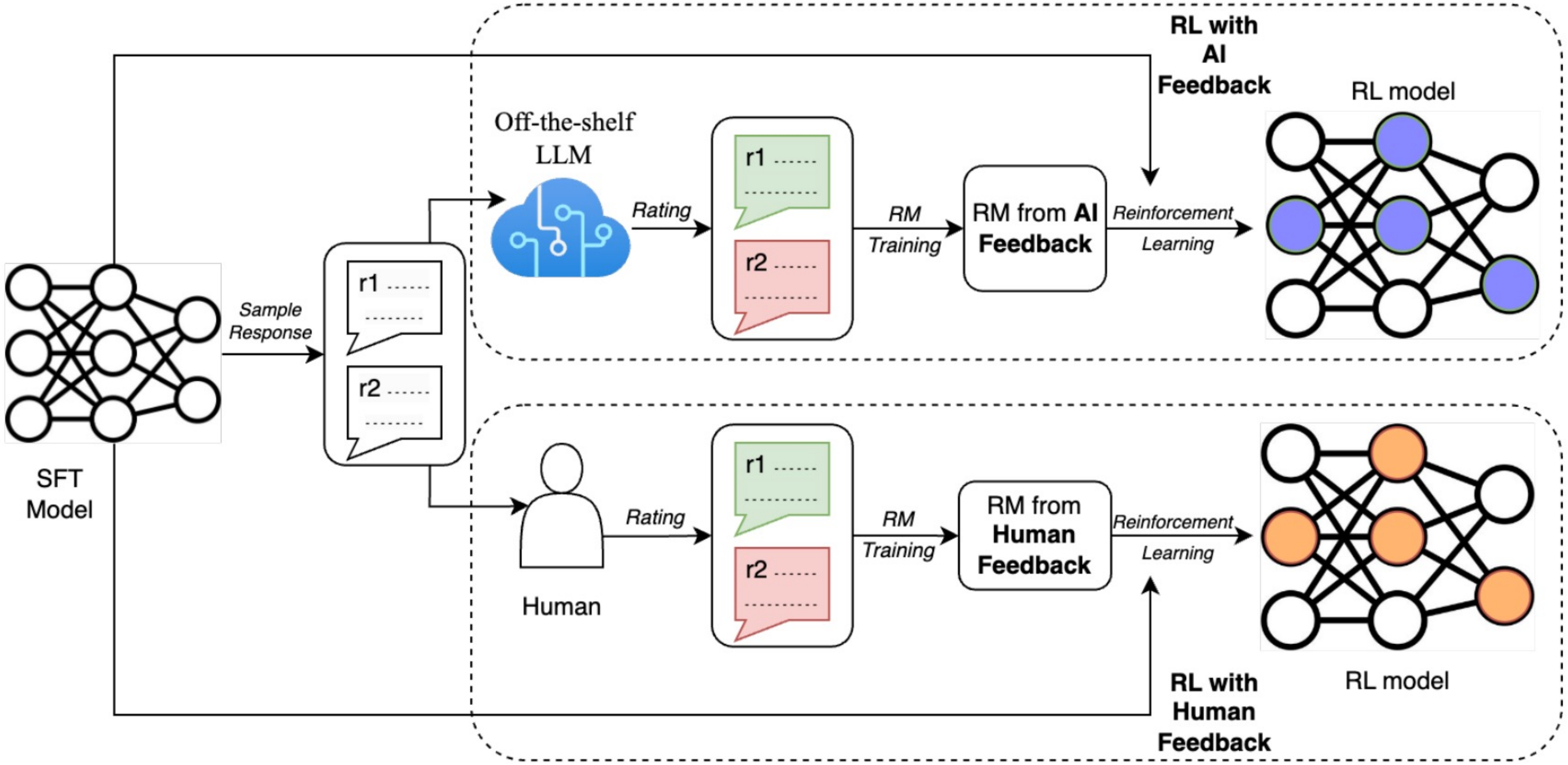
$$\min(\, yL_{sim} + (1-y)L_{diff})$$

Considerations in Human-in-the-Loop

- *Foundation models* for scalability
- Transfer learning/domain adaptation for scalability

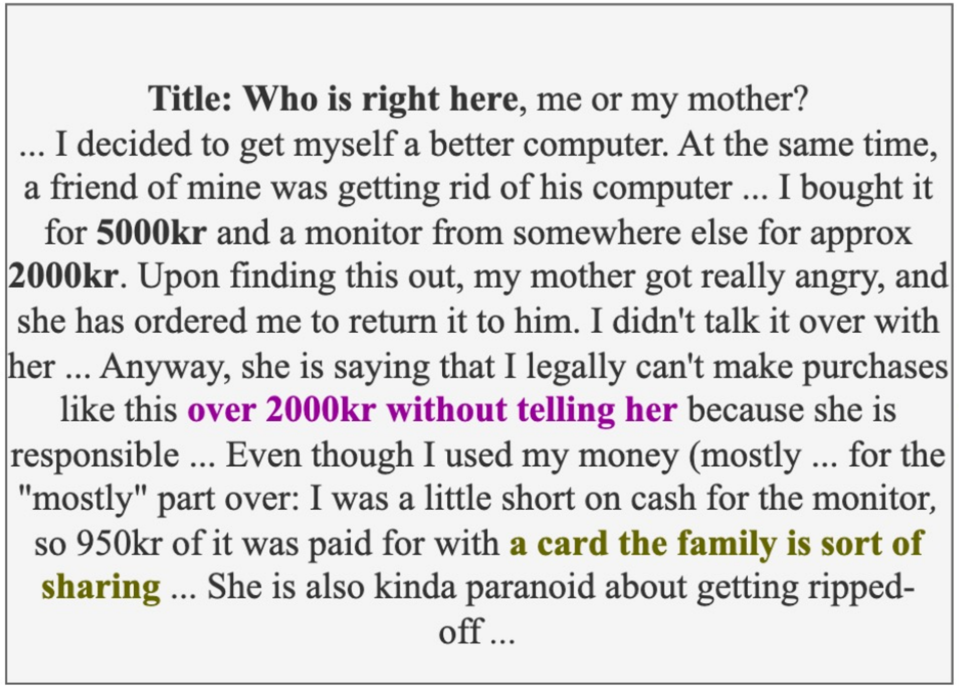
Reinforcement Learning with AI Feedback (RLAIF)

- Human evaluations are costly



Lee et al., RLAIF: Scaling Reinforcement Learning from Human Feedback with AI Feedback, 2023

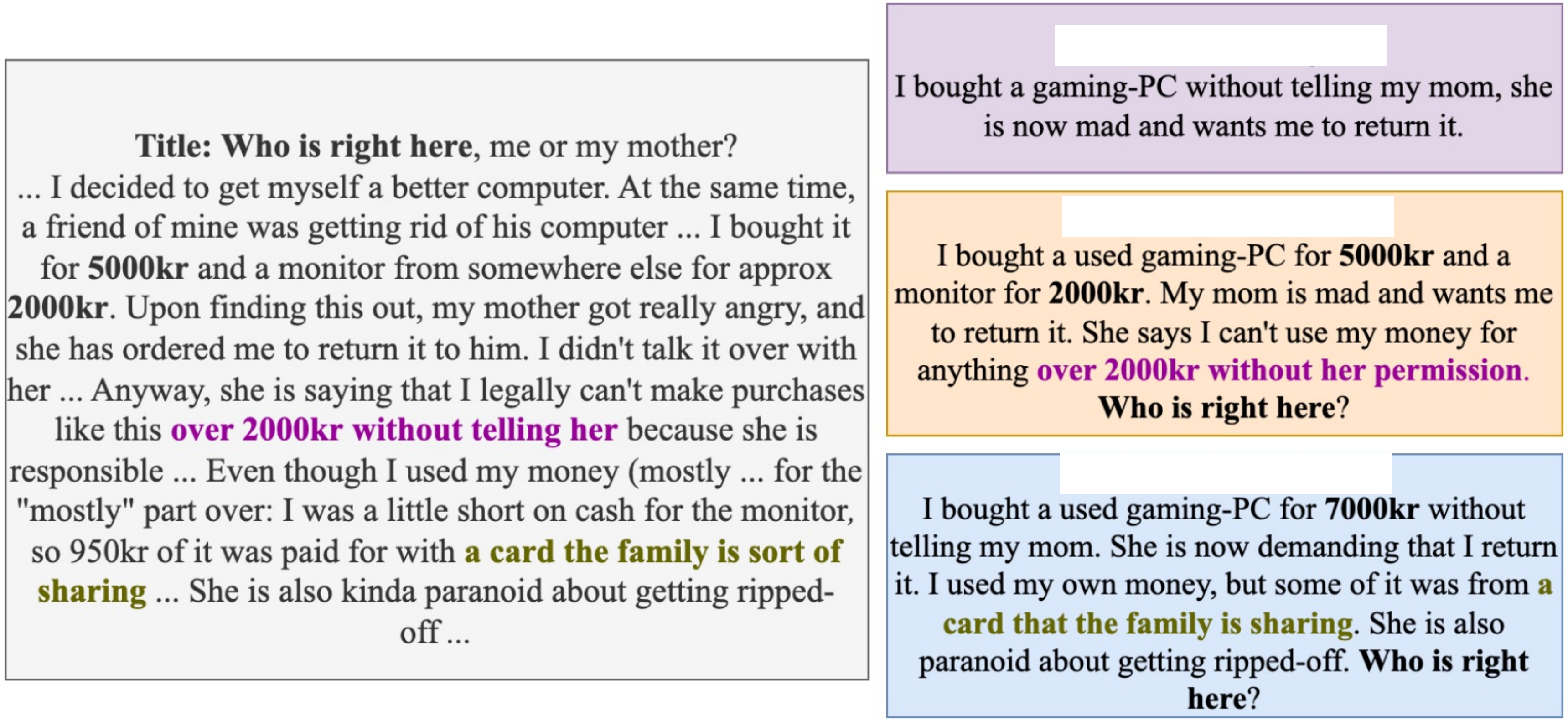
Which summaries do human prefer?



Title: Who is right here, me or my mother?
... I decided to get myself a better computer. At the same time, a friend of mine was getting rid of his computer ... I bought it for **5000kr** and a monitor from somewhere else for approx **2000kr**. Upon finding this out, my mother got really angry, and she has ordered me to return it to him. I didn't talk it over with her ... Anyway, she is saying that I legally can't make purchases like this **over 2000kr without telling her** because she is responsible ... Even though I used my money (mostly ... for the "mostly" part over: I was a little short on cash for the monitor, so 950kr of it was paid for with **a card the family is sort of sharing** ... She is also kinda paranoid about getting ripped-off ...

Lee et al., RLAIF: Scaling Reinforcement Learning from Human Feedback with AI Feedback, 2023

Which summaries do human prefer?



Title: Who is right here, me or my mother?
... I decided to get myself a better computer. At the same time, a friend of mine was getting rid of his computer ... I bought it for **5000kr** and a monitor from somewhere else for approx **2000kr**. Upon finding this out, my mother got really angry, and she has ordered me to return it to him. I didn't talk it over with her ... Anyway, she is saying that I legally can't make purchases like this **over 2000kr without telling her** because she is responsible ... Even though I used my money (mostly ... for the "mostly" part over: I was a little short on cash for the monitor, so 950kr of it was paid for with **a card the family is sort of sharing** ... She is also kinda paranoid about getting ripped-off ...

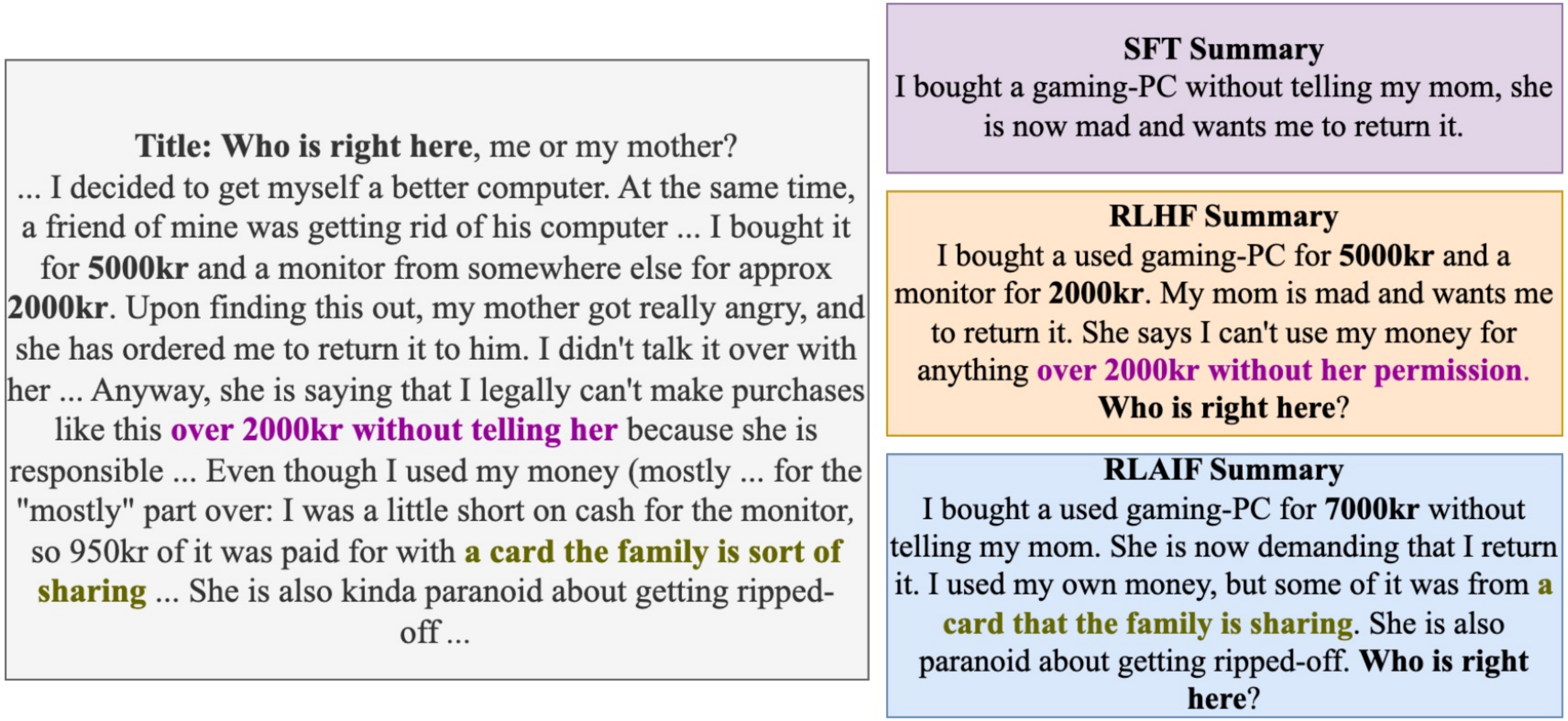
I bought a gaming-PC without telling my mom, she is now mad and wants me to return it.

I bought a used gaming-PC for **5000kr** and a monitor for **2000kr**. My mom is mad and wants me to return it. She says I can't use my money for anything **over 2000kr without her permission**. **Who is right here**?

I bought a used gaming-PC for **7000kr** without telling my mom. She is now demanding that I return it. I used my own money, but some of it was from **a card that the family is sharing**. She is also paranoid about getting ripped-off. **Who is right here**?

Lee et al., RLAIF: Scaling Reinforcement Learning from Human Feedback with AI Feedback, 2023

Which summaries do human prefer?

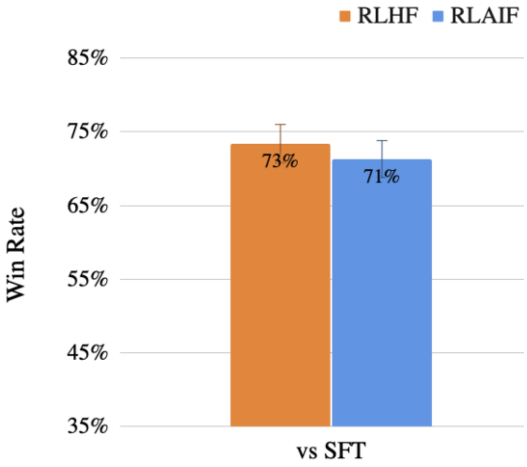


Title: Who is right here, me or my mother?
... I decided to get myself a better computer. At the same time, a friend of mine was getting rid of his computer ... I bought it for **5000kr** and a monitor from somewhere else for approx **2000kr**. Upon finding this out, my mother got really angry, and she has ordered me to return it to him. I didn't talk it over with her ... Anyway, she is saying that I legally can't make purchases like this **over 2000kr without telling her** because she is responsible ... Even though I used my money (mostly ... for the "mostly" part over: I was a little short on cash for the monitor, so 950kr of it was paid for with **a card the family is sort of sharing** ... She is also kinda paranoid about getting ripped-off ...

SFT Summary
I bought a gaming-PC without telling my mom, she is now mad and wants me to return it.

RLHF Summary
I bought a used gaming-PC for **5000kr** and a monitor for **2000kr**. My mom is mad and wants me to return it. She says I can't use my money for anything **over 2000kr without her permission**. **Who is right here**?

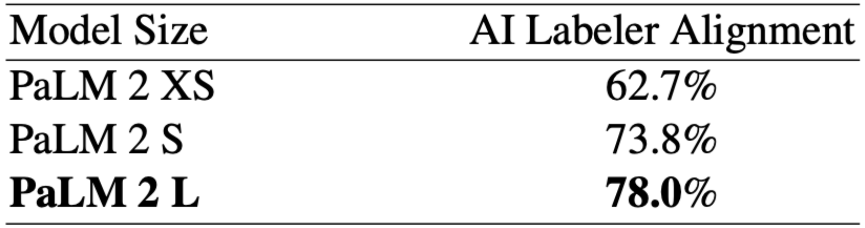
RLAIF Summary
I bought a used gaming-PC for **7000kr** without telling my mom. She is now demanding that I return it. I used my own money, but some of it was from **a card that the family is sharing**. She is also paranoid about getting ripped-off. **Who is right here**?



Lee et al., RLAIF: Scaling Reinforcement Learning from Human Feedback with AI Feedback, 2023

Factors that Affect

Size of the LLM



Model Size	AI Labeler Alignment
PaLM 2 XS	62.7%
PaLM 2 S	73.8%
PaLM 2 L	**78.0%**

Table 4: AI Labeler Alignment increases as the size of the LLM labeler increases.

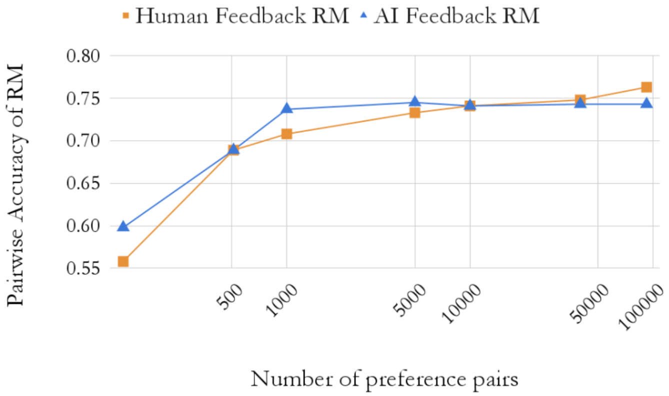
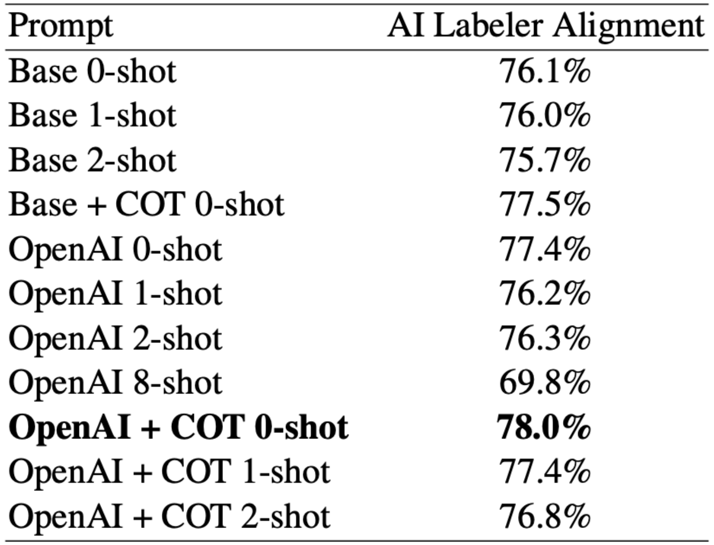


Figure 5: RM accuracy on a held-out set of human preferences increases rapidly as more preference pairs are used in training. After training on a few thousand examples, performance is close to training on the full dataset. The x-axis is in log-scale.

How many prompts?

Quality of prompts



Prompt	AI Labeler Alignment
Base 0-shot	76.1%
Base 1-shot	76.0%
Base 2-shot	75.7%
Base + COT 0-shot	77.5%
OpenAI 0-shot	77.4%
OpenAI 1-shot	76.2%
OpenAI 2-shot	76.3%
OpenAI 8-shot	69.8%
OpenAI + COT 0-shot	**78.0%**
OpenAI + COT 1-shot	77.4%
OpenAI + COT 2-shot	76.8%

Table 2: We observe that prompting with the detailed OpenAI preamble and eliciting chain-of-thought reasoning gives the highest AI Labeler Alignment. In-context learning does not improve accuracy, and possibly even makes it worse.

Lee et al., RLAIF: Scaling Reinforcement Learning from Human Feedback with AI Feedback, 2023