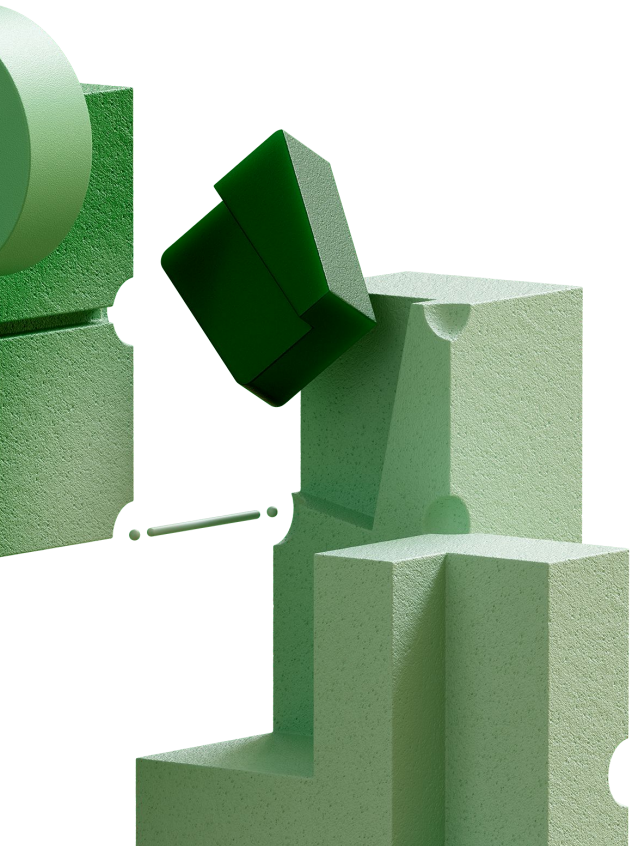
Google DeepMind



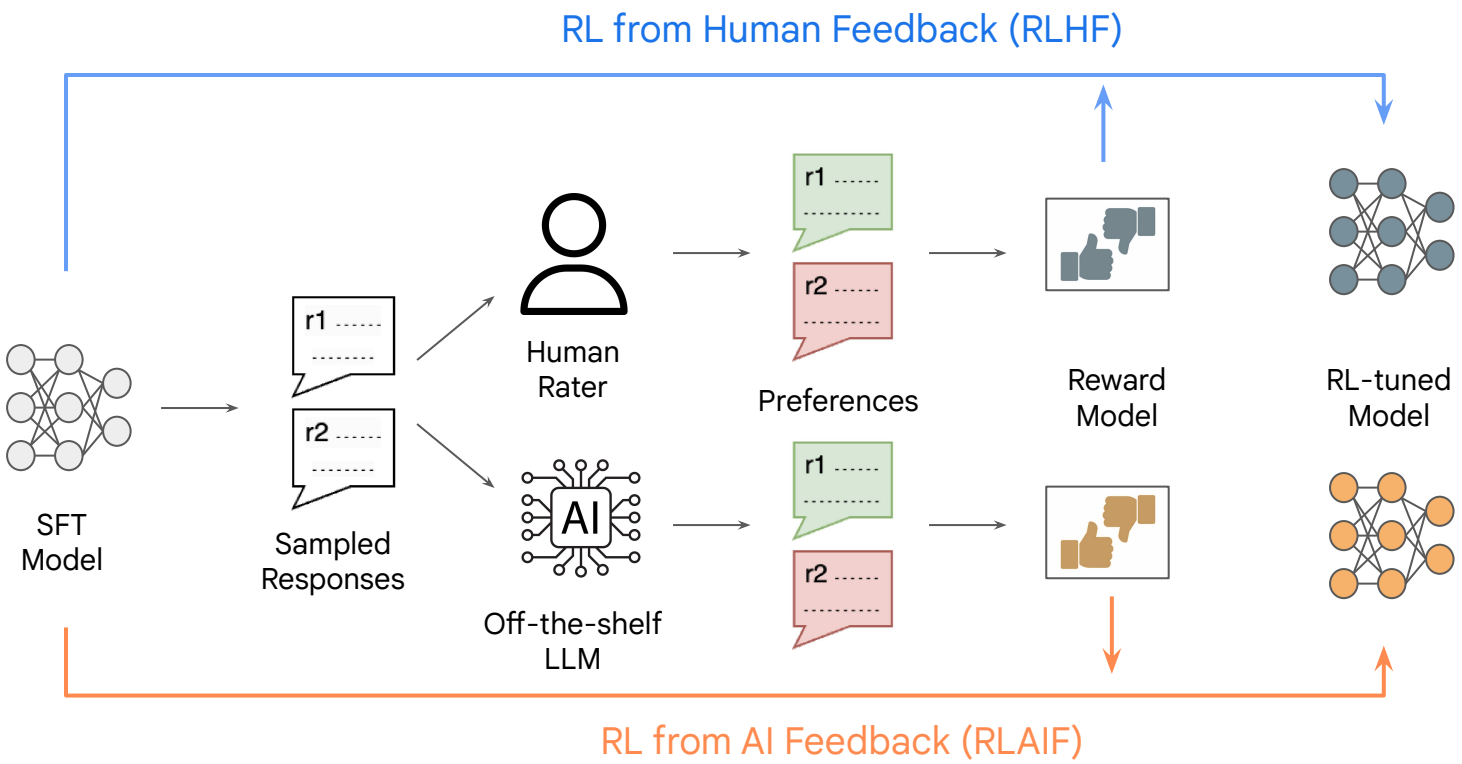
RLAIF vs. RLHF: Scaling Reinforcement Learning from Human Feedback with AI Feedback

Harrison Lee, Samrat Phatale, Hassan Mansoor, Thomas Mesnard, Johan Ferret, Kellie Lu, Colton Bishop, Ethan Hall, Victor Cărbune, Abhinav Rastogi, Sushant Prakash

ICML 2024

Contents

Overview



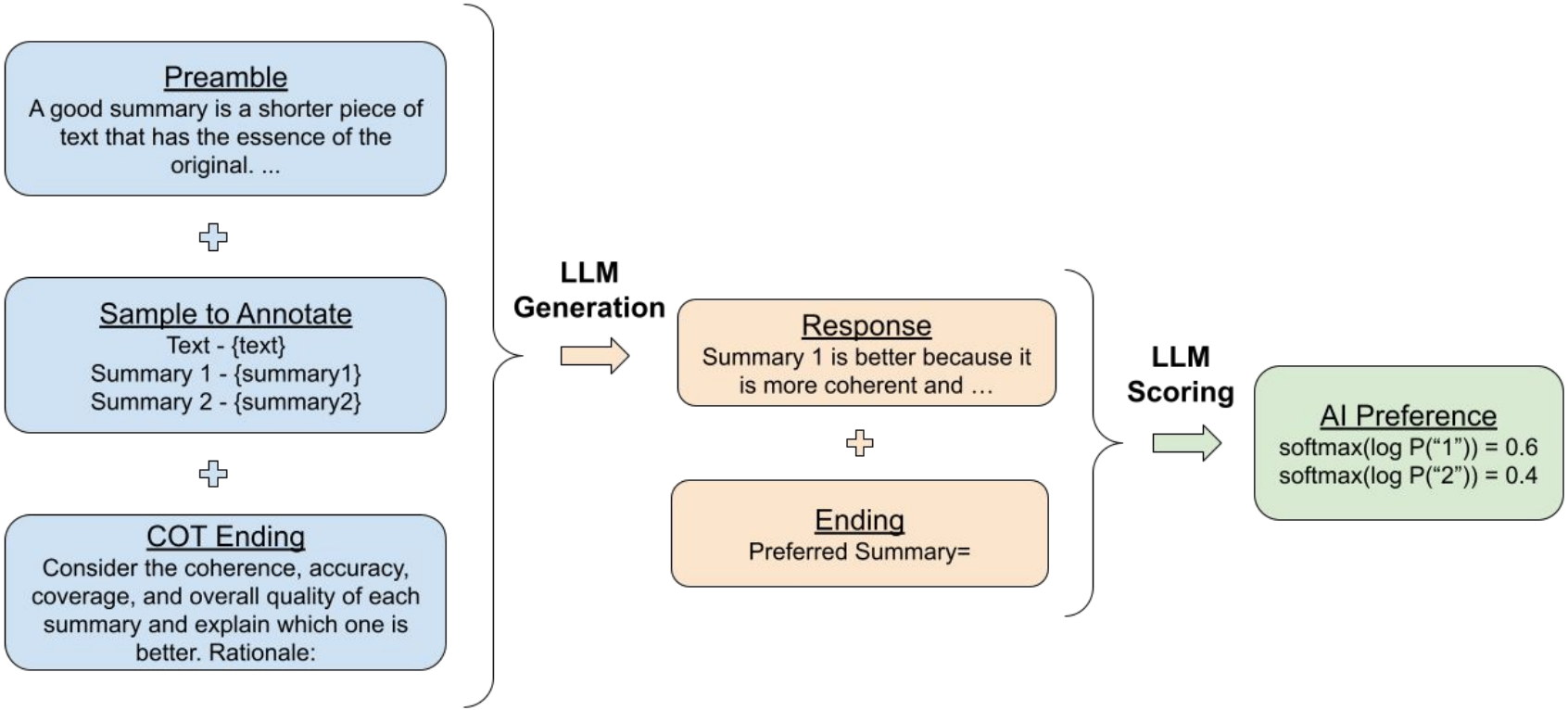
Research Questions

RLAIF first introduced in Bai et al. 2022. This work seeks to answer...

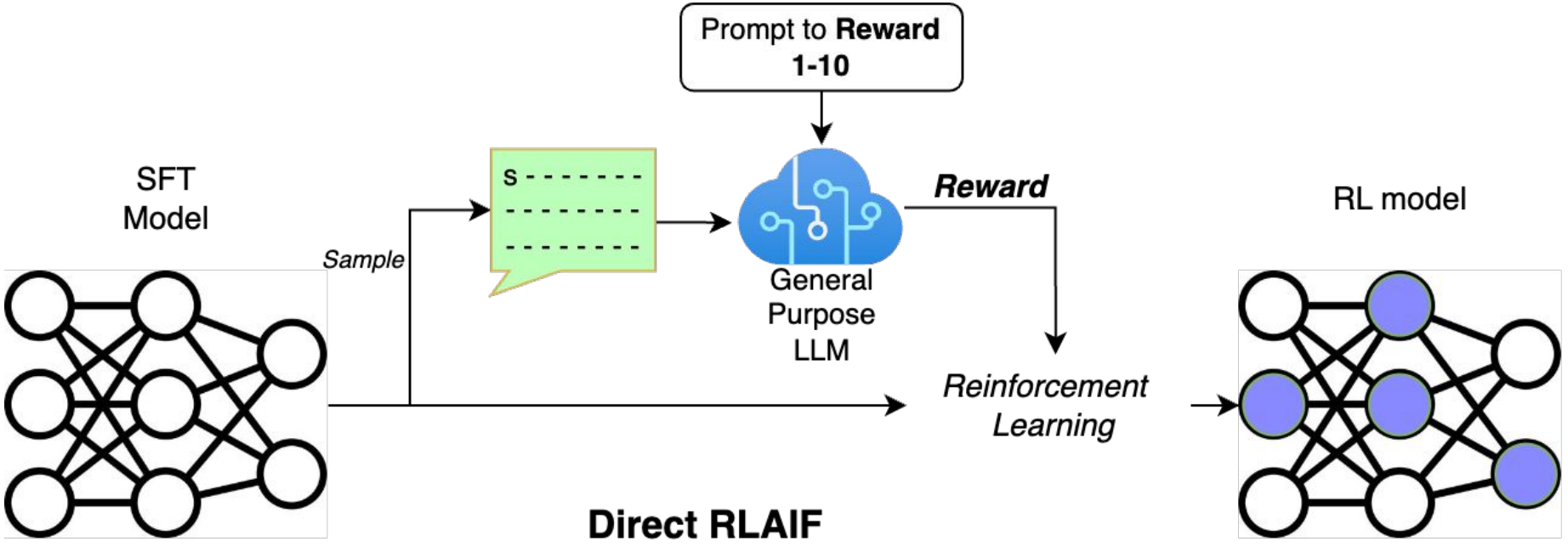
- **“Can RLAIF replace RLHF?”**
- “Can RLAIF be used for self-improvement?”
- “Can we leverage LLMs to directly produce a reward signal during RL?”
- ...and more

Contents

AI Preference Generation



Direct RLAIF



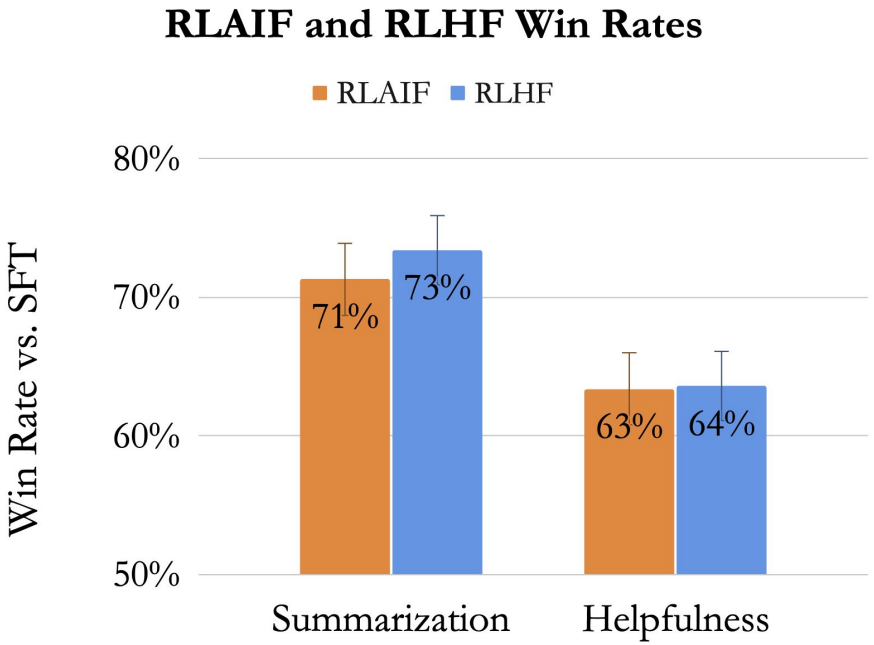
Experimental Details

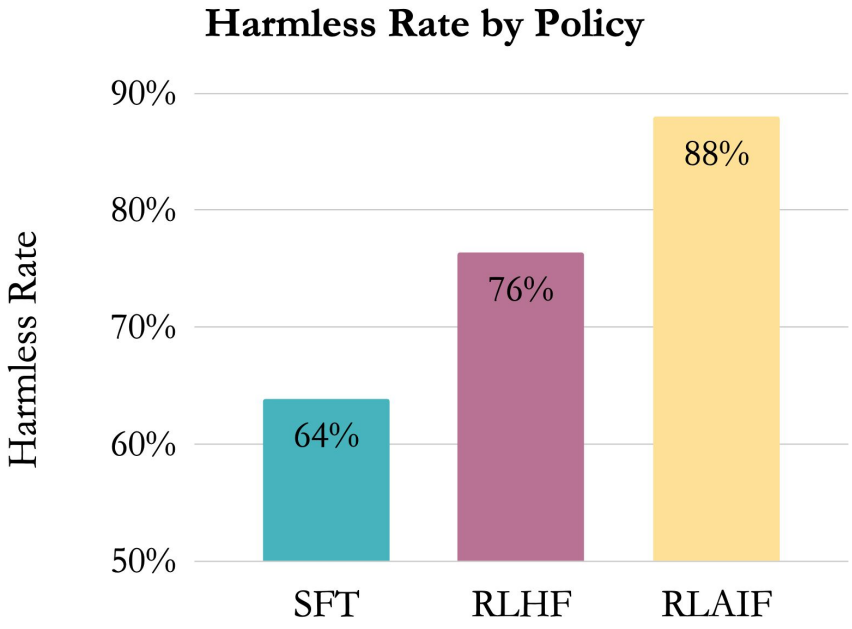
- AI Labeler
  - PaLM 2 Large
  - or PaLM 2 XS for self-improvement
- Policy Model - PaLM 2 XS
- Reward Model - PaLM 2 XS
- Algorithm - REINFORCE with value function

Contents

RLAIF vs. RLHF

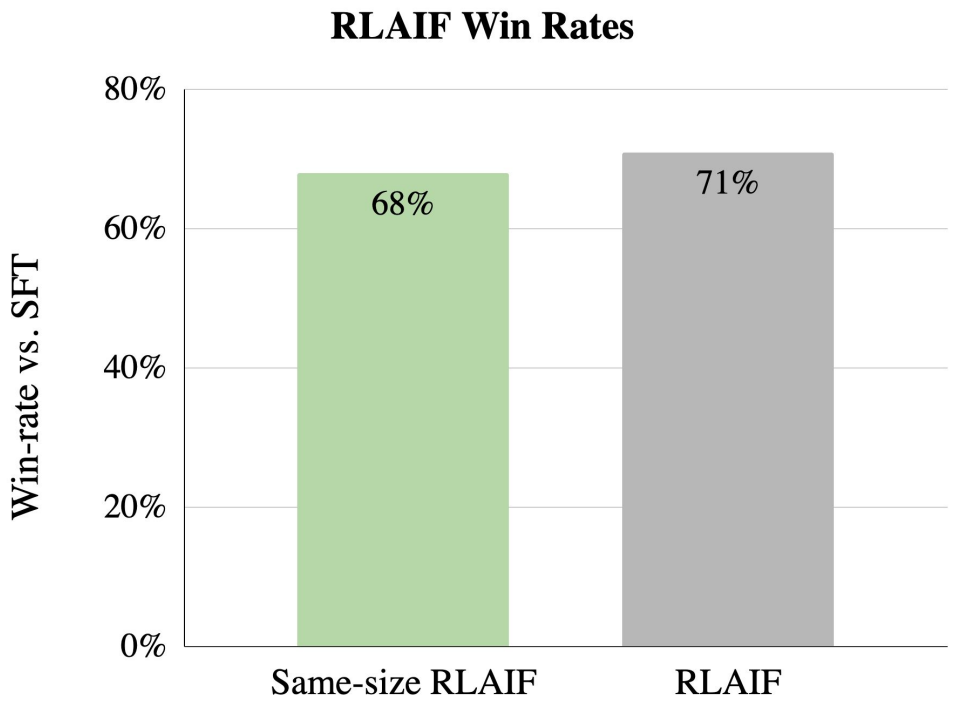
RLAIF performs **on par** with RLHF on summarization, helpfulness, and harmlessness





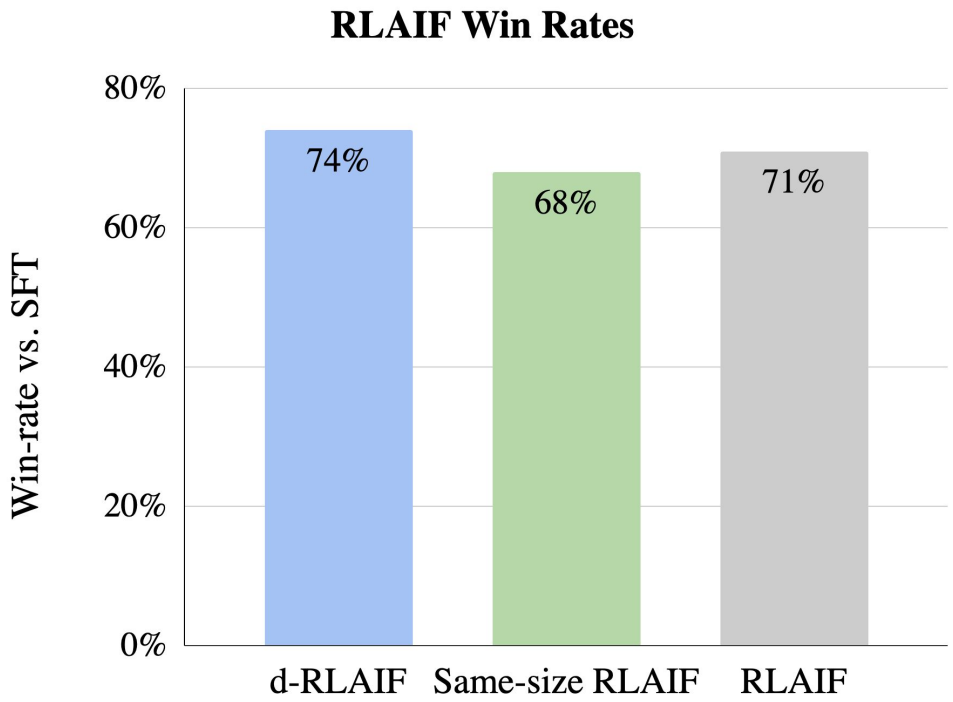
Towards Self-Improvement

RLAIF improves over SFT even when AI labeler and policy are **the same size**



Direct RLAIF (d-RLAIF)

Directly prompting the LLM for rewards performs **even better**



** d-RLAIF was also used on the helpfulness task, where it achieved a 66% win rate over SFT (compared to 63% from canonical RLAIF)*

Thank you.

For more, see
https://arxiv.org/pdf/2309.00267.pdf

Appendix

Example Prompt (summarization)

Preamble	A good summary is a shorter piece of text that has the essence of the original. ... Given a piece of text and two of its possible summaries, output 1 or 2 to indicate which summary best adheres to coherence, accuracy, coverage, and overall quality as defined above.
1-Shot Exemplar	»»»»» Example »»»»» Text - We were best friends over 4 years ... Summary 1 - Broke up with best friend, should I wish her a happy birthday... And what do you think of no contact? Summary 2 - should I wish my ex happy birthday, I broke no contact, I'm trying to be more patient, I'm too needy, and I don't want her to think I'll keep being that guy. Preferred Summary=1 »»»»» Follow the instructions and the example(s) above »»»»»
Sample to Annotate	Text - **{text}** Summary 1 - **{summary1}** Summary 2 - **{summary2}**
Ending	Preferred Summary=

Prompting Techniques

Key Findings

- Chain-of-Thought consistently improves alignment
- Few-shot prompting doesn't always improve the alignment.

	AI Labeler Alignment		
Prompt	Summary	H1	H2
Base 0-shot	76.1%	67.8%	69.4%
Base 1-shot	76.0%	67.1%	71.7%
Base 2-shot	75.7%	66.8%	**72.1%**
Base + CoT 0-shot	77.5%	**69.1%**	70.6%
Detailed 0-shot	77.4%	67.6%	70.1%
Detailed 1-shot	76.2%	67.6%	71.5%
Detailed 2-shot	76.3%	67.3%	71.6%
Detailed 8-shot	69.8%	–	–
Detailed + CoT 0-shot	**78.0%**	67.8%	70.1%
Detailed + CoT 1-shot	77.4%	67.4%	69.9%
Detailed + CoT 2-shot	76.8%	67.4%	69.2%

** Note: **H1** refers to Helpful dialogue while **H2** refers to Harmless dialogue*

Size of AI Labeler

Alignment with human preferences **improves with AI labeler size**

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

Example of SFT vs. RLHF vs. RLAIF summaries

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**?

Cost Analysis

We estimate AI preference labeling to be **10x cheaper** than human preference labeling