Jiaying Liang

DS5690 Transformers Fall 2023

RLAIF – Overview

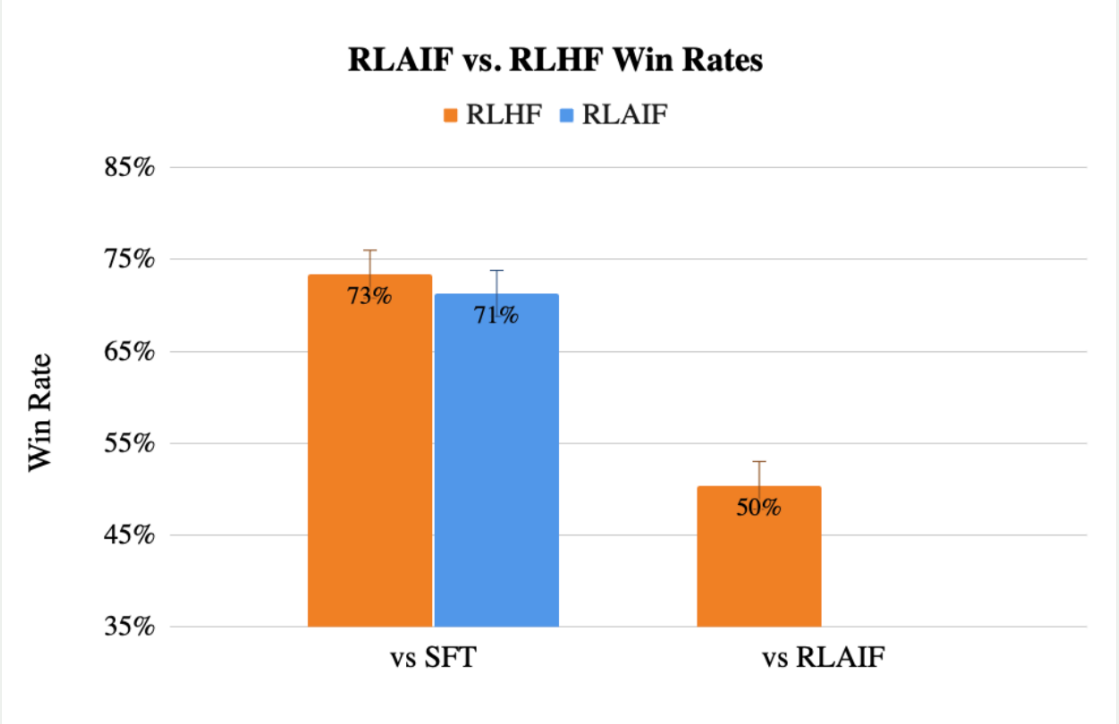
RLAIF: Scaling Reinforcement Learning from Human Feedback with AI Feedback

Harrison Lee, Samrat Phatale, Hassan Mansoor, Kellie Lu, Thomas Mesnard, Colton Bishop, Victor Carbune, Abhinav Rastogi
Google Research
{harrisonlee,samratph,hassan}@google.com

Source: https://arxiv.org/abs/2309.00267

Main Result 1:

- Figure 1: Human evaluators strongly prefer RLHF and RLAIF summaries over the supervised fine-tuned (SFT) baseline. The differences in win rates between *RLAIF vs. SFT* and *RLHF vs. SFT* are not statistically significant. Additionally, when compared head-to-head, RLAIF is equally preferred to RLHF by human evaluators. Error bars denote 95% confidence intervals.



Main Result 2:

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%

Chain-of-thought: COT

Self-Consistency	AI Labeler Alignment
1 sample, T=0	**78.0%**
4 samples, T=1	72.6%
16 samples, T=1	72.8%

Self-Consistency

Main Result 3:

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.

Thank you!

Contact info: jiaying.liang@vanderbilt.edu

References

- Lee, H., Phatale, S., Mansoor, H., Lu, K., Mesnard, T., Bishop, C., ... & Rastogi, A. (2023). Rlaif: Scaling reinforcement learning from human feedback with ai feedback. arXiv preprint arXiv:2309.00267.