university of koblenz
continue discovering

IWVI

Combined Preference-Based & Absolute Reward Signals for RLHF Fine-tuning

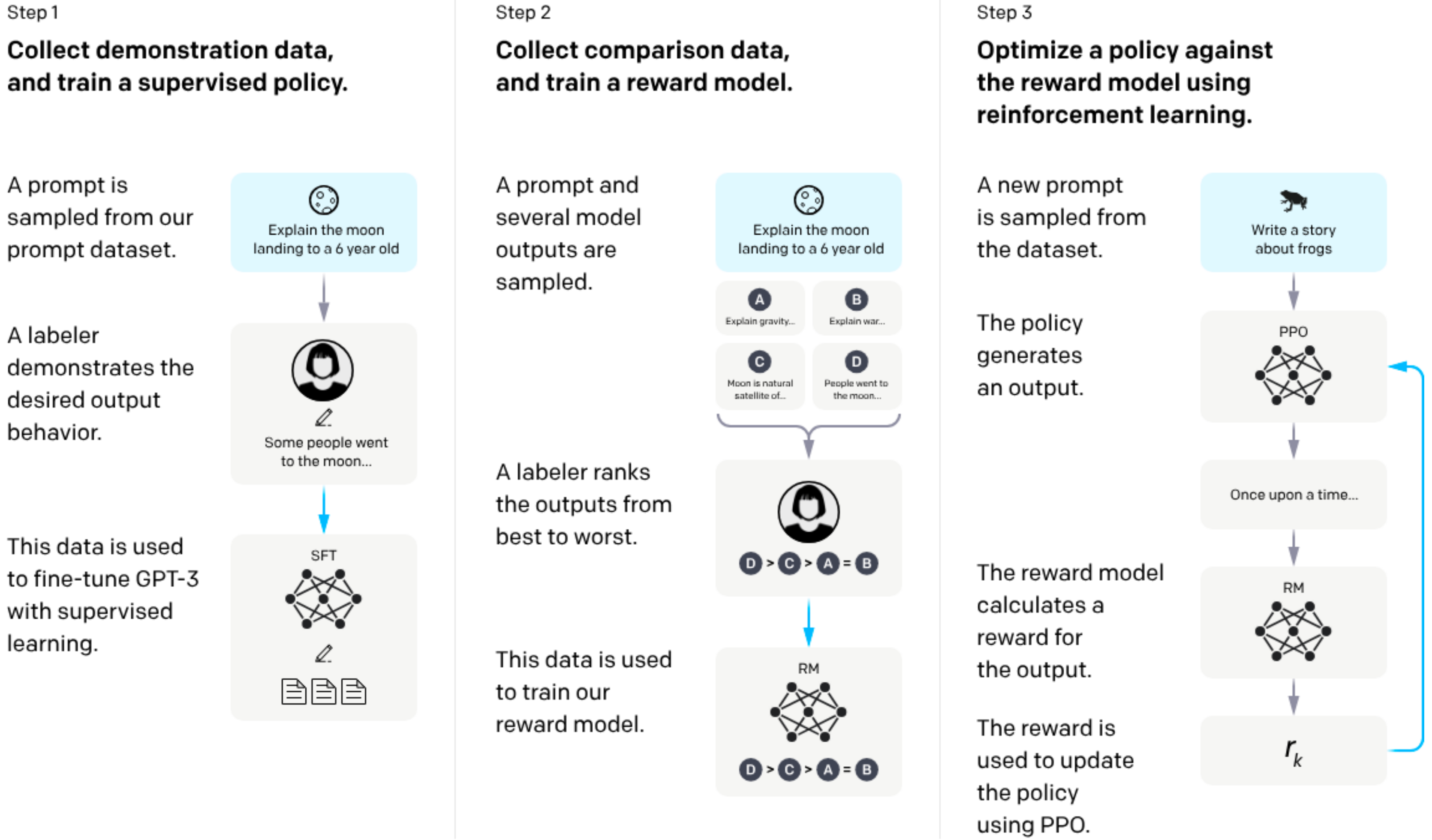
Master Thesis

AmeerAli Khan | Web & Data Science

Table of Content

Background

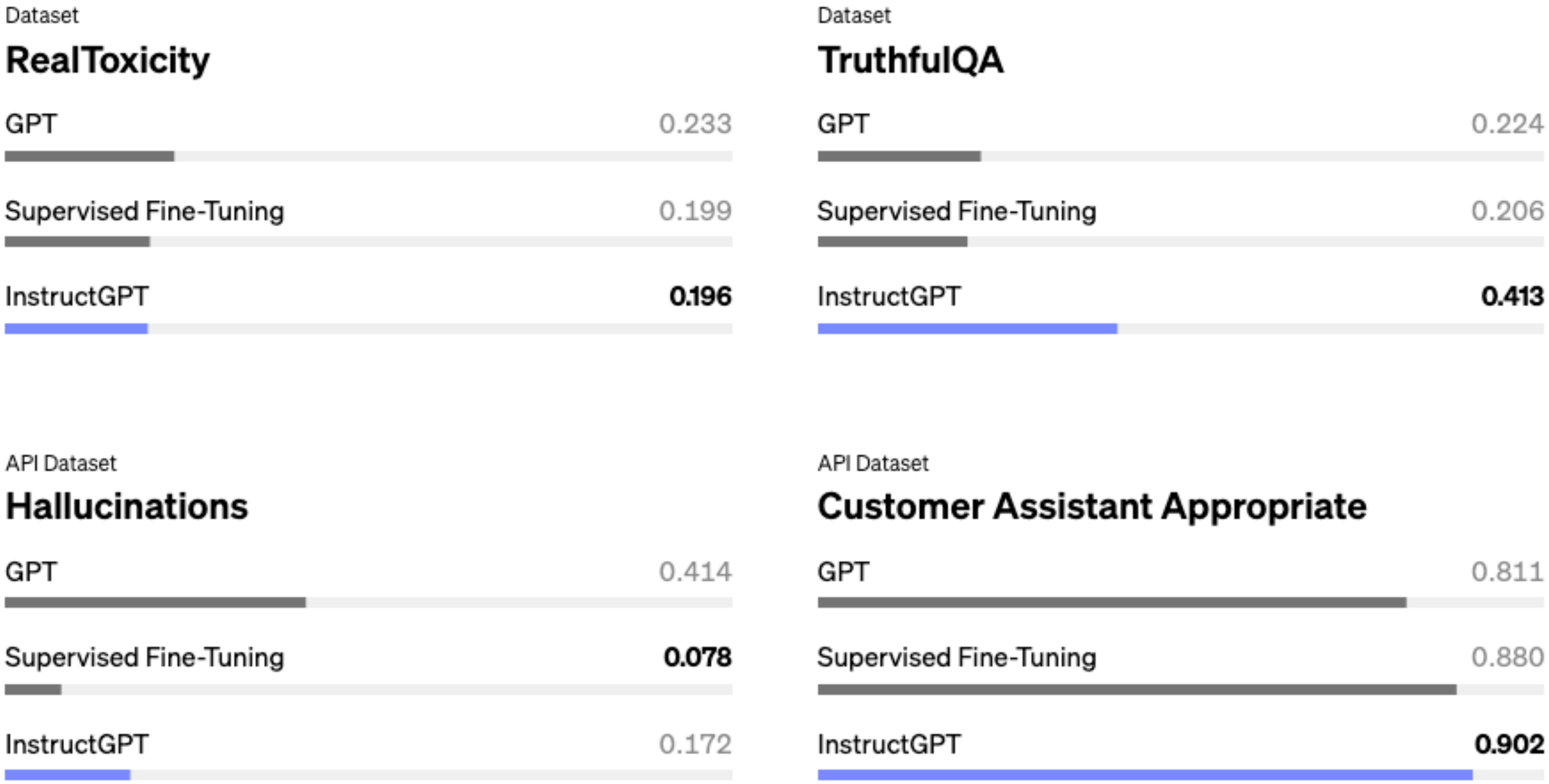
Reinforcement Learning from Human Feedback (RLHF)



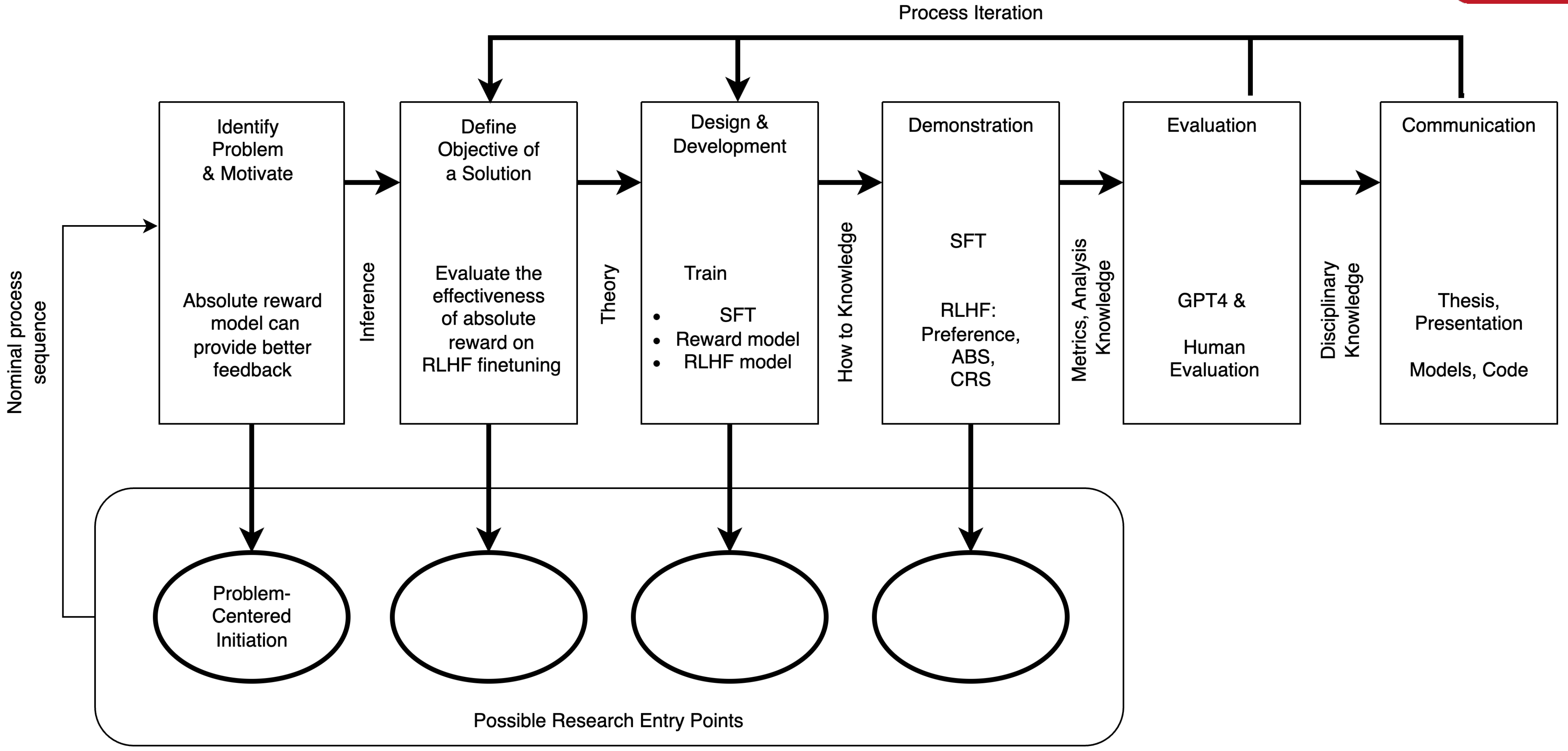
(Ouyang et al., 2022), Ziegler et al. (2020)

Background

Result of RLHF Fine-tuning



Design Science Research Methodology



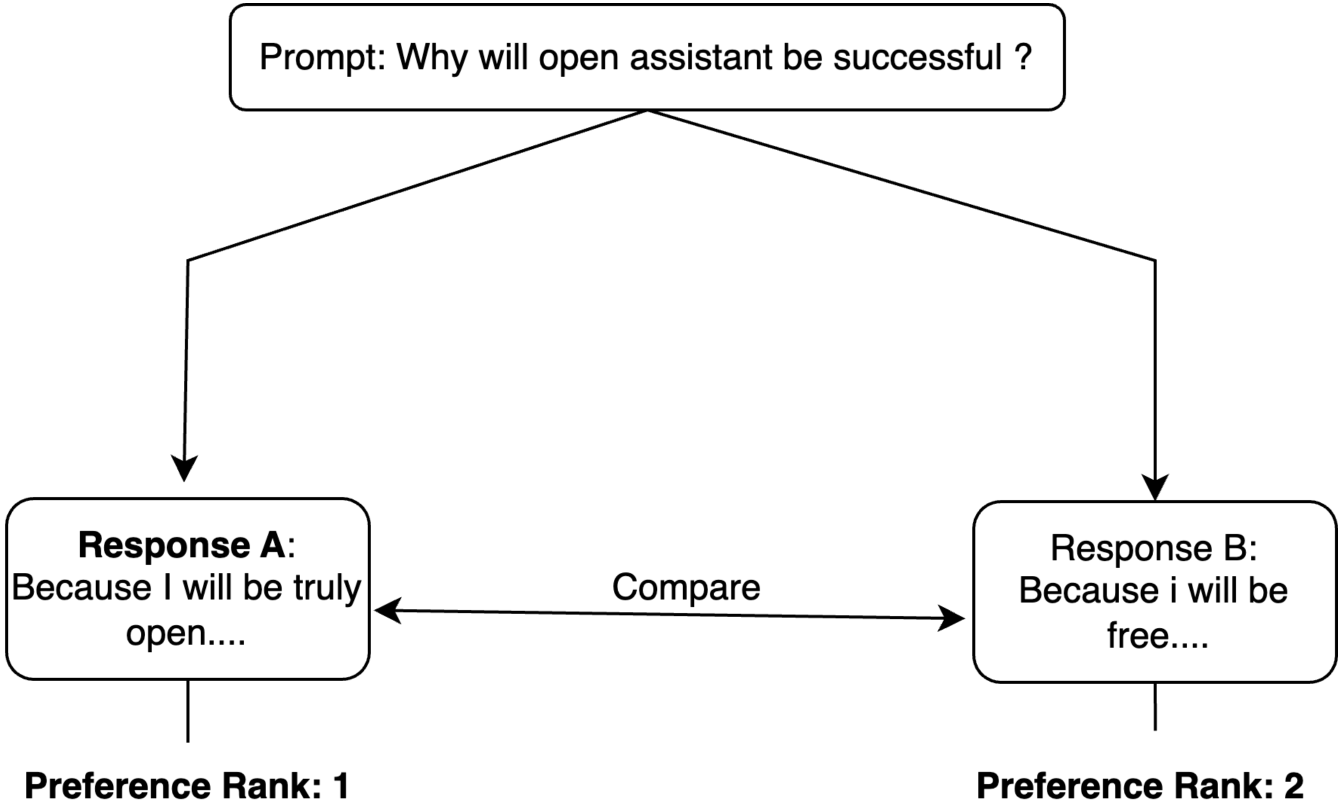
(Ouyang et al., 2022)

Research Method - Motivation

- Preference rank dataset
- Loss function by Ouyang et al. 2022

$$\operatorname{loss}(\theta) = -\frac{1}{\binom{K}{2}} \mathbb{E}_{(x, y_w, y_l) \sim \mathcal{D}} \left[\log \left(\sigma \left(r_\theta(x, y_w) - r_\theta(x, y_l) \right) \right) \right]$$

Sample from reward modelling dataset (Köpf et al. 2023)



(Ouyang et al., 2022)

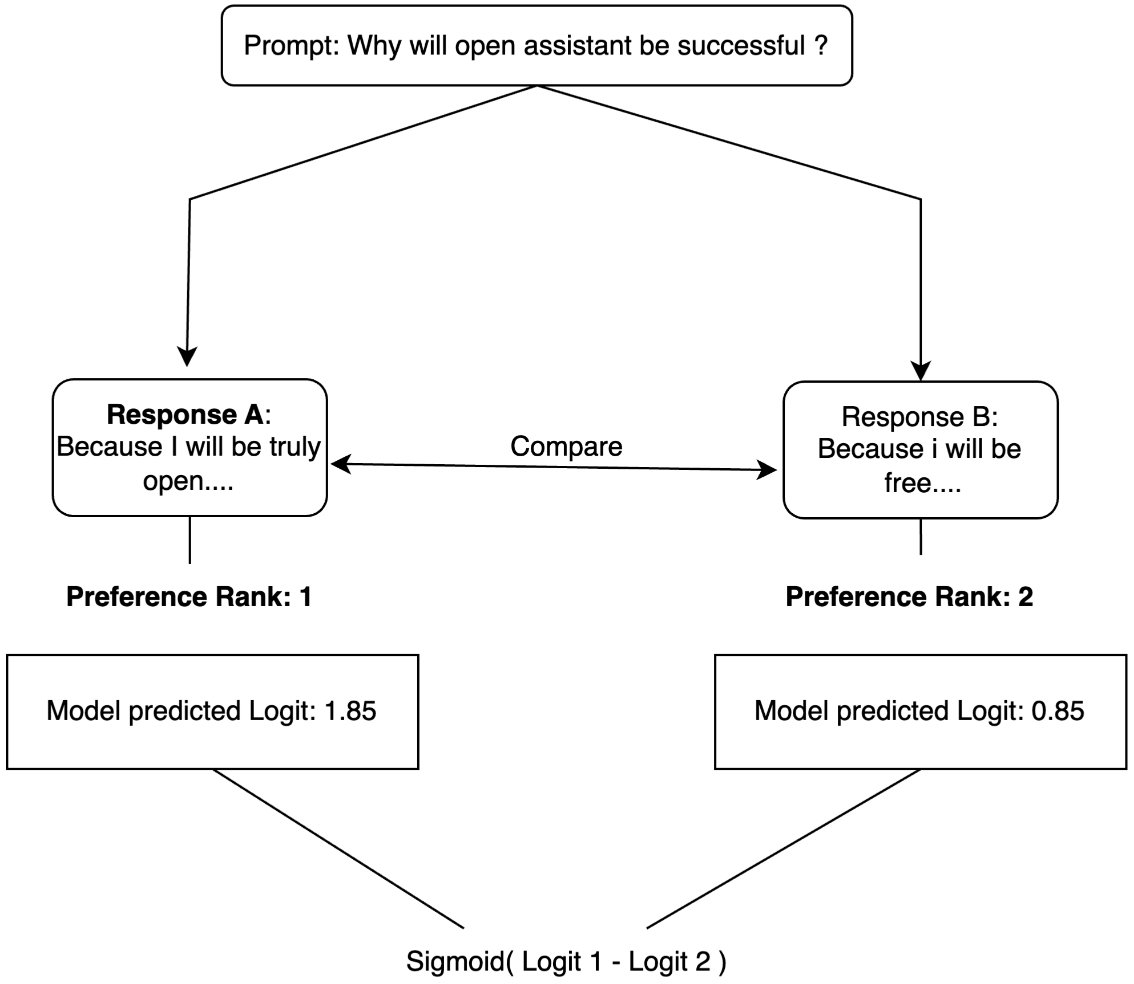
Research Method - Motivation

- Effectively maximize
  - Probability(Logit 1 > Logits)
  - Mean Resposne A is preferred over B

- Preference reward is implicitly learned

- Doesn’t consider each response independently

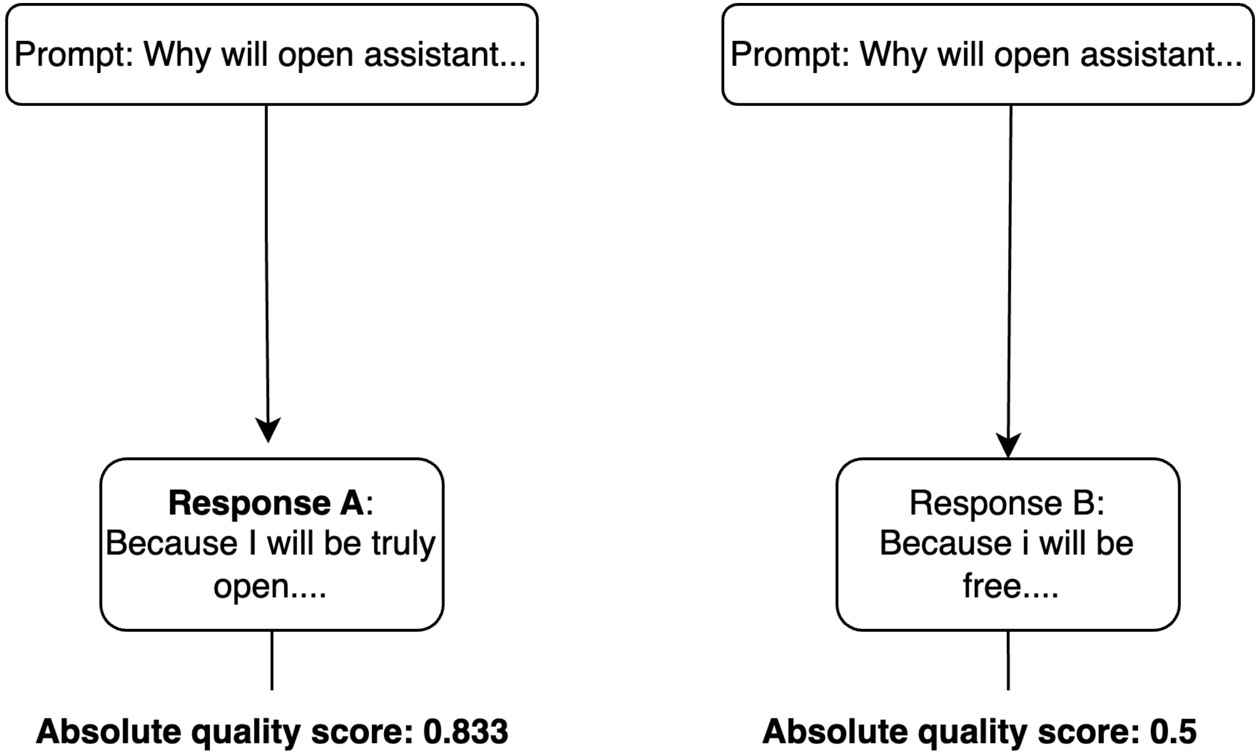
Sample from reward modelling dataset (Köpf et al. 2023)



Research Method - Motivation

- Consider each response independently and learn to provide feedback directly
- Softmax layer to constrain the predicted score
- Loss function: Binary cross entropy

Sample from reward modelling dataset (Köpf et al. 2023)



Research Method - Objective

- How do preference and absolute reward modelling impact the performance and generalisability of RLHF models on various datasets?
- What is the effect of varying the relative weights of preference-based and absolute reward signals during the RL fine-tuning process?
- What is the response quality and training efficiency of the \gls{RLHF} model, when using only preference-based reward, only absolute reward or a combination of both?

Research Method - Design & Development

Reward Modelling

Preference reward model

- Train from preference ranked data
- Implicity learn to provide feedback
- Custom loss function

Abs reward model

- Trained using absolute feedback dataset
- Directly learn to provide feedback
- Binary cross entroy loss

Research Method - Demonstration

The proposed solution will be demonstrated by training and fine-tuning several variants of the RLHF model. The variants include:

	Preference reward model	Abs reward model
CRS-RLHF	✓	✓
Preference-RLHF (baseline)	✓	✗
Abs-RLHF	✗	✓
SFT (baseline)	✗	✗

Research Method - Evaluation

GPT4 Evaluation

- 100 prompts sampled from
  - OASST
  - Koala
  - Vicuna
  - Helpful_base
- Evaluate RLHF + SFT
- Pairwise competition using GPT4

Human Evaluation

- 25 prompts sampled from
  - OASST
  - Koala
  - Vicuna
  - Helpful_base
- Evaluate only RLHF
- Preference ranking
- Repeat it 3 time to reduce variability

Research Method

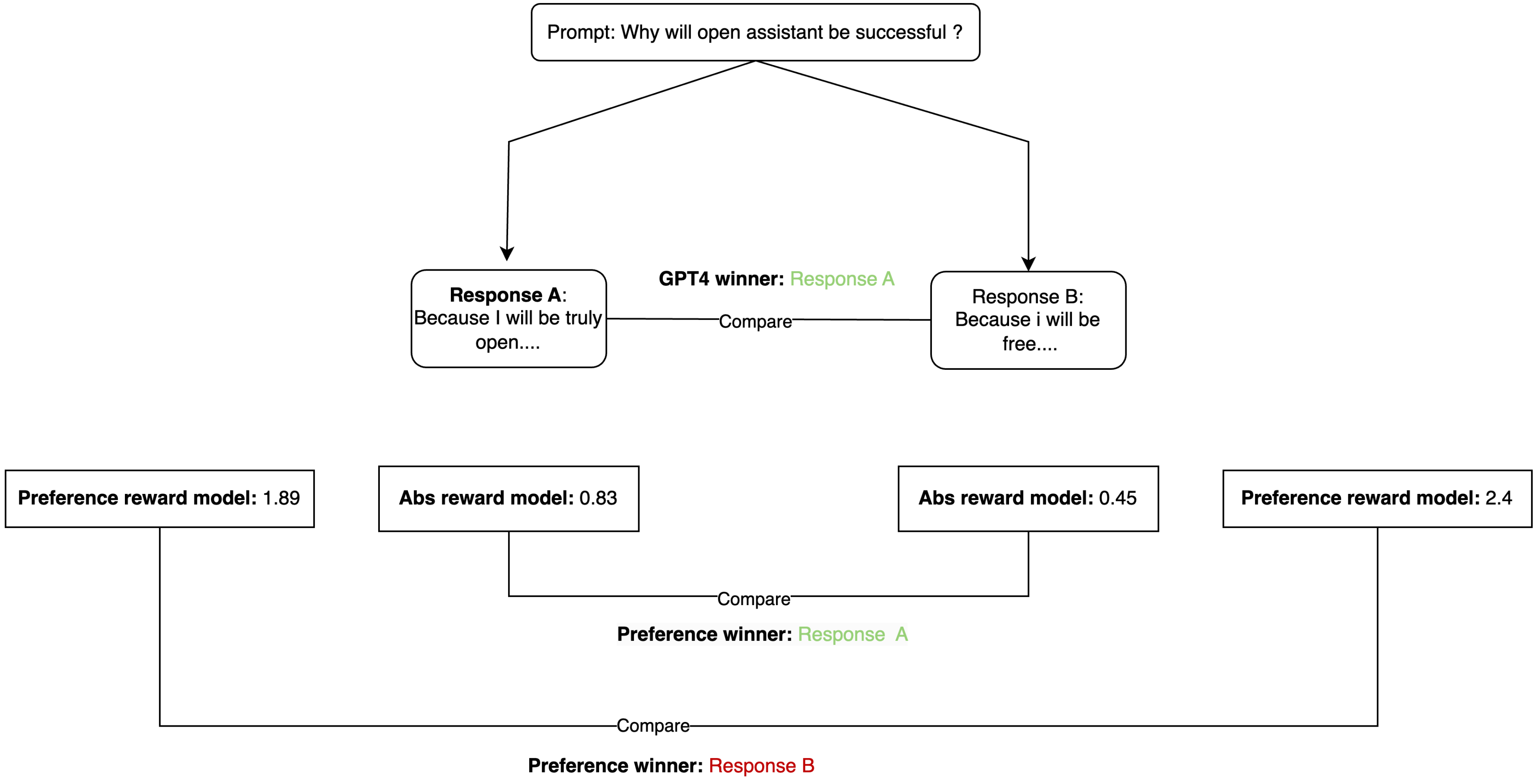
Communication & Contribution

- Communication through
  - Research paper
  - Presentation
  - Open-source models (on HuggingFace)
  - Code Repository

- Contribution:
  - Evaluate the impact of reward signals
  - Enhance the quality of responses

Discussion - Which reward model genralize better ?

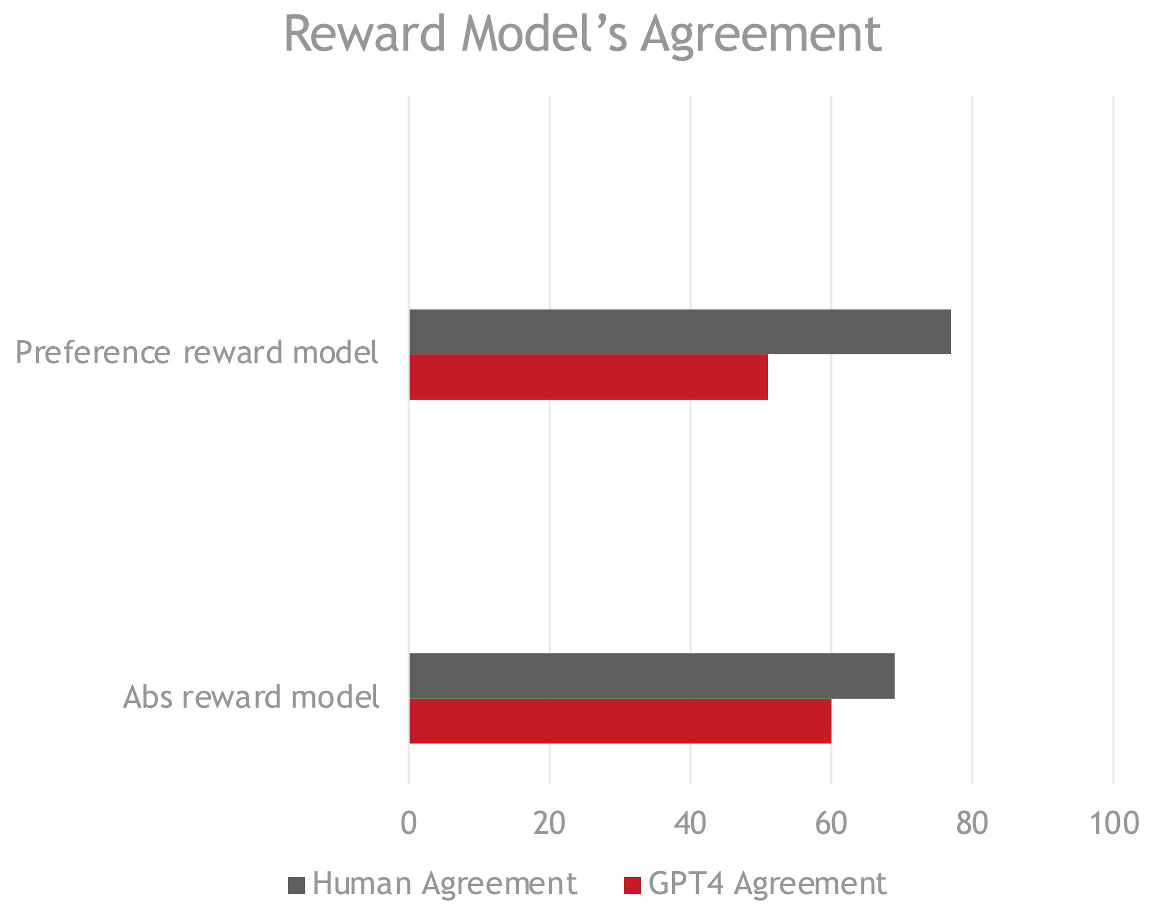
GPT4 aggrement



Discussion - Which reward model genralize better ?

Aggrement Result

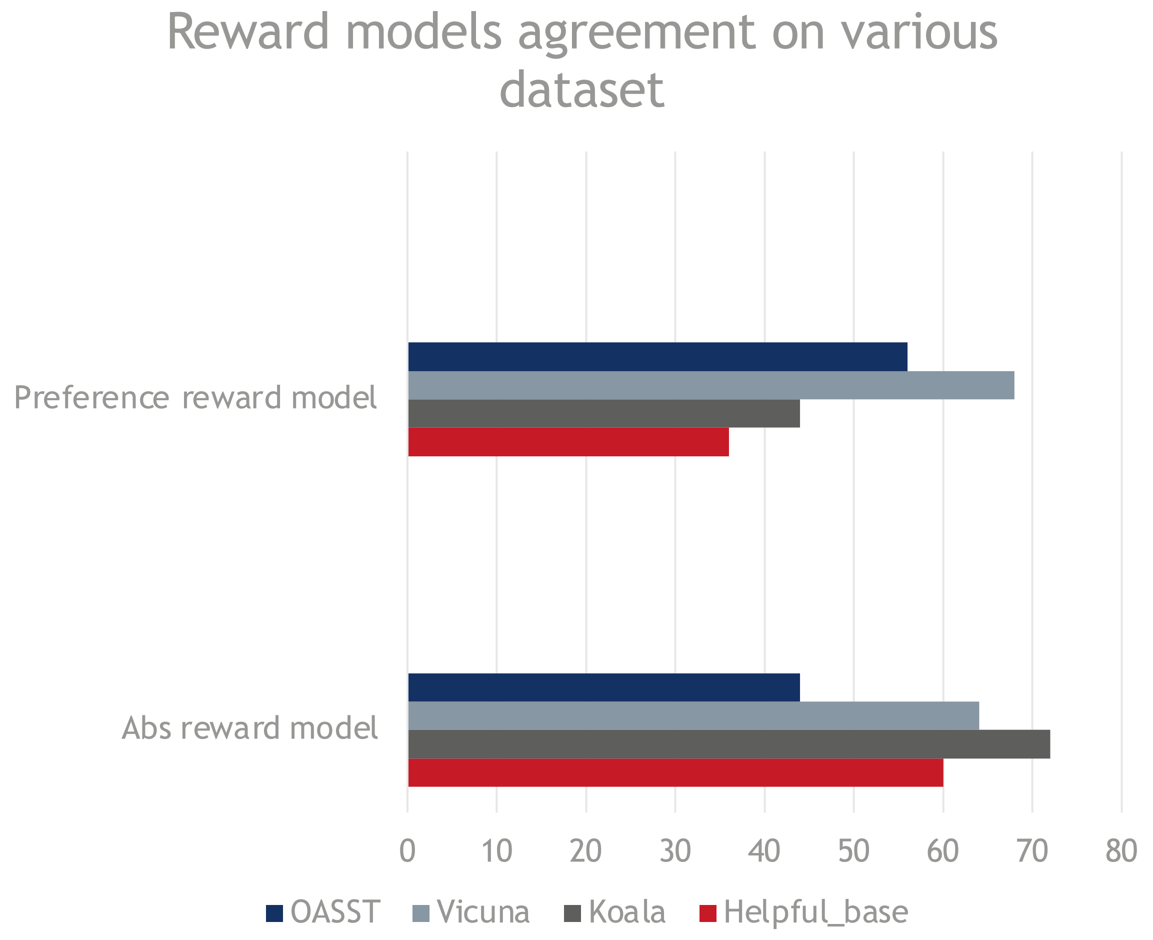
- GPT4 Agreement on **final_eval** dataset.
- Human Agreement on OASST eval dataset.
- Discrepancy in performance.



Discussion - Which reward model genralize better ?

Agreement on indvidual dataset

- GPT4 Agreement on individual dataset
- Preference only perform good on OASST
- Abs model general perform well except on OASST



Discussion - Which reward model genralize better ?

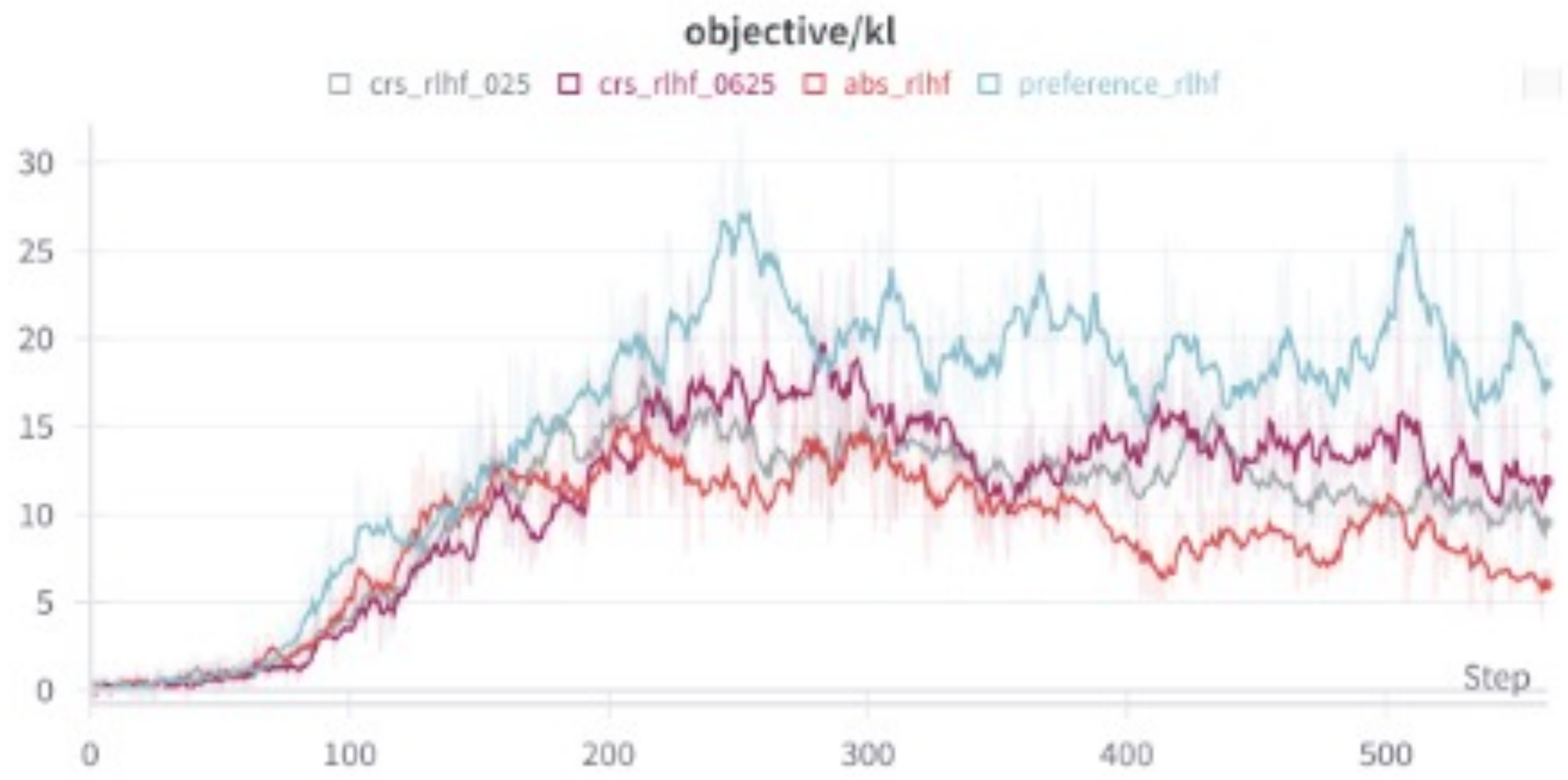
- Observation
  - Preference reward model performs better when prompts and generated response style is same as OASST dataset.
  - Abs reward model is able to consistently provide roboust signal but under fit on OASST due to noise.
- Hypothesis:
  - Implicit learning learn feature which are specific to a particular dataset.
  - Explict learning learn objective features aplicable to other dataset.

Discussion - Impact of varying weight of each model ?

- Combining reward score from both reward models.
- Train several RLHF model

	Preference reward weight	Abs reward weight
ABS_RLHF	0	1
CRS_RLHF_025	0.25	0.75
CRS_RLHF_0625	0.625	0.375
Prefernce_RLHF	1	0

Discussion - Impact of varying weight of each model ?



(Peffers et al., 2022)

Discussion - Comparative analysis

GPT4 Evaluation

Model (vs)	Preference_Rlhf	Abs_Rlhf	Crs_rlhf_025	SFT
Preference_Rlhf	-	34%	39%	29%
Abs_Rlhf	66%	-	63%	45%
Crs_rlhf_025	61%	37%	-	39%
SFT	71%	55%	61%	-

Discussion - Comparative analysis

Human Evaluation

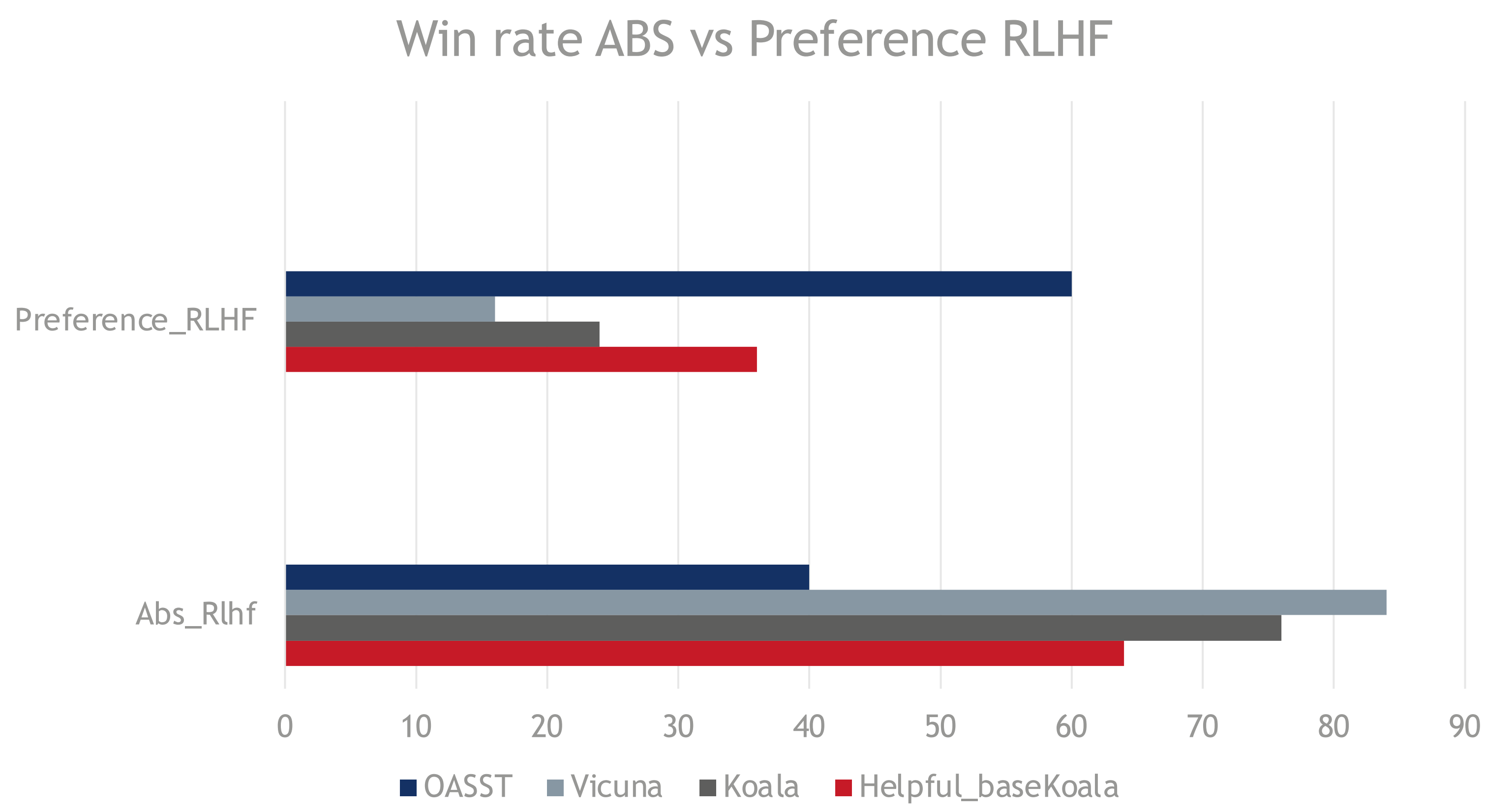
- Average winning point
  - Preference_RLHF → 0.54
  - Abs_RLHF → 0.79
  - CRS_RLHF_025 → 0.66
- High correlation with GPT4 evaluation

Model (vs)	Group 1	Group 2	Group 3
Preference_Rlhf	0.63	0.48	0.52
Abs_Rlhf	0.73	0.87	0.78
Crs_Rlhf_025	0.63	0.65	0.7

(Peffers et al., 2022)

Discussion - Comparative analysis of RLHF ?

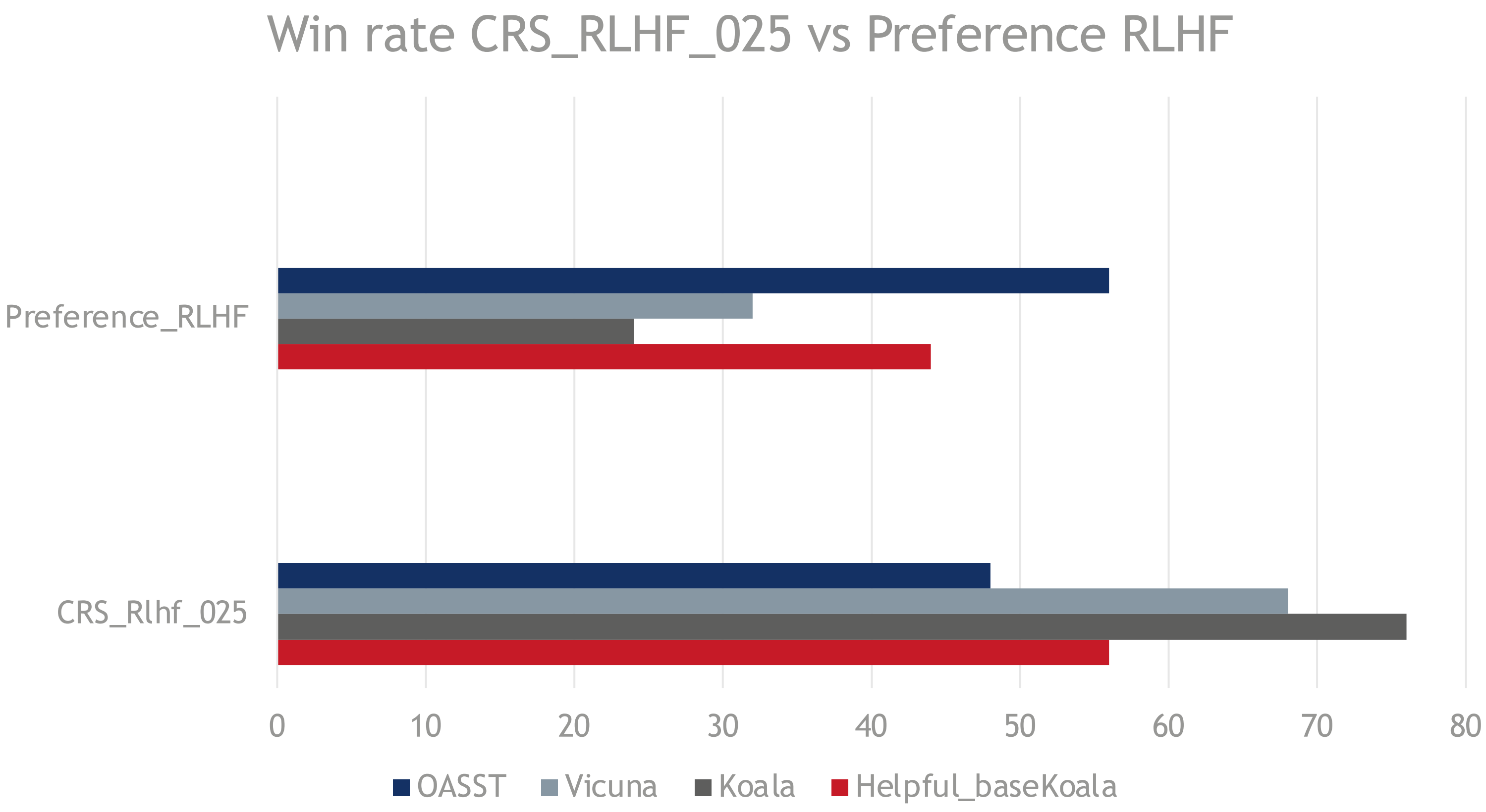
- Individual dataset win rate



(Peffers et al., 2022)

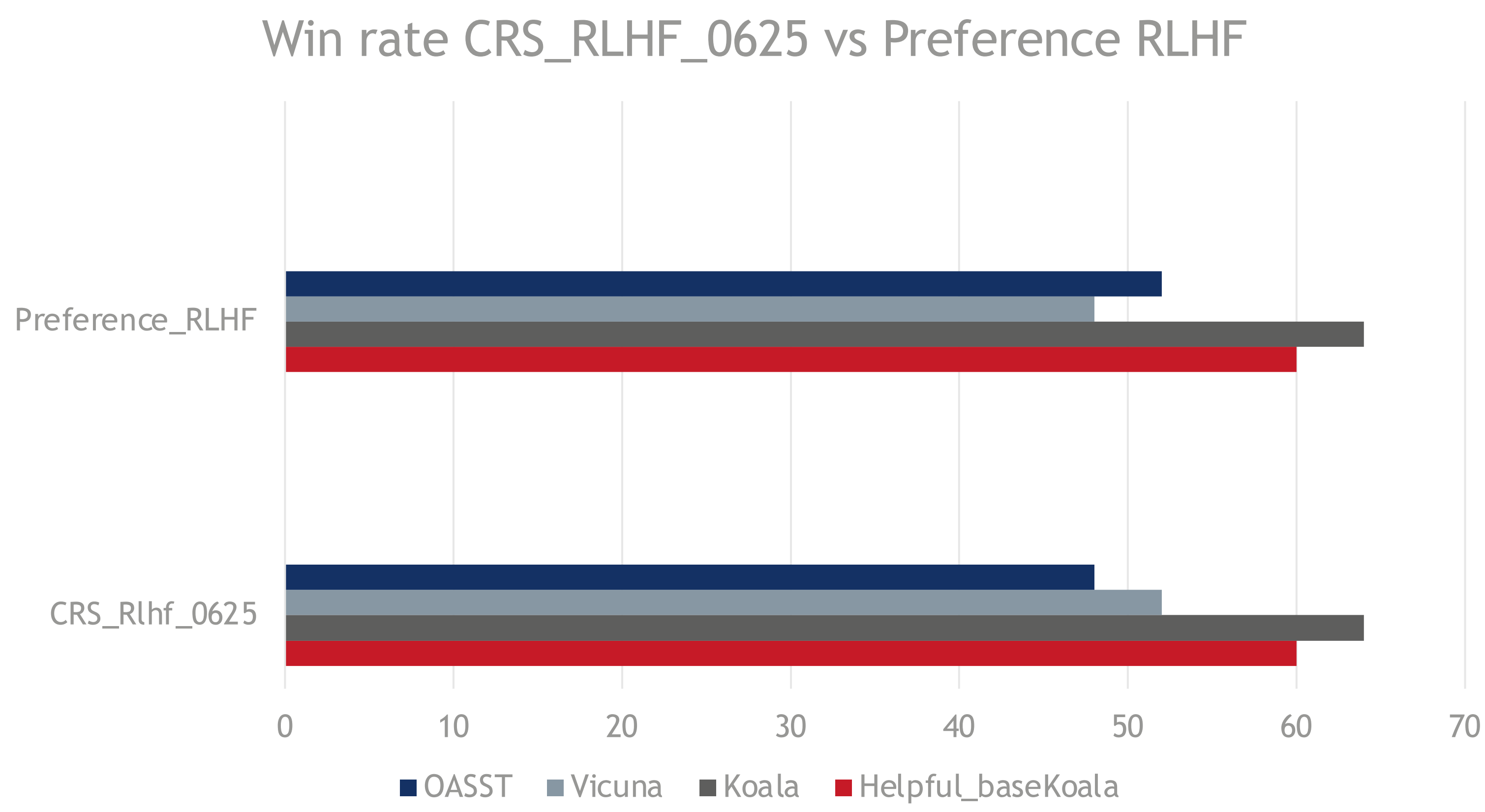
Discussion - Comparative analysis of RLHF ?

- Individual dataset win rate



Discussion - Comparative analysis of RLHF ?

- Individual dataset win rate



(Peffers et al., 2022)

Discussion - Comparative analysis of RLHF ?

Model (vs)	Preference_Rlhf	Abs_Rlhf	Crs_rlhf_025	SFT
Preference_Rlhf	-	34%	39%	29%
Abs_Rlhf	66%	-	63%	45%
Crs_rlhf_025	61%	37%	-	39%
SFT	71%	55%	61%	-

- SFT is better than RLHF
- Contrary to the work of Ouyang et al. (2022); Askell et al. (2021); Bai et al. (2022),

Discussion - Comparative analysis of RLHF ?

Confidence Interal with 95% confidence

Model (vs)	Preference_RLHF	CRS_RLHF_025	SFT
Abs_RLHF	66% ± 9.33%	63% ± 9.33%	45% ± 9.8%
Preference_RLHF		39% ± 9.6%	29% ± 8.93%

Conclusion

- Absolute reward model provide more roboust reward
- Model train purely with absolute reaward model perfrom better
- Combine both reward model work better for preference RLHF
- Abs_RLHF model better than all

Summary

- Background
- Motivation
- DSRM
- Discuss – Abs generalize better than preference
- Discussion combining preference and Abs result into worsen the performance.
- Discussion RLHF model trained with absolute reward model perform best but not against SFT

References

- Askell, A., Bai, Y., Chen, A., Drain, D., Ganguli, D., Henighan, T., Jones, A., Joseph, N., Mann, B., DasSarma, N., Elhage, N., Hatfield-Dodds, Z., Hernandez, D., Kernion, J., Ndousse, K., Olsson, C., Amodei, D., Brown, T., Clark, J., ... Kaplan, J. (2021). *A General Language Assistant as a Laboratory for Alignment* (arXiv:2112.00861). arXiv. https://doi.org/10.48550/arXiv.2112.0086
- Ouyang, L., Wu, J., Jiang, X., Almeida, D., Wainwright, C. L., Mishkin, P., Zhang, C., Agarwal, S., Slama, K., Ray, A., Schulman, J., Hilton, J., Kelton, F., Miller, L., Simens, M., Askell, A., Welinder, P., Christiano, P., Leike, J., & Lowe, R. (2022). *Training language models to follow instructions with human feedback* (arXiv:2203.02155; Version 1). arXiv. https://doi.org/10.48550/arXiv.2203.02155
- Peffers, K., Tuunanen, T., Rothenberger, M., & Chatterjee, S. (2007). A Design Science Research Methodology for Information Systems Research. *Journal of Management Information Systems*, *24*(3), 45–77. https://doi.org/10.2753/MIS0742-1222240302

university of koblenz
continue discovering

AmeerAli Khan

alikhan@uni-Koblenz.de