## Summary

Nash Learning from Human Feedback is an improvement on RLHF, in which the reward function is learned from preference. And the final goal for learning is not optimizing the reward function, instead, our goal is to find a strategy which satisfies nash equilibrium.

## Questions

1. In the experiment(Figure 2), we can find that the result of learning from preference is on par with or may be even better than the reward model, but if we take the efficiency into consideration, will learning from preference cost more computations, especially when they are introducing multiple strategies?
2. If the preference model is biased, could it mislead the training dynamics of the strategies? Is there any mechanism to reduce this risk?