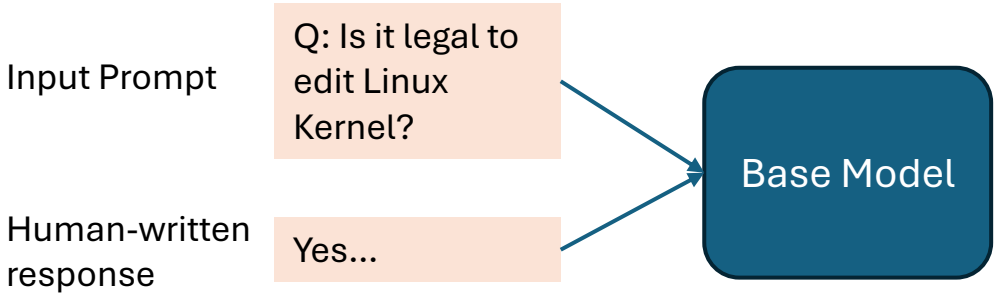
BRAIn – **B**ayesian **R**eward-conditioned **A**mortized **In**ference for natural language learning from feedback

Gaurav, Yatin, Tahira, Mayank, Gx, Dinesh, Sachin, Asim, Ramon

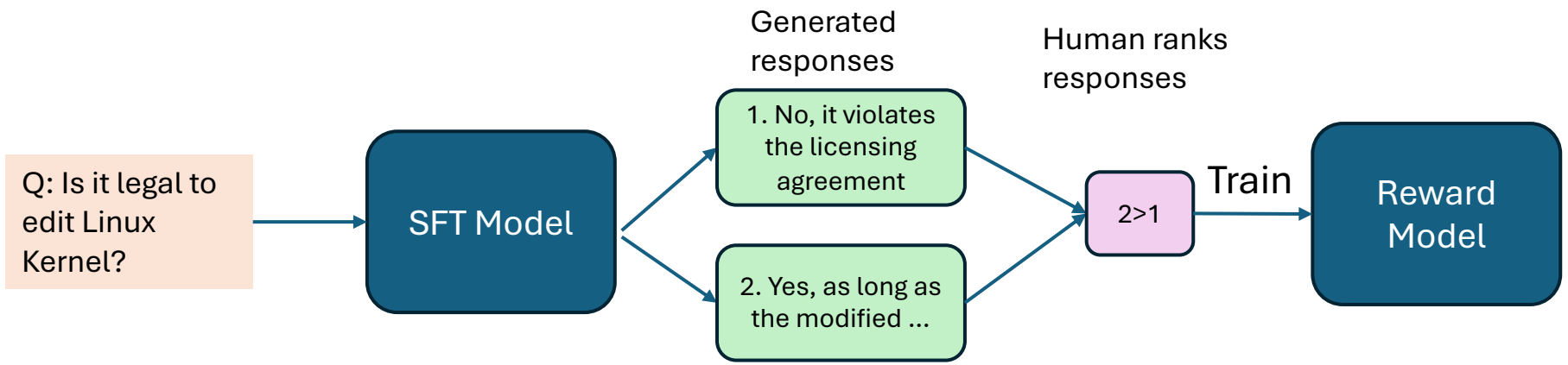
IBM Research AI

Reinforcement Learning from human feedback

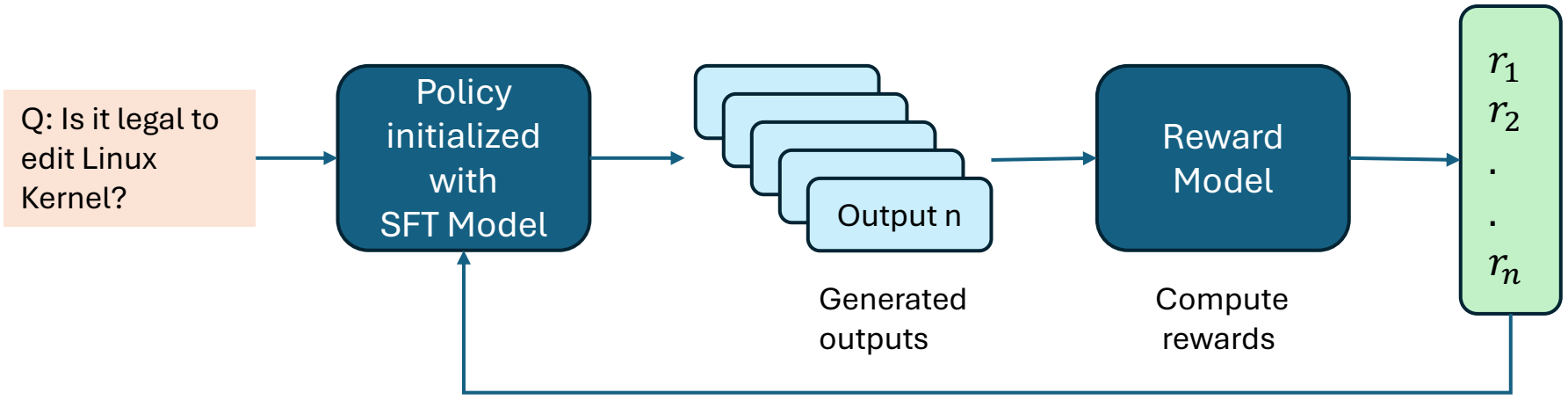
1. SFT - Finetune base LLM on human-written responses to create SFT model



2. Reward Model - Train a reward model based on human ranking of SFT model outputs



The RL stage of PPO-RLHF



Update policy based on rewards of generated outputs

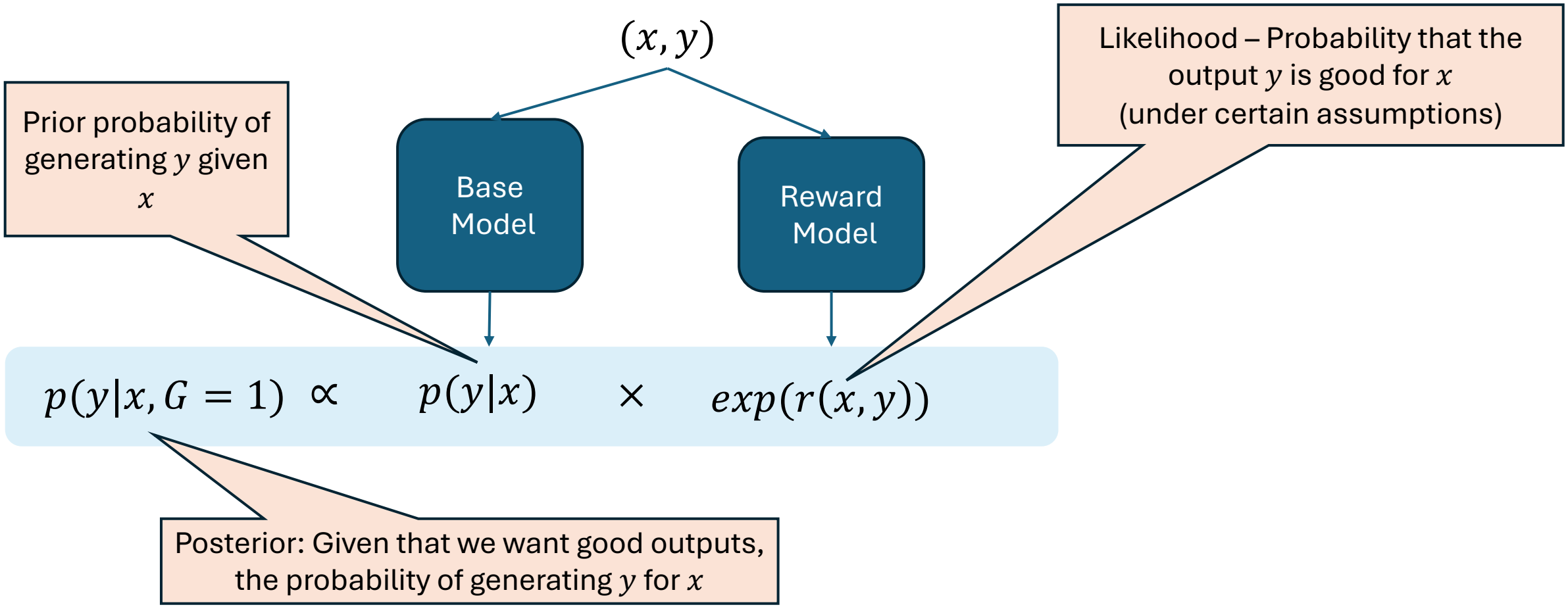
Maximize reward - The generated outputs should have high reward as decided by the reward model

- Train the policy to maximize the reward

Avoid catastrophic forgetting – The policy shouldn't deviate too much from the SFT model

- Minimize divergence between the policy and the SFT model.

Idea 1 – Use Bayes' rule to obtain target

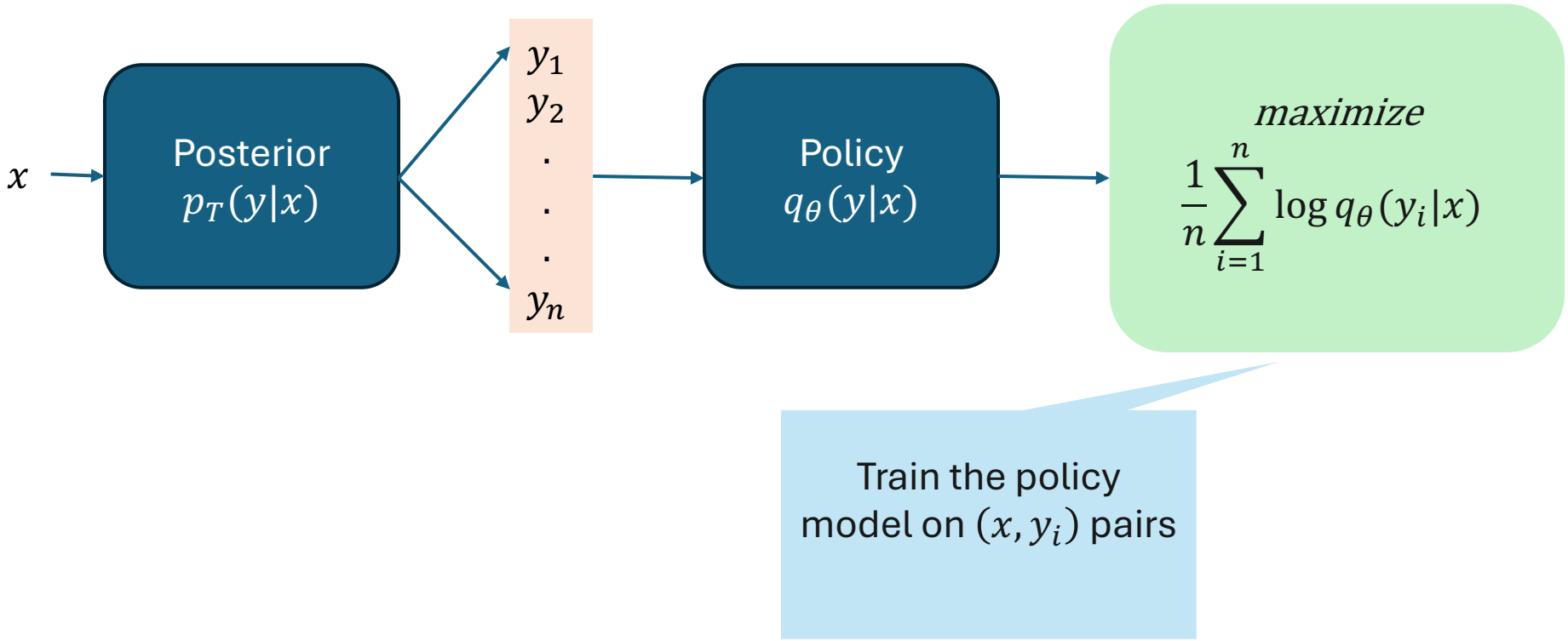


The normalization constant is intractable. Direct sampling is infeasible.

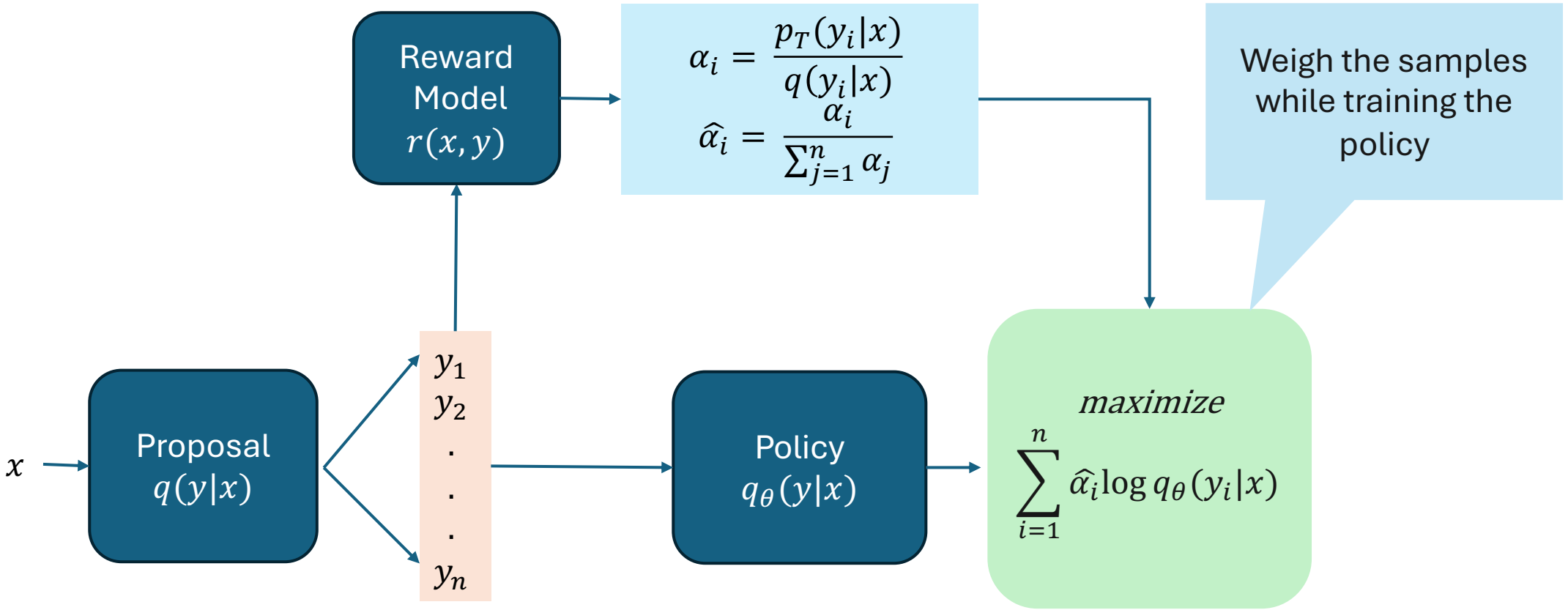
Train an LLM to mimic the posterior

Amortized Inference - Ideal

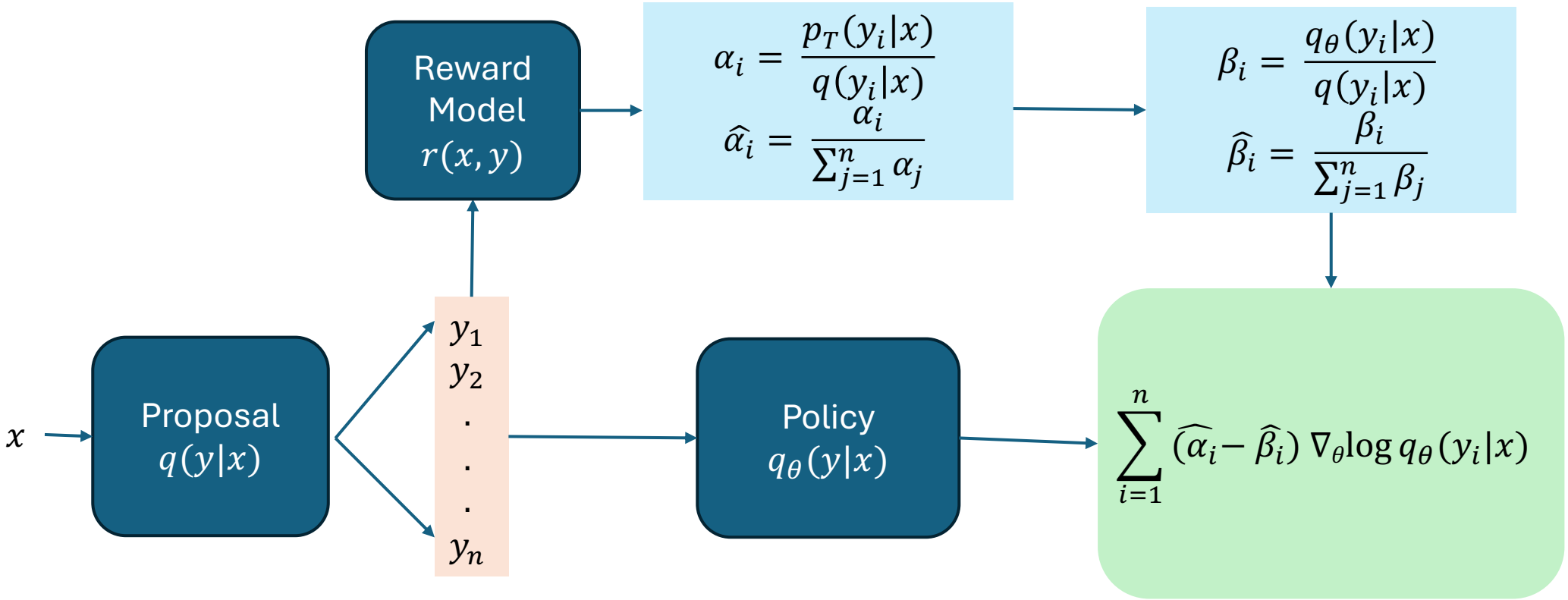
- Train an LLM to mimic the posterior.



Amortized Inference with importance weights



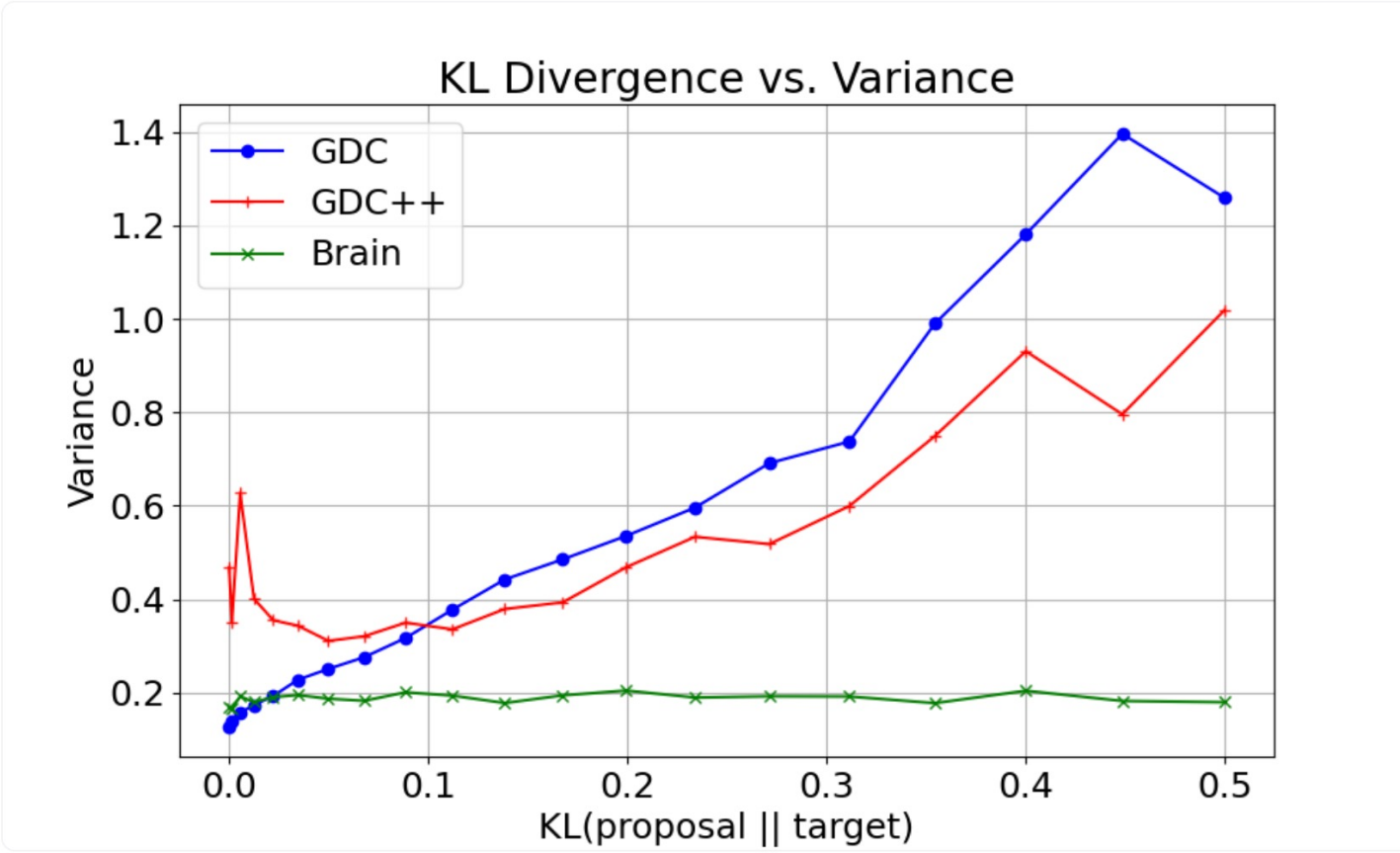
Idea 2 - Variance reduction with a self-normalized baseline



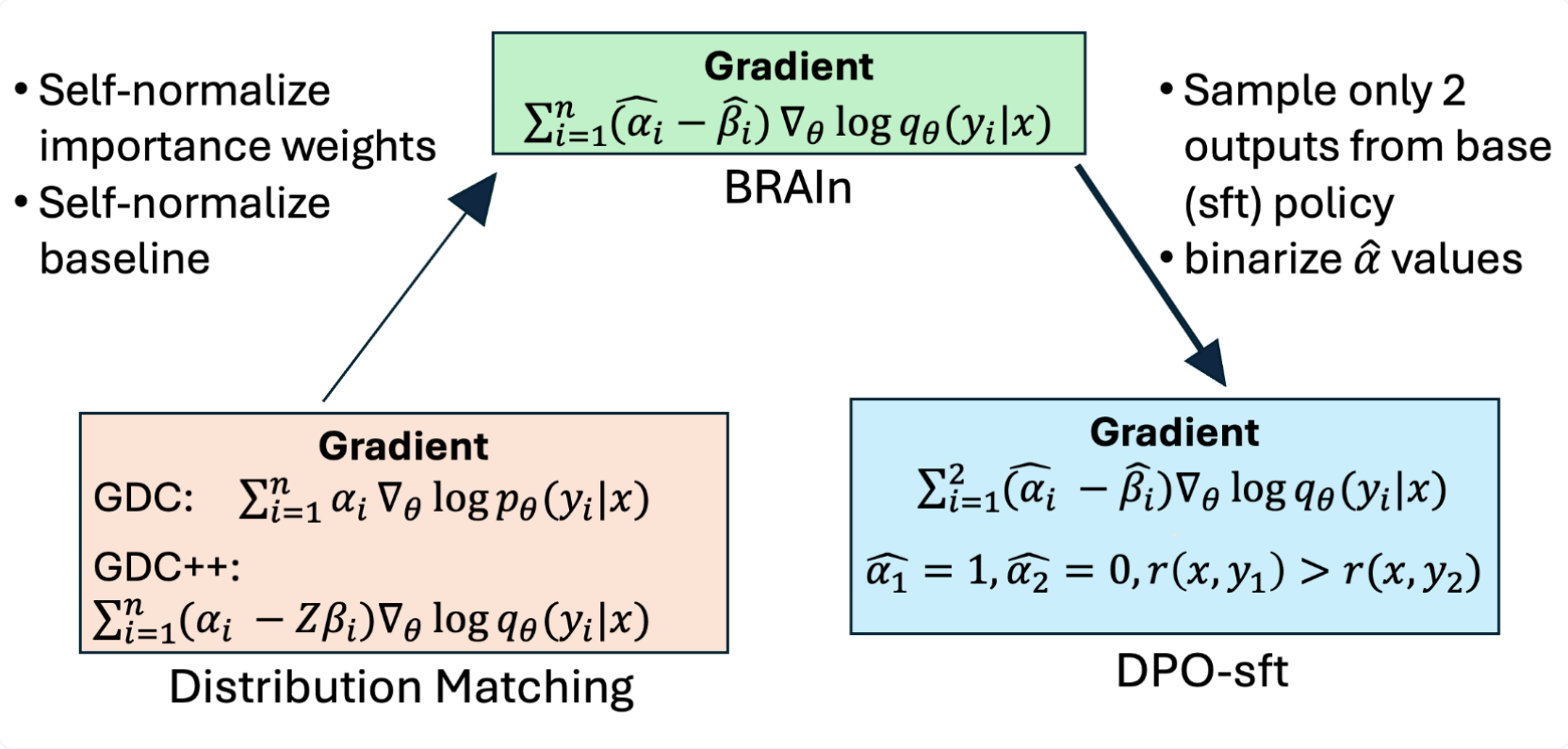
How important is baseline?

	BRAIn	w/o self-norm	w/o baseline
TL;DR	95.2	61.4	61.1
AnthropicHH	95.4	59.1	58.3

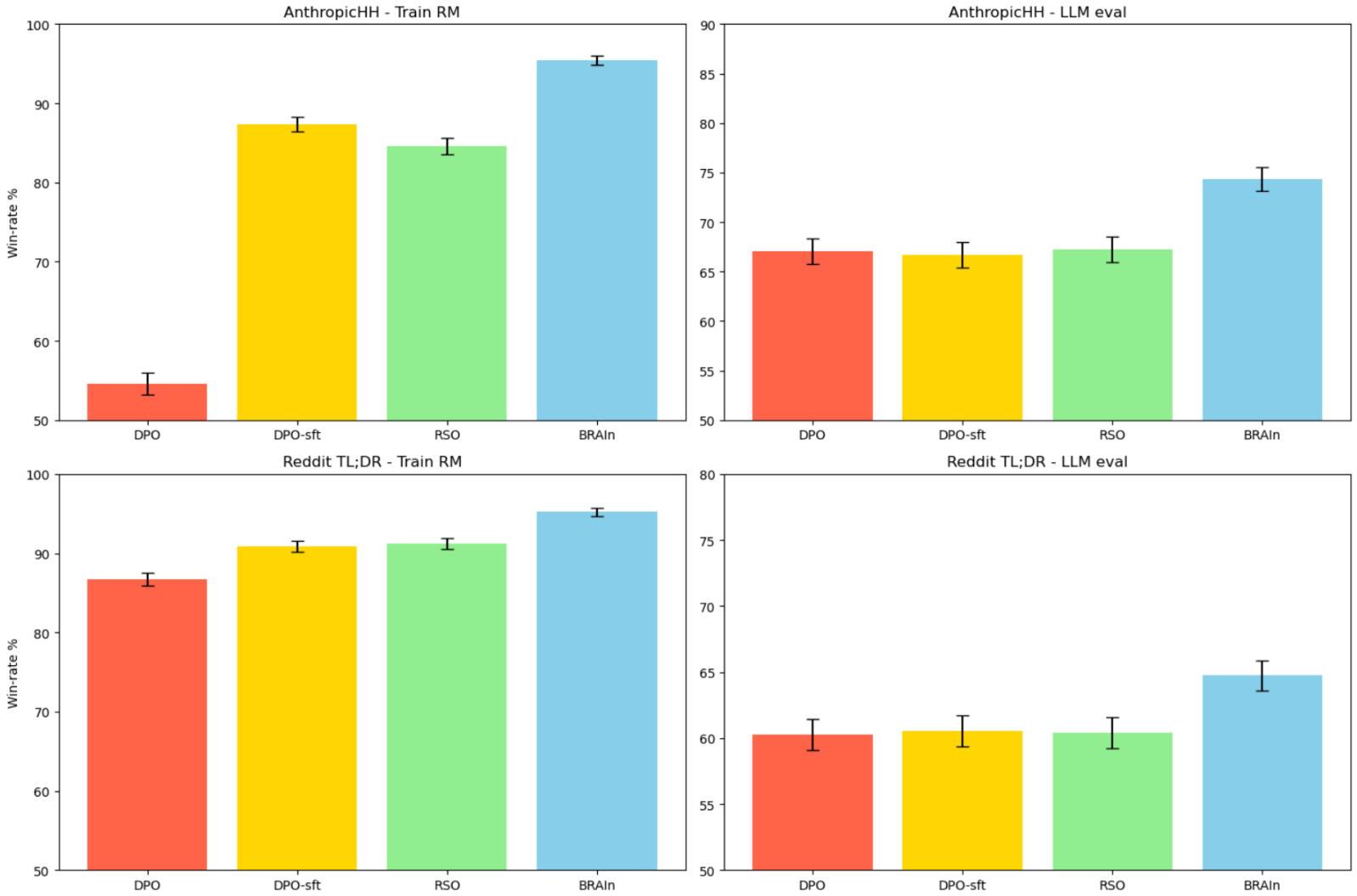
Table 3. Effect of self-normalized baseline on the performance of various models.



DPO, BRAIn & Distribution Matching

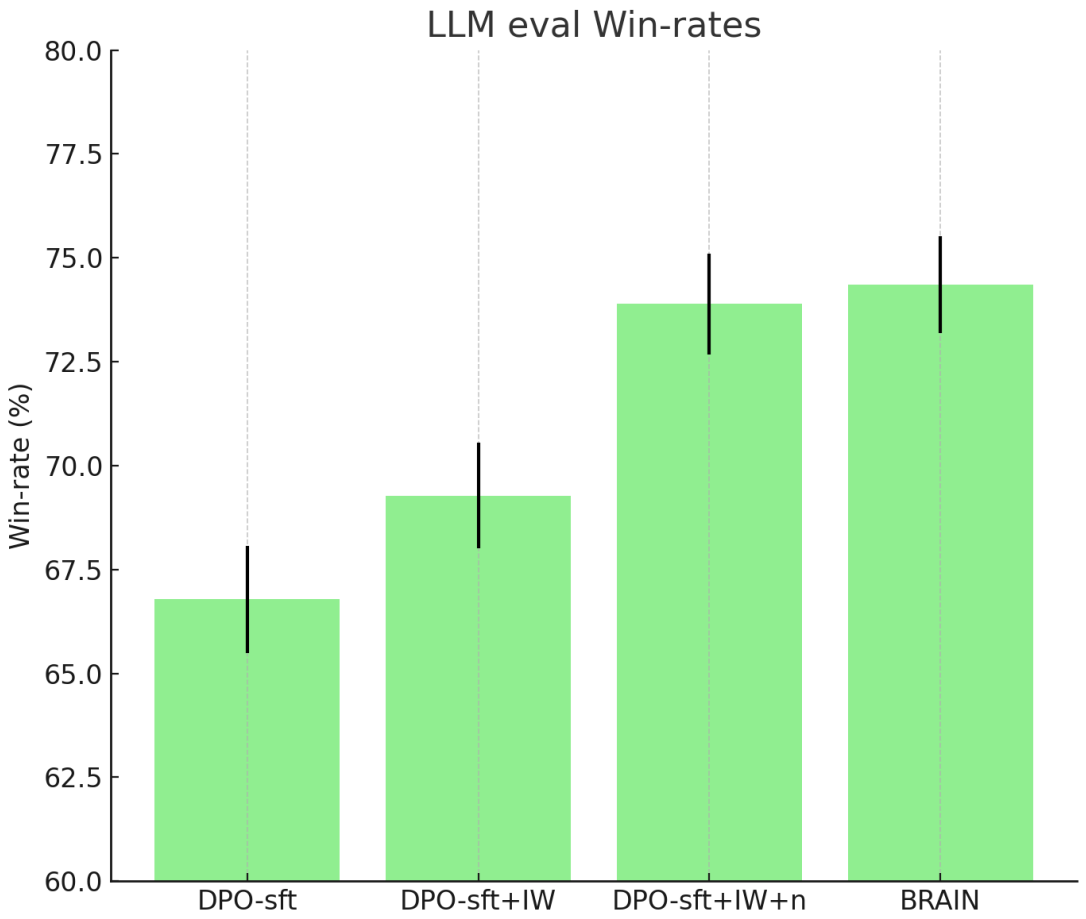


Performance comparison

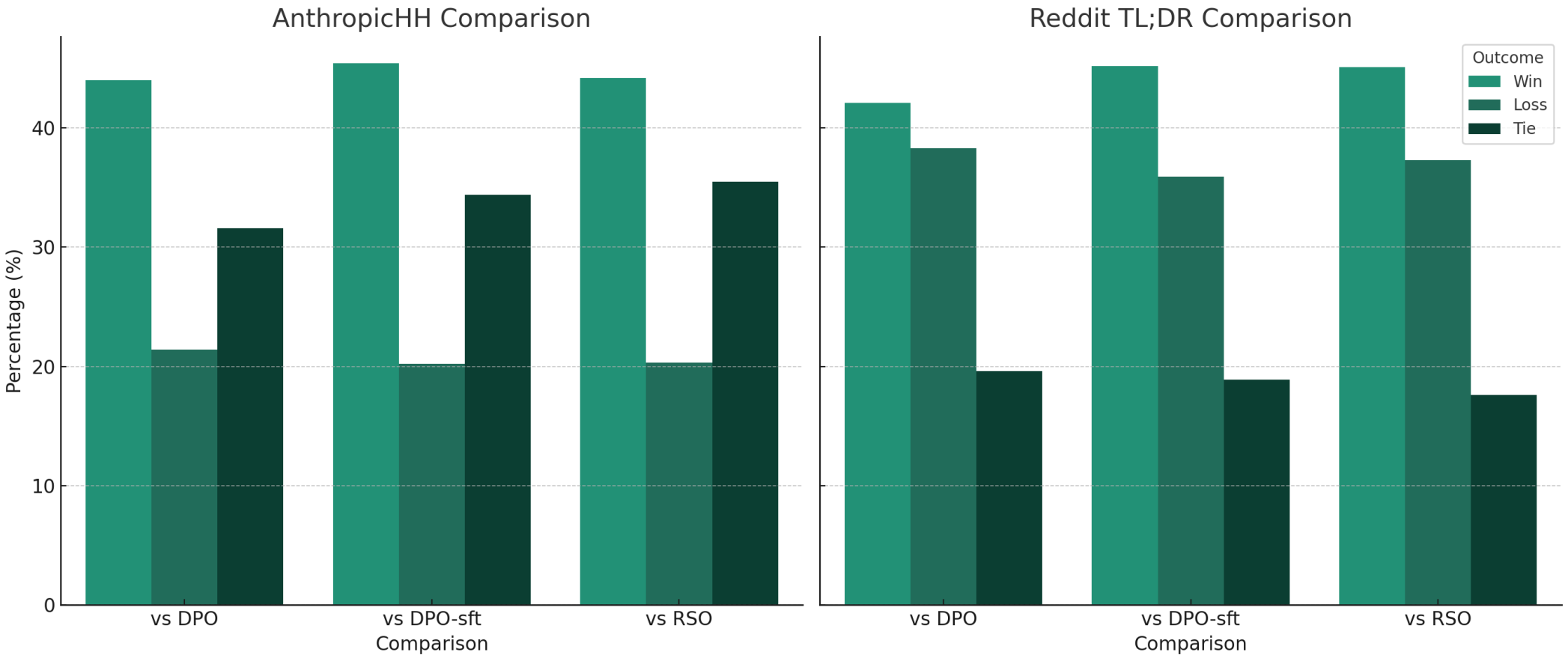


Significantly outperforms DPO and RSO on AnthropicHH and Reddit TL;DR

From DPO to BRAIn



GPT-4 evaluation (Head-to-Head)



https://arxiv.org/pdf/2402.02479.pdf

Fewer samples per prompt

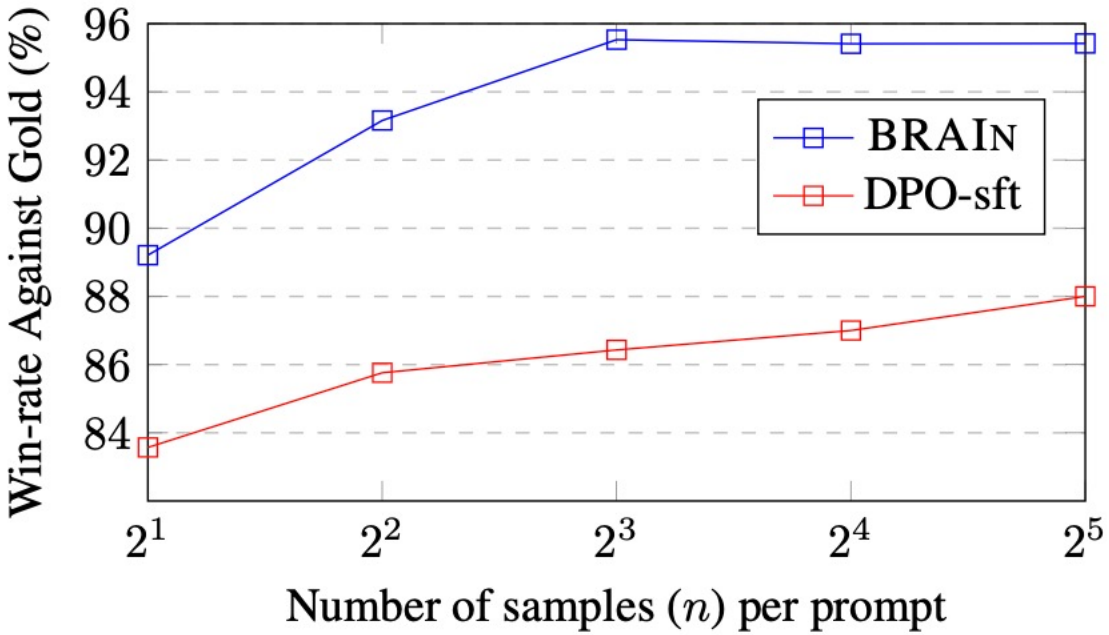


Figure 2: Plot of Win-rate Against Gold as a function of the Number of Samples per Prompt.

Conclusions

- Bayesian posterior can be an effective choice for the target distribution in RLHF
  - It accounts for the reward modelling assumptions made.
- Variance reduction is key for distribution matching to be effective.
  - A self-normalized baseline results in reduced variance.
- The DPO loss can be derived a special case of distribution matching with a self-normalized baseline, aka, BRAIN