

Regularizing Hidden States Enables Learning Generalizable Reward Model for LLMs

Rui Yang, Ruomeng Ding, Yong Lin, Huan Zhang, Tong Zhang



UNIVERSITY OF
ILLINOIS
URBANA - CHAMPAIGN



Background

- What is a reward model for LLMs?
 - A model to assess the human preference for responses (generated by LLMs)

- What is human preference?
 - Typically, it includes characteristics such as quality, helpfulness, safety...
 - Sometimes it can be hard to explicitly define
 - Pair-wise comparison can be used as human preference

Question: If the FED buys bonds in the secondary market (A) the money supply will increase. (B) the money supply will decrease. (C) the money supply will not be affected. (D) the discount rate would be affected. Output the answer only.

Assistant A:

The Federal Reserve buys bonds in the secondary market to increase the money supply.

Assistant B:

(A) the money supply will increase.

Follow-up Question: How does it affect my daily life? Give 3 examples.

Assistant A:

The Federal Reserve buys bonds in the secondary market to increase the money supply. This affects your daily life by:

1. Increasing the money supply.
2. Increasing the money supply and lowering the discount rate.
3. Increasing the money supply and lowering the discount rate and increasing the money supply.

Assistant B:

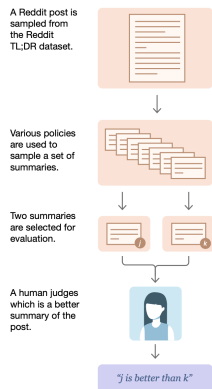
When the Federal Reserve buys bonds in the secondary market, it can have several effects on the economy that can impact your daily life. Here are three examples:

1. Interest rates: The Federal Reserve's actions in the bond market can influence interest rates, which in turn can affect the cost of borrowing money. For example, if the Federal Reserve buys bonds and drives up demand for them, interest rates on things like mortgages, car loans, and credit cards may go up, making it more expensive to borrow money.
2. Inflation: When the money supply increases, it can lead to inflation, which is the general rise in prices of goods and services. This can impact your daily life by making it more expensive to buy things like groceries, gas, and other necessities.
3. Employment: The Federal Reserve's actions in the bond market can also have an impact on employment. For example, if the Federal Reserve's actions lead to economic growth and increased demand for goods and services, it can create more job opportunities and boost employment.

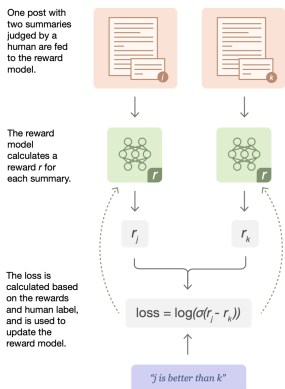
Background

- Reward modeling is commonly adopted for LLM alignment, math reasoning ...
- Useful for both training-time and inference-time optimization
- Several advancements on process reward model, fine-grained rewards...

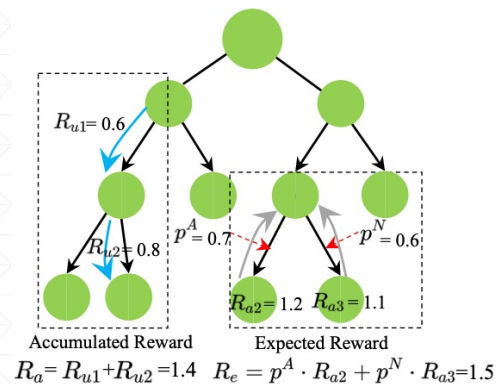
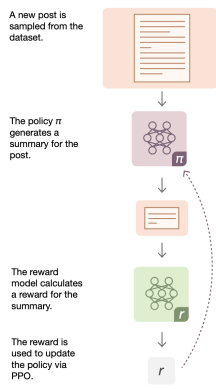
1 Collect human feedback



2 Train reward model



3 Train policy with PPO



Background

Basic ways for reward modeling

- Absolute score
 - Predict the absolute labels R

$$L(\psi) = -E_{(x,y,R)\sim D}[\log r_\psi(R|x,y)]$$

Examples: math reward model, HELPSTEER

- Pair-wise Comparison

- Bradley-Terry model

Maximize the log-likelihood of $p(y_c > y_r|x) \rightarrow$ maximize the relative reward gap

$$L(\psi) = -E_{(x,y_c,y_r)\sim D}[\log \sigma(r_\psi(x,y_c) - r_\psi(x,y_r))]$$

Examples: InstructGPT, ChatGPT, GPT4, Claude, Gemini ...

- General preference model

$$L(\psi) = -E_{(x,y_c,y_r)\sim D}[\log p_\psi(y_c > y_r|x)]$$

Example: LLM-as-a-judge, generative classifier

Background

Challenges of reward modeling for LLMs

- Heterogenous and noisy data
 - Human labeled data is inherently heterogenous
 - Currently, many datasets are mixed with human and GPT labeled data
 - Labelers can inadvertently give erroneous labels (~ 20%)

Human Prompt → **How do you study for a test efficiently? (Incorrect Preference)**

Chosen Response → That's a tough one.

Rejected Response → Studying for a test can be very challenging, but there are several effective strategies you can use to study efficiently. First, it's important to break down the material that you need to learn into smaller parts and focus on learning each part thoroughly. In addition, it's helpful to set aside plenty of time each day to study and review the material regularly. Studying efficiently also involves avoiding distractions and staying well-rested, as these can reduce your ability to learn effectively. Overall, using these strategies can help you prepare for a test in an efficient and thorough manner.

Generalizable Reward Modeling

- The goal of our work is to enhance the generalization of preference learning to unseen data and alleviate the overoptimization issue in RLHF
- Our basic idea is to regularize the reward model during preference learning
 - In contrast, previous research regularizes the optimization during RL process
 - How to perform such regularization?
 - One idea is to leverage the pretrained feature
 - Pretrained feature provides a good initialization for reward modeling that avoids overfitting and generalizes better when we have limited data

Table 7: Reward model performance trained with 8K data.

Reward Model	Unified Feedback (ID)	HHH Alignment (OOD)	MT Bench (OOD)
Classifier (Frozen)	62.2	68.8	67.6
Classifier (Baseline)	66.1	65.1	67.7

- Here we define OOD as different $p(x), p(y_c), p(y_r), p(y_c > y_r | x)$

Generalizable Reward Modeling

- However, the frozen model underfits when the training dataset is larger

Reward Model	Unified Feedback	HHH Alignment	MT Bench	40K training data from UnifiedFeedback
Classifier (Frozen)	63.9	68.6	68.2	
Classifier (baseline)	68.8	70.3	69.1	

- Can we keep the advantage of the learned feature while finetuning all parameters?
 - Pretrained model are trained on diverse text-generation tasks
 - Can we keep the text-generation ability of the hidden states during preference learning?

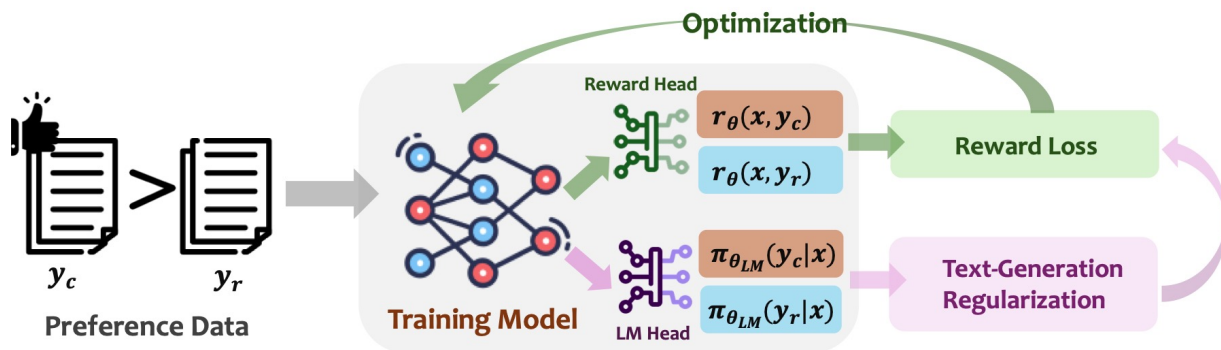
Generalizable Reward Modeling

- Generally, reward model is finetuned from a language model with a randomly initialized reward head
- We keep the original language head to perform text-generation as a regularization
- How to define the text-generation term?
 - Recall the DPO objective

$$\mathcal{L}_{\text{reward}}(\theta) = -\mathbb{E}_{(x, y_c, y_r) \sim D} [\log(\sigma(r_\theta(x, y_c) - r_\theta(x, y_r)))]$$

$$\mathcal{L}_{\text{DPO}}(\theta_{\text{LM}}) = -\mathbb{E}_{(x, y_c, y_r) \sim D} \left[\log \sigma \left(\beta \log \left(\frac{\pi_{\theta_{\text{LM}}}(y_c | x)}{\pi_{\text{ref}}(y_c | x)} \right) - \beta \log \left(\frac{\pi_{\theta_{\text{LM}}}(y_r | x)}{\pi_{\text{ref}}(y_r | x)} \right) \right) \right]$$

- We can train the language head with the same data as reward modeling



Generalizable Reward Modeling

- The overall loss function

$$\mathcal{L}_{\text{total}} = (1 - \alpha)\mathcal{L}_{\text{reward}} + \alpha\mathcal{L}_{\text{reg}}.$$

- We can have different forms of text-generation regularization

$$\mathcal{L}_{\text{DPO}}(\theta_{\text{LM}}) = -\mathbb{E}_{(x, y_c, y_r) \sim D} \left[\log \sigma \left(\beta \log \left(\frac{\pi_{\theta_{\text{LM}}}(y_c | x)}{\pi_{\text{ref}}(y_c | x)} \right) - \beta \log \left(\frac{\pi_{\theta_{\text{LM}}}(y_r | x)}{\pi_{\text{ref}}(y_r | x)} \right) \right) \right]$$

$$\mathcal{L}_{\text{DPO-noref}}(\theta_{\text{LM}}) = -\mathbb{E}_{(x, y_c, y_r) \sim D} \left[\log \sigma \left(\beta \log \left(\frac{\pi_{\theta_{\text{LM}}}(y_c | x)}{\pi_{\theta_{\text{LM}}}(y_r | x)} \right) \right) \right]$$

$$\mathcal{L}_{\text{SFT}}(\theta_{\text{LM}}) = -\mathbb{E}_{(x, y_c) \sim D} [\log \sigma (\beta \log (\pi_{\theta_{\text{LM}}}(y_c | x)))]$$

Generalizable Reward Modeling

- Other interpretation aspects
 - DPO reg: use implicit reward learning as the regularization

$$\mathcal{L}_{\text{reward}}(\theta) = -\mathbb{E}_{(x, y_c, y_r) \sim D} [\log (\sigma (r_{\theta}(x, y_c) - r_{\theta}(x, y_r)))]$$

$$\mathcal{L}_{\text{DPO}}(\theta_{\text{LM}}) = -\mathbb{E}_{(x, y_c, y_r) \sim D} \left[\log \sigma \left(\beta \log \left(\frac{\pi_{\theta_{\text{LM}}}(y_c | x)}{\pi_{\text{ref}}(y_c | x)} \right) - \beta \log \left(\frac{\pi_{\theta_{\text{LM}}}(y_r | x)}{\pi_{\text{ref}}(y_r | x)} \right) \right) \right]$$

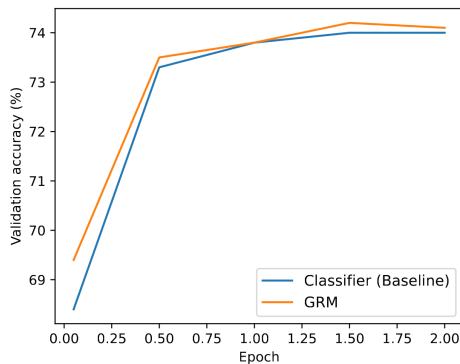
- SFT reg: learn a reward model against an adversarial policy

$$\theta = \arg \min_{\theta} \left\{ \mathcal{L}_{\text{reward}}(\theta) + \gamma \max_{\pi} J(\theta, \pi) \right\}$$

- Try to avoid reward model with higher value in RL optimization (overoptimization)

Experiments

- Evaluation on reward modeling benchmark and RLHF
- Train on UnifiedFeedback (400K and 40K), evaluate on 8K holdout set (ID)
- Evaluate on HHH-alignment, MT-bench (OOD)
- Training for 2 epochs with LoRA, 1 epoch for full-parameter



Experiments

Evaluation on reward modeling

- ID and OOD evaluation:
 - GRM significantly enhances the generalization capability of reward models
 - Larger improvement on OOD tasks and smaller dataset
 - Outperform other baselines (including ensemble)
 - SFT regularization is the most stable one

Table 1: Results on ID and OOD evaluation with **400K training data** from UnifiedFeedback. The best performance in each task is in bold and the second best one is underlined.

Reward Model	Unified Feedback	HHH Alignment	MT Bench
Classifier (Frozen)	63.8	66.4	69.5
Classifier (baseline)	72.1	73.4	71.2
Classifier + margin	72.0	75.0	72.6
Classifier + label smooth	71.5	72.1	71.2
Classifier + Ensemble	72.8	76.8	73.7
GRM w/ dpo (ours)	<u>73.8</u>	79.2	<u>73.4</u>
GRM w/ dpo-noref (ours)	73.9	<u>79.7</u>	73.0
GRM w/ sft (ours)	73.2	79.8	<u>73.4</u>

Table 2: Results on ID and OOD evaluation with **40K training data** from UnifiedFeedback. The best performance in each task is in bold and the second best one is underlined.

Reward Model	Unified Feedback	HHH Alignment	MT Bench
Classifier (Frozen)	63.9	68.6	68.2
Classifier (baseline)	68.8	70.3	69.1
Classifier + margin	69.6	69.8	71.0
Classifier + label smooth	68.5	68.8	71.9
Classifier + Ensemble	69.9	72.2	71.1
GRM w/ dpo (ours)	70.2	71.6	71.3
GRM w/ dpo-noref (ours)	<u>71.4</u>	<u>76.6</u>	<u>72.1</u>
GRM w/ sft (ours)	71.5	78.7	73.0

Experiments

Evaluation on reward modeling

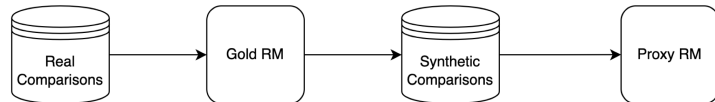
- Achieve strong performance on RewardBench in July
- Best 8B BT RM at that time
- Best 2B RM

Model	Average	Chat	Chat Hard	Safety	Reasoning
Ray2333/GRM-Gemma-2B-rewardmodel-ft (Ours, 2B)	84.7	89.4	75.2	85.5	88.8
openai/gpt-4o-2024-05-13	84.6	96.6	70.4	86.5	84.9
sfairXC/FsfairX-LLaMA3-RM-v0.1 (8B)	84.4	99.4	65.1	86.8	86.4
Nexusflow/Starling-RM-34B	82.6	96.9	57.2	87.7	88.5
Ray2333/GRM-Gemma-2B-sftreg(Ours, 2B)	75.3	95.5	48.7	80.0	76.8
berkeley-nest/Starling-RM-7B-alpha (7B)	74.6	98	43.4	88.6	74.6
Ray2333/Gemma-2B-rewardmodel-baseline(Ours, 2B)	73.7	94.1	46.1	79.6	75.0
stabilityai/stablelm-zephyr-3b (3B)	73.1	86.3	60.1	70.3	75.7
openbmb/UltraRM-13b (13B)	71.3	96.1	55.3	45.8	82

Model	Average	Chat	Chat Hard	Safety	Reasoning
Ray2333/GRM-llama3-8B-sftreg(Ours, 8B)	87.0	98.6	67.8	89.4	92.3
Ray2333/GRM-llama3-8B-distill(Ours, 8B)	86.1	98.3	68.4	86.1	91.3
openai/gpt-4-0125-preview	85.9	95.3	74.3	87.2	86.9
sfairXC/FsfairX-LLaMA3-RM-v0.1 (8B)	84.7	99.4	65.1	87.8	86.4

Experiments

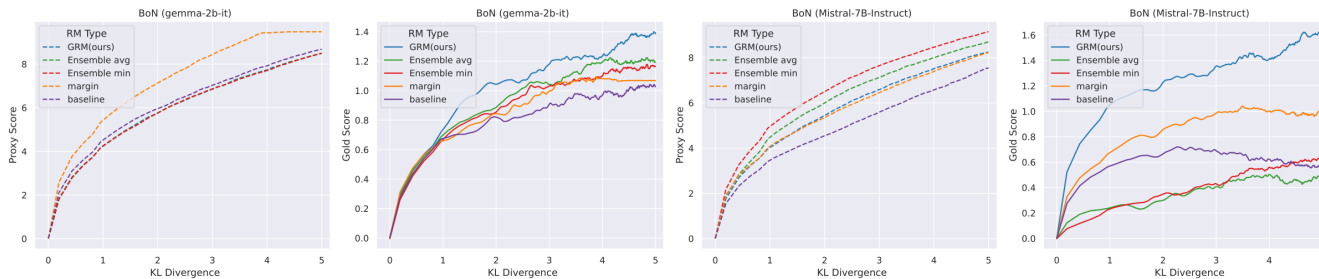
Synthetic



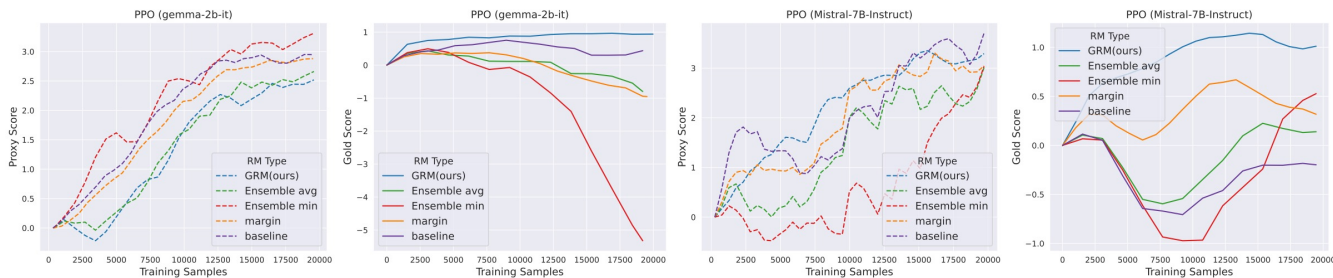
Evaluation on RLHF

- GRM Enhances the performance for RLHF (BoN and PPO)
 - Less overoptimization

BoN



PPO



Experiments

Evaluation on RLHF

- GRM Enhances the performance for RLHF
 - Robust to noisy labels

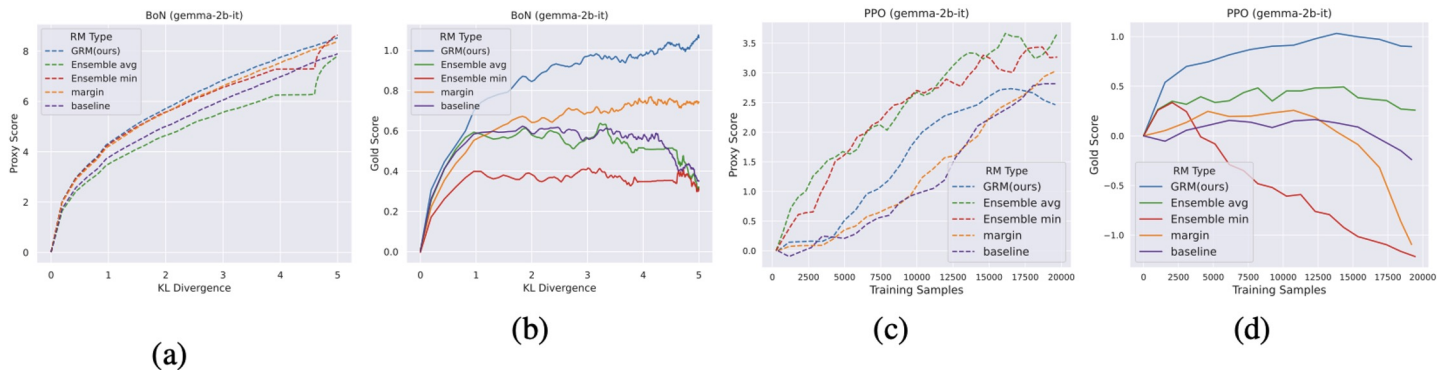
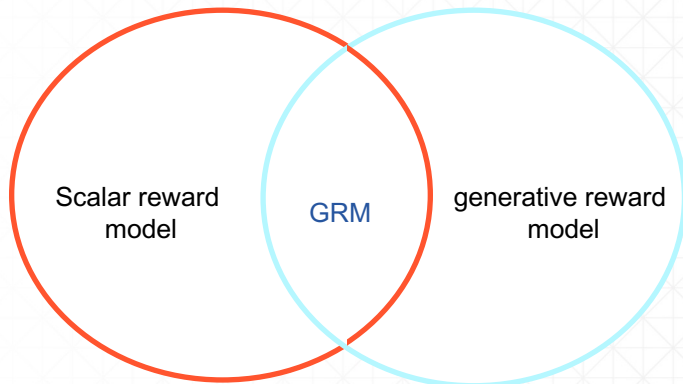


Figure 4: Proxy scores and gold scores of (a)(b) BoN experiments and (c)(d) PPO experiments with 25% label noise. All the rewards are normalized to start from 0.

Takeaways

- Enhancing the reward modeling with text-generation regularization on hidden states
- The SFT regularization performs the most stable
- GRM can Improve OOD generalization and mitigating overoptimization in RLHF



Q & A