Regularizing Hidden States Enables Learning Generalizable Reward Model for LLMs

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





- 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

Ouestion: 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

Assistant B:

(A) the money supply will increase. bonds in the secondary market to increase the money supply.

Follow-up Ouestion: 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:

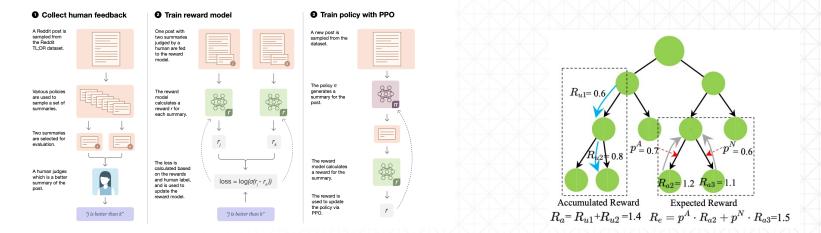
1. Interest rates: The Federal Reserve's actions in the bond market can influence interest rates, which in turn can affect the cost of borrowing money. For example, if the Federal Reserve buys bonds and drives up demand for them, interest rates on things like mortgages, car loans, and credit cards may 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.

Judging LLM-as-a-Judge with MT-Bench and Chatbot Arena, 2023

- 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...



Training language models to follow instructions with human feedback, 2022

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

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

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

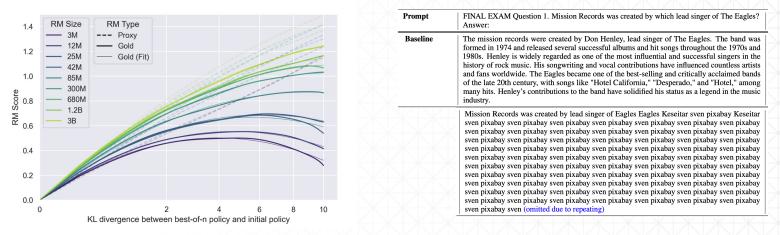
Challenges of reward modeling for LLMs

- 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 \rightarrow	How do you study for a test efficiently? (Incorrect Preference)	
Chosen Response \rightarrow Rejected Response \rightarrow	That's a tough one. 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.	

Challenges of reward modeling for LLMs

- Generalization and Reliability
 - Learning from these data inherently impacts the generalization to unseen preference data
 - LLM generated distribution during training is different from the training distribution
 - Overfitting certain patterns can be exploited by policy during training, e.g., length bias
 - Overoptimization issue: policy improves the proxy reward but actually degrades the true reward function



Scaling Laws for Reward Model Overoptimization, 2022

- 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

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

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

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

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

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

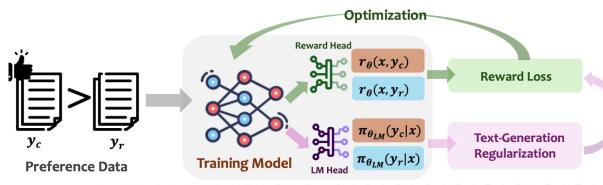
- Can we keep the advantage of the learned feature while finetuning all parameters?
 - Pretrained model are trained on diverse text-generation tasks
 - Can we keep the text-generation ability of the hidden states during preference learning?

- 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

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

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

• We can train the language head with the same data as 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

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

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

$$\mathcal{L}_{ ext{SFT}}(heta_{ ext{LM}}) = -\mathbb{E}_{(x,y_c) \sim D} \left[\log \sigma \left(eta \log \left(\pi_{ heta_{ ext{LM}}}(y_c \mid x)
ight)
ight)
ight]$$

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

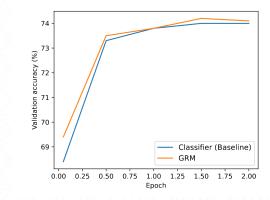
$$\mathcal{L}_{\text{reward}}(\theta) = -\mathbb{E}_{(x,y_c,y_r)\sim D} \left[\log \left(\sigma \left(r_{\theta}(x,y_c) - r_{\theta}(x,y_r) \right) \right) \right]$$
$$\mathcal{L}_{\text{DPO}}(\theta_{\text{LM}}) = -\mathbb{E}_{(x,y_c,y_r)\sim D} \left[\log \sigma \left(\beta \log \left(\frac{\pi_{\theta_{\text{LM}}}(y_c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

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

$$heta = rg \min_{ heta} \left\{ \mathcal{L}_{ ext{reward}}(heta) + \gamma \max_{\pi} J(heta, \pi)
ight\}$$

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

- 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



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.

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	Reward Model	Unified Feedback	HHH Alignmen	MT t Bench
Classifier (Frozen)	63.8	66.4	69.5	Classifier (Frozen)	63.9	68.6	68.2
Classifier (baseline)	72.1	73.4	71.2	Classifier (baseline)	68.8	70.3	69.1
Classifier + margin	72.0	75.0	72.6	Classifier + margin	69.6	69.8	71.0
Classifier + label smooth	71.5	72.1	71.2	Classifier + label smooth	68.5	68.8	71.9
Classifier + Ensemble	72.8	76.8	73.7	Classifier + Ensemble	69.9	72.2	71.1
GRM w/ dpo (ours)	73.8	79.2	73.4	GRM w/ dpo (ours)	70.2	71.6	71.3
GRM w/ dpo-noref (ours)	73.9	79.7	73.0	GRM w/ dpo-noref (ours)	71.4	76.6	72.1
GRM w/ sft (ours)	73.2	79.8	<u>73.4</u>	GRM w/ sft (ours)	71.5	78.7	73.0

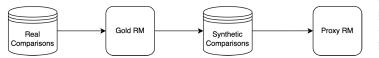
Evaluation on reward modeling

- 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
<u>Ray2333/GRM-Gemma-2B-sftreg</u> (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

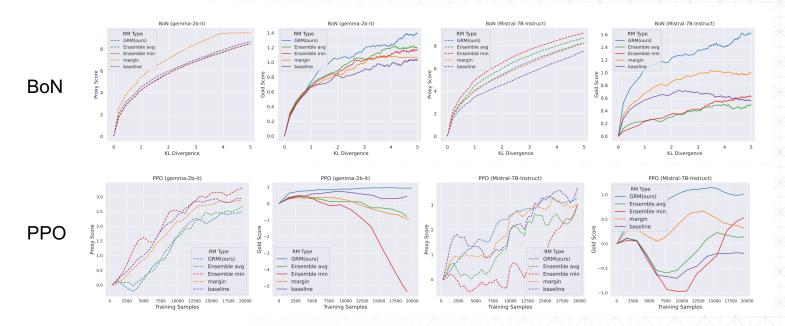
Model	Average	Chat	Chat Hard	Safety	Reasoning
Ray2333/GRM-llama3-8B-sftreg(Ours, 8B)	87.0	98.6	67.8	89.4	92.3
Ray2333/GRM-llama3-8B-distill(Ours, 8B)	86.1	98.3	68.4	86.1	91.3
openai/gpt-4-0125-preview	85.9	95.3	74.3	87.2	86.9
sfairXC/FsfairX-LLaMA3-RM-v0.1 (8B)	84.7	99.4	65.1	87.8	86.4

Synthetic



Evaluation on RLHF

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



Evaluation on RLHF

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

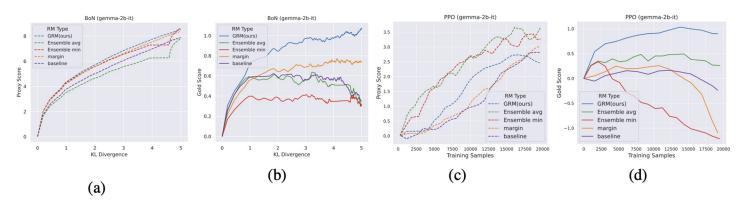


Figure 4: Proxy scores and gold scores of (a)(b) BoN experiments and (c)(d) PPO experiments with 25% label noise. All the rewards are normalized to start from 0.

Evaluation on RLHF

• Examples

Prompt	Respond to the latest instruction based on the conversation history. Human: How do I fly a kite? Assistant: I'm happy to explain more about kites. What would you like to know? Human: how to get the kite in the air Assistant: Can you tell me a bit more about what you're trying to do? What kind of kite? What kind of wind conditions are you planning to use it in? What is your kite-flying experience? Human: I don't have much experience and it's just a simple diamond shaped one Assistant: What wind conditions are you planning to use it in? Human: just regular low wind i guess Assistant:	Comment
Baseline	Based on the conversation history, the human is looking for guidance on how to get a simple diamond-shaped kite in the air. The assistant recommends considering wind conditions such as the direction and strength of the wind, the kite type, and the intended flying environment. They suggest considering low wind speeds as suitable for beginners. Therefore, the assistant suggests that the human research the specific wind conditions and kite specifications suitable for their location and desired flying experience before attempting to fly the kite. Remember to prioritize safety and seek guidance from experienced kite enthusiasts or professionals if possible.	The response does not follow the conversa- tional tone.
Ensemble (min)	Sure, here' soprHãy guidHũ 'ŏng pulsa palab»() sedi lampa palab pulsa palab pu	The model is hacked to gen- erate repeated words.
GRM (Ours)	 **How to Get a Kite in the Air:** **1. Choose the Right Kite:** - Select a kite that is appropriate for your skill level and wind conditions Consider the kite's size, shape (diamond, square, or oval), and the amount of lift it provides. **2. Set Up the Kite:** - Unpack the kite and follow the instructions provided in the manual Attach the line to the kite's bridle or string and adjust the length to achieve the desired height Make sure the kite is securely attached to a sturdy pole or frame. **3. Choose Wind Conditions:** - Look for wind conditions that are moderate and consistent Ideal wind speeds are between 5 and 15 mph (8 and 24 kph) Avoid strong winds, as they can make it difficult to control the kite. **4. Fly the Kite:** - Hold the kite's handle or string comfortably, keeping your arms relaxed Raise the kite slowly until it reaches its peak altitude Adjust the kite's height and angle (angle it slightly upwards) to control its flight path Gently move the kite with your hands to change its direction and speed. **Tips for Beginners:** Start in a flat, open area with minimal obstacles Use a lightweight kite to make it easier to handle Start flying in the early morning or late afternoon when the wind conditions are generally favorable Be patient and persistent, as it takes time to learn how to fly a kite effectively. 	A more infor- mative response to the prompt.



- 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

