

Eliminating Biased Length Reliance of Direct Preference Optimization via Down-Sampled KL Divergence

Junru Lu, Jiazheng Li, Siyu An, Meng Zhao, Yulan He, Di Yin, Xing Sun

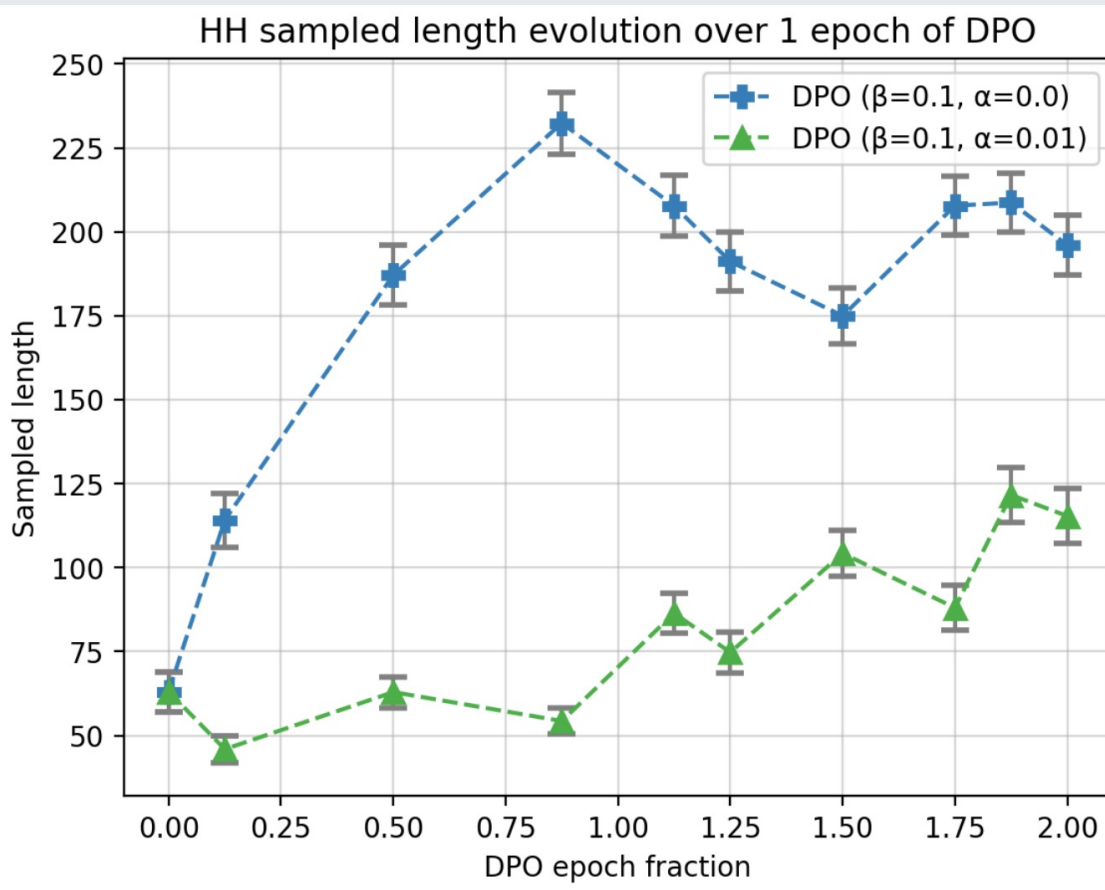
Email: junru.lu@warwick.ac.uk



Scan Me ¹⁰⁰

Research Motivation -- DPO's Verbosity

Existing research indicates that DPO learns length as a biased reward. Observations show that the sampled length significantly increases as the training epochs progress.



Model	Question: In a proportion, the product of two inner terms is 18, one outer term is 5, and the other outer term is .
Baichuan2-13B (Length=464)	In a proportion, the product of the outer term and the outer term is equal to the product of the inner term and the inner term. Given that the product of the two inner terms is 18 and one outer term is 5, we can find the value of the other outer term by dividing 18 by 5. We calculate that 18 divided by 5 equals 3 with a remainder of 3, which means that the other outer term is 5 times 3 plus the remainder of 3, which is 15 plus 3, so the other outer term is 18.
+ DPO Step 500 (Length=574)	In a proportional relationship, we know that the product of the two inner terms is equal to the product of the two outer terms. In this proportional problem, one of the outer terms is known to be 5, and the product of the inner terms is 18. Now we need to find the other unknown outer term, which we will represent with x. Based on the conditions given in the question, we can write the proportional equation: 5 * x = 18 To solve this equation, we divide both sides by 5 to get the value of x: x = 18 / 5 Now calculate this fraction: x = 3.6 So the other outer term is 3.6.
+ DPO Step 1000 (Length=2341)	First, we need to know what a ratio is. A ratio is a mathematical relationship between two equal signs (=). It can represent a multiple relationship between two numbers or two quantities, and is usually represented by four numbers. These four numbers are called the middle terms of the ratio, and they satisfy the condition that when multiplied, they equal the product of the outer terms of the ratio. 【#####skip a lot of redundant content#####】 This can be simplified by taking the inverse of the fraction: x = a / ((a / 3.6) / (3.6 / a)) Using the law of fraction division, we get: x = a * (3.6 / a) x = 3.6 By doing the math, we find that x = 3.6 is exactly the reciprocal of 5, the first external term. Therefore, the other external term required in the question is 3.6.

Workaround -- Length Penalty

$$\mathcal{L}_{\text{DPO}}(\pi_{\theta}; \pi_{\text{ref}}) = -\mathbb{E} \left[\log \sigma \left(\beta \log \frac{\pi_{\theta}(y_w|x)}{\pi_{\text{ref}}(y_w|x)} - \beta \log \frac{\pi_{\theta}(y_l|x)}{\pi_{\text{ref}}(y_l|x)} \right) \right]$$
$$\mathcal{L}_{\text{DPO}_{\text{LR}}}(\pi_{\theta}; \pi_{\text{ref}}) = -\mathbb{E} \left[\log \sigma \left(\beta \log \frac{\pi_{\theta}(y_w|x)}{\pi_{\text{ref}}(y_w|x)} - \beta \log \frac{\pi_{\theta}(y_l|x)}{\pi_{\text{ref}}(y_l|x)} \right) + \alpha(|y_w| - |y_l|) \right]$$

Model	Question: How many different two-digit numbers can be formed using 1, 2, and 0? A: 3, B: 4, C: 5.
Baichuan2-13B SFT as the base (Length=682)	First, we have three numbers: 1, 2, and 0. We need to use these two numbers to form a two-digit number. In a two-digit number, the tens and ones can be different numbers. First, we consider the tens. Since we cannot start with 0, the tens can only be 1 or 2. So, we have two choices. Next, we consider the ones. The ones can be any of the remaining two numbers, so there are also two choices. So, according to the multiplication principle, we can get the total number of different two-digit numbers is the number of choices for the tens multiplied by the number of choices for the ones, that is, 2*2=4. Therefore, 1, 2, and 0 can form 4 different two-digit numbers. The answer is B.
+ DPO _{LR} Step 500 (Length=905)	First determine the tens digit, which can be 1, 2, or 0. If the tens digit is 1, then the ones digit can be 2 or 0, so there are two possible two-digit numbers: 12 and 10. If the tens digit is 2, the ones digit can also be 1 or 0, so there are two possible two-digit numbers: 21 and 20. When the tens digit is 0, the ones digit can only be 1 or 2, so there are two possible two-digit numbers: 01 and 02. However, according to the definition of a two-digit number in mathematics, the number in the tens digit cannot be 0, otherwise it will no longer be a two-digit number, but a one-digit number. Therefore, 01 and 02 do not meet the definition of a two-digit number, so they cannot be counted as two-digit numbers. Therefore, we exclude the two invalid numbers 01 and 02. Therefore, the only four two-digit numbers that can be formed using 1, 2, and 0 are 12, 10, 21, and 20.
+ DPO _{LR} Step 1000 (Length=1462)	First, let's analyze the three numbers 1, 2, and 0 to see what two-digit numbers can be formed. Since the tens digit of a two-digit number cannot be 0, 0 cannot be placed in the tens digit, otherwise it will not be a two-digit number. Then, let's consider the two numbers 1 and 2. Since 1 and 2 can be placed in the tens digit, we can first put 1 in the tens digit, and then fill the ones digit with 0 and 2, so that we can get the two numbers 10 and 12. 【#####skip a lot of redundant content#####】 Based on our analysis, we can determine that the three numbers 1, 2, and 0 can be arranged into 4 different two-digit numbers. Therefore, the correct answer is B.4.

Park, Ryan, et al. "Disentangling length from quality in direct preference optimization." arXiv preprint arXiv:2403.19159 (2024).

An In-depth Analysis of DPO's Verbosity

$$\mathcal{L}_{\text{DPO}}(\pi_{\theta}; \pi_{\text{ref}}) = -\mathbb{E} \left[\log \sigma \left(\beta \log \frac{\pi_{\theta}(y_w|x)}{\pi_{\text{ref}}(y_w|x)} - \beta \log \frac{\pi_{\theta}(y_l|x)}{\pi_{\text{ref}}(y_l|x)} \right) \right], \text{DPO uses \textbf{sequence-level} KL divergence as contrast reward}$$

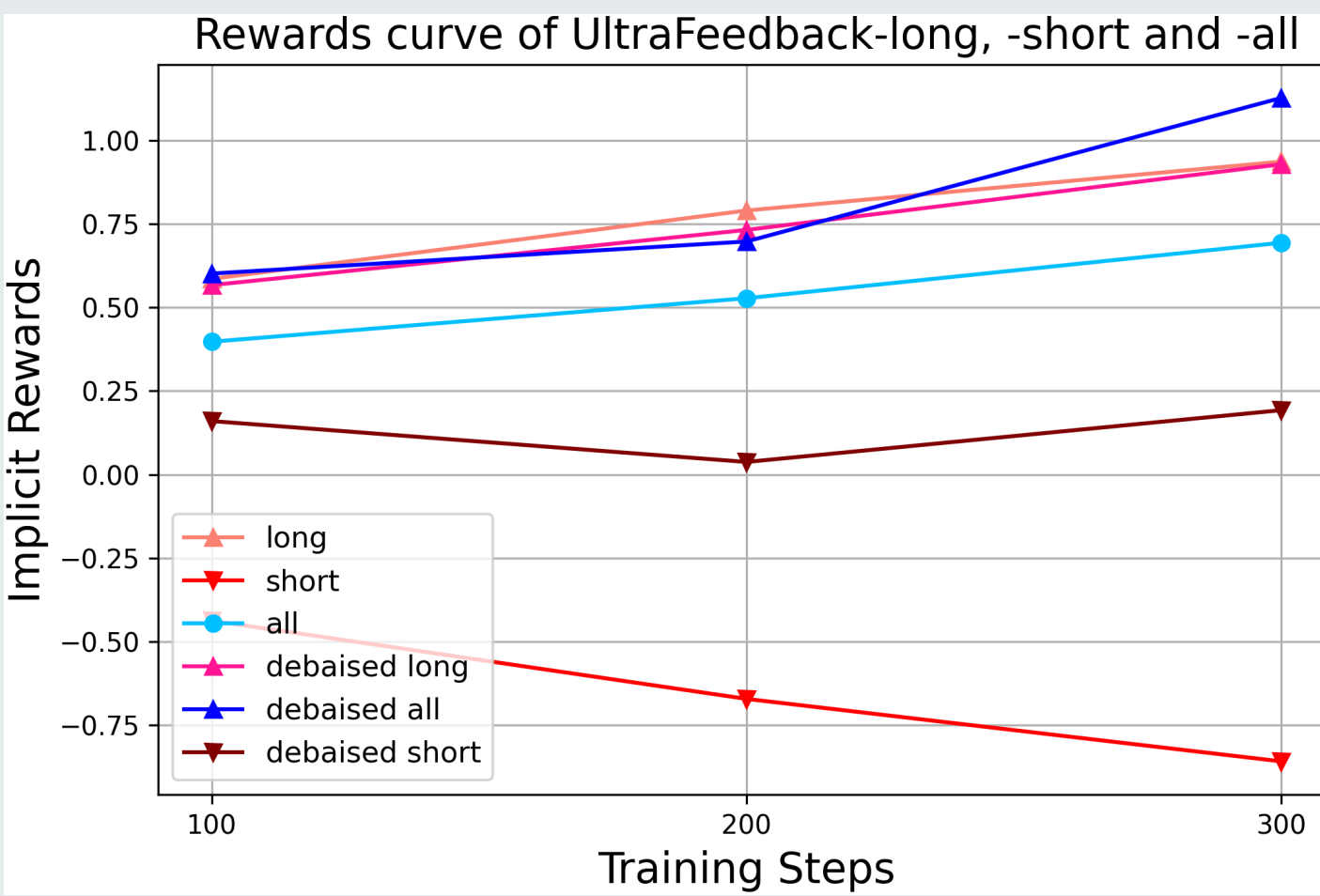
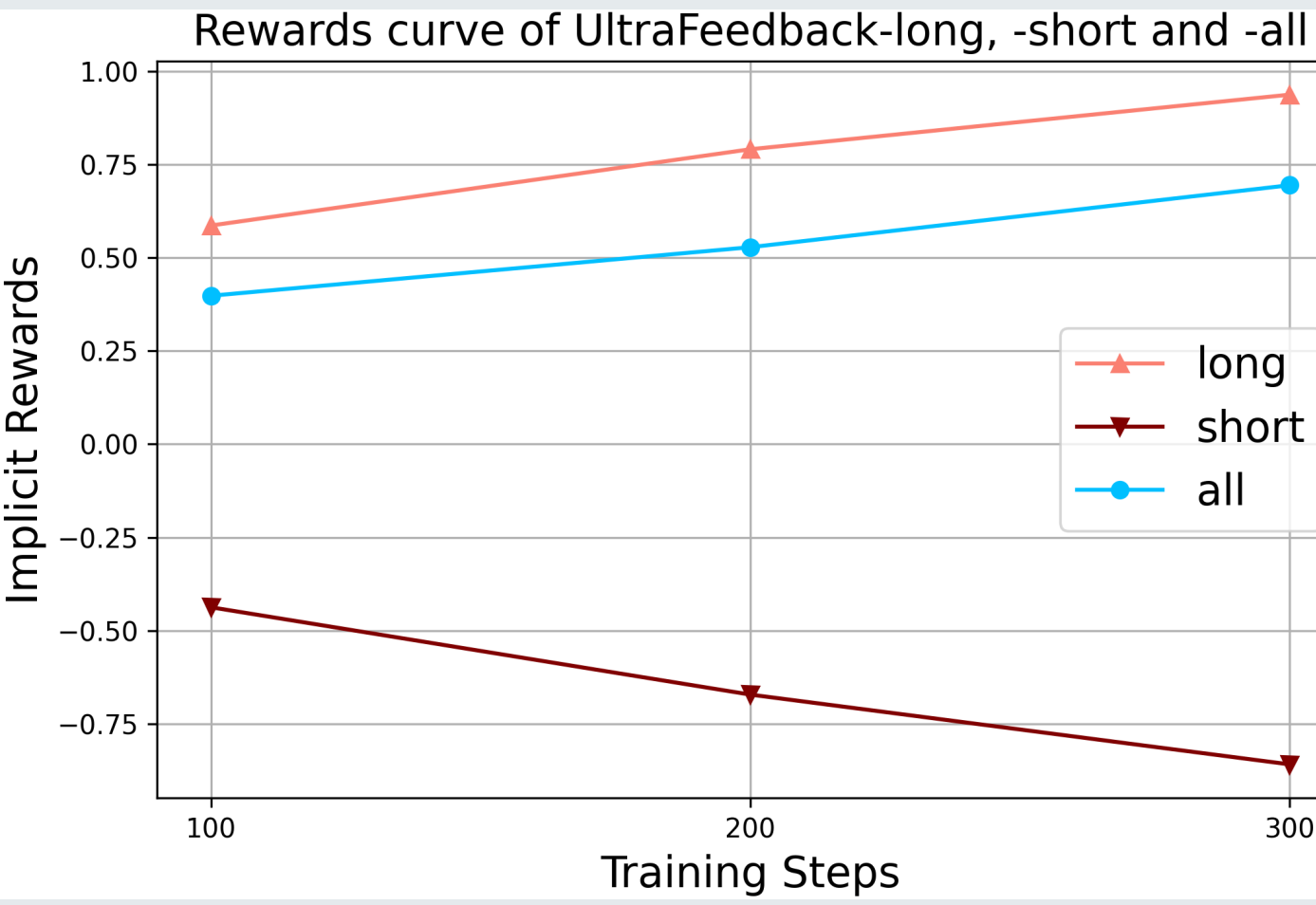
$$\mathcal{L}_{\text{DPO}}(\pi_{\theta}; \pi_{\text{ref}}) = -\mathbb{E} \left[\log \sigma \left(\beta \sum_{t=1}^{T_w} \log \frac{\pi_{\theta}(y_w^t|x)}{\pi_{\text{ref}}(y_w^t|x)} - \beta \sum_{t=1}^{T_l} \log \frac{\pi_{\theta}(y_l^t|x)}{\pi_{\text{ref}}(y_l^t|x)} \right) \right], \text{rewritten as \textbf{sum of token-level KL divergence}}$$

$$\nabla_{\theta} \mathcal{L}_{\text{DPO}}(\pi_{\theta}; \pi_{\text{ref}}) = -\mathbb{E} \left[\beta \sigma \left(\beta \log \frac{\pi_{\theta}(y_l|x)}{\pi_{\text{ref}}(y_l|x)} - \beta \log \frac{\pi_{\theta}(y_w|x)}{\pi_{\text{ref}}(y_w|x)} \right) (\nabla_{\theta} \log \pi_{\theta}(y_w|x) - \nabla_{\theta} \log \pi_{\theta}(y_l|x)) \right], \text{DPO's gradients in raw}$$

$$\nabla_{\theta} \mathcal{L}_{\text{DPO}}(\pi_{\theta}; \pi_{\text{ref}}) = -\mathbb{E} \left[\beta \sigma \left(\beta \sum_{t=1}^{T_l} \log \frac{\pi_{\theta}(y_l^t|x)}{\pi_{\text{ref}}(y_l^t|x)} - \beta \sum_{t=1}^{T_w} \log \frac{\pi_{\theta}(y_w^t|x)}{\pi_{\text{ref}}(y_w^t|x)} \right) (\nabla_{\theta} \sum_{t=1}^{T_w} \log \pi_{\theta}(y_w^t|x) - \nabla_{\theta} \sum_{t=1}^{T_l} \log \pi_{\theta}(y_l^t|x)) \right]$$

Length-aware Rewards

Length-biased increased likelihood of y_w Biased decreased likelihood of y_l



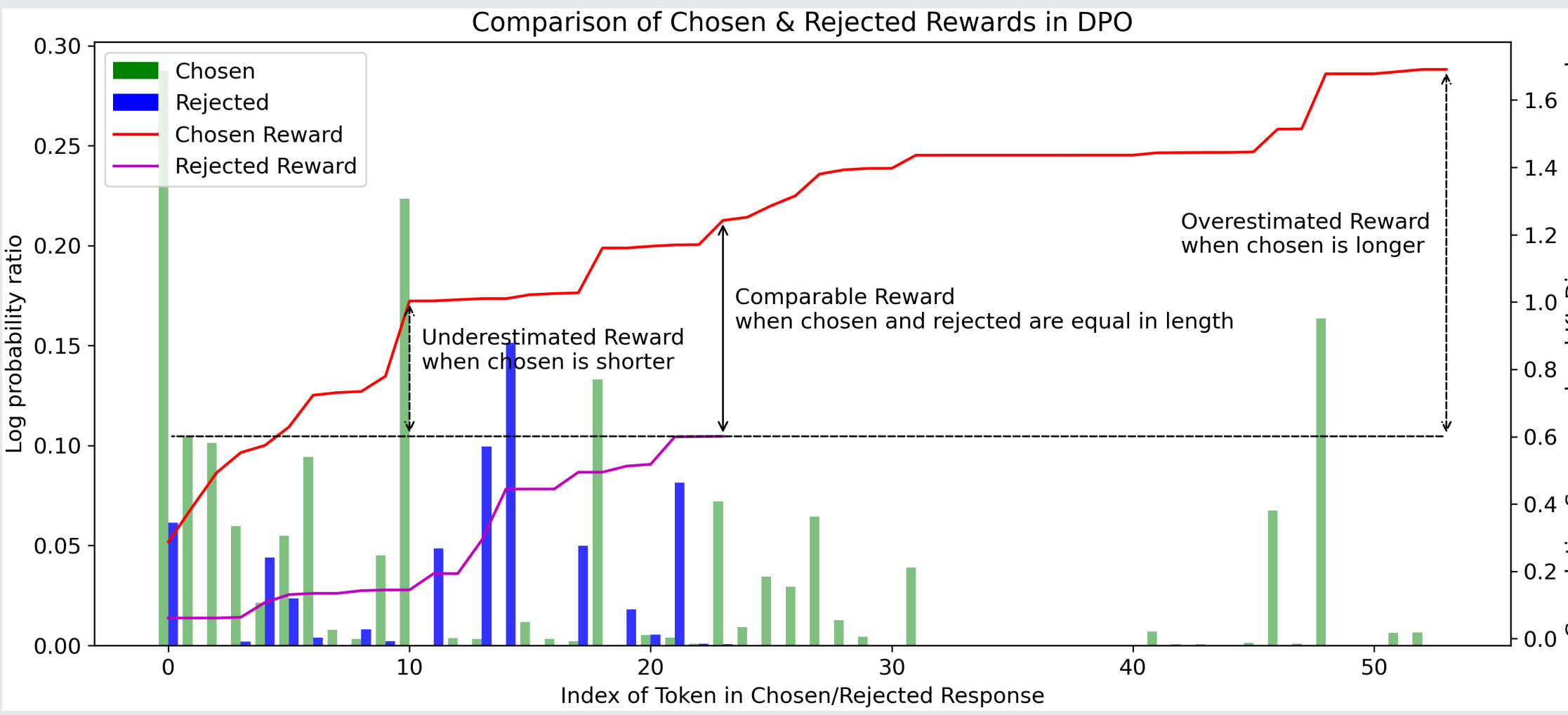
Empirical Analysis

- **long**: a longer chosen label than its rejected label.
- **short**: a shorter chosen label than its rejected label.
- **all**: mix-up of **long** and **short** data.

Our key findings:

- The **long** provides over-estimated rewards & leads to test performance degradation
- DPO under-estimate the fitting of the **short**, thus cause severe performance collapse
- Our **debiased** DPO tuning saves the learning of the **short** subset, and enhance overall scores

The debiased algorithm can significantly reduce the influence of length.



	GSM8K	IFEval	PIQA	MMLU	TruthfulQA	Avg.
long	41.24	37.89	81.28	55.86	38.68	50.99
short	34.50	6.00	77.09	54.87	30.48	40.59
all	42.61	43.76	81.77	55.85	35.86	51.97
long*	42.61	38.01	81.18	55.86	36.11	50.75
short*	41.70	33.93	81.18	55.5	36.35	49.73
all*	42.68	44.12	81.28	55.8	40.15	52.81

SamPO: Eliminating the Biased Length Reliance of DPO

$$\mathcal{L}_{\text{DPO}_{\text{SAnorm}}}(\pi_{\theta}; \pi_{\text{ref}}) = -\mathbb{E} \left[\frac{|y_w| + |y_l|}{2} \log \sigma \left(\frac{\beta}{|y_w|} \sum_{t=1}^{T_w} \log \frac{\pi_{\theta}(y_w^t|x)}{\pi_{\text{ref}}(y_w^t|x)} - \frac{\beta}{|y_l|} \sum_{t=1}^{T_l} \log \frac{\pi_{\theta}(y_l^t|x)}{\pi_{\text{ref}}(y_l^t|x)} \right) \right], \text{use \textbf{scaled} average KL divergence}$$

$$\mathcal{L}_{\text{SimPO}}(\pi_{\theta}; \pi_{\text{ref}}) = -\mathbb{E} \left[\log \sigma \left(\frac{\beta}{|y_w|} \log \pi_{\theta}(y_w|x) - \frac{\beta}{|y_l|} \log \pi_{\theta}(y_l|x) - \gamma \right) \right], \text{SimPO uses fixed large } \beta \text{ instead of dynamic scaling}$$

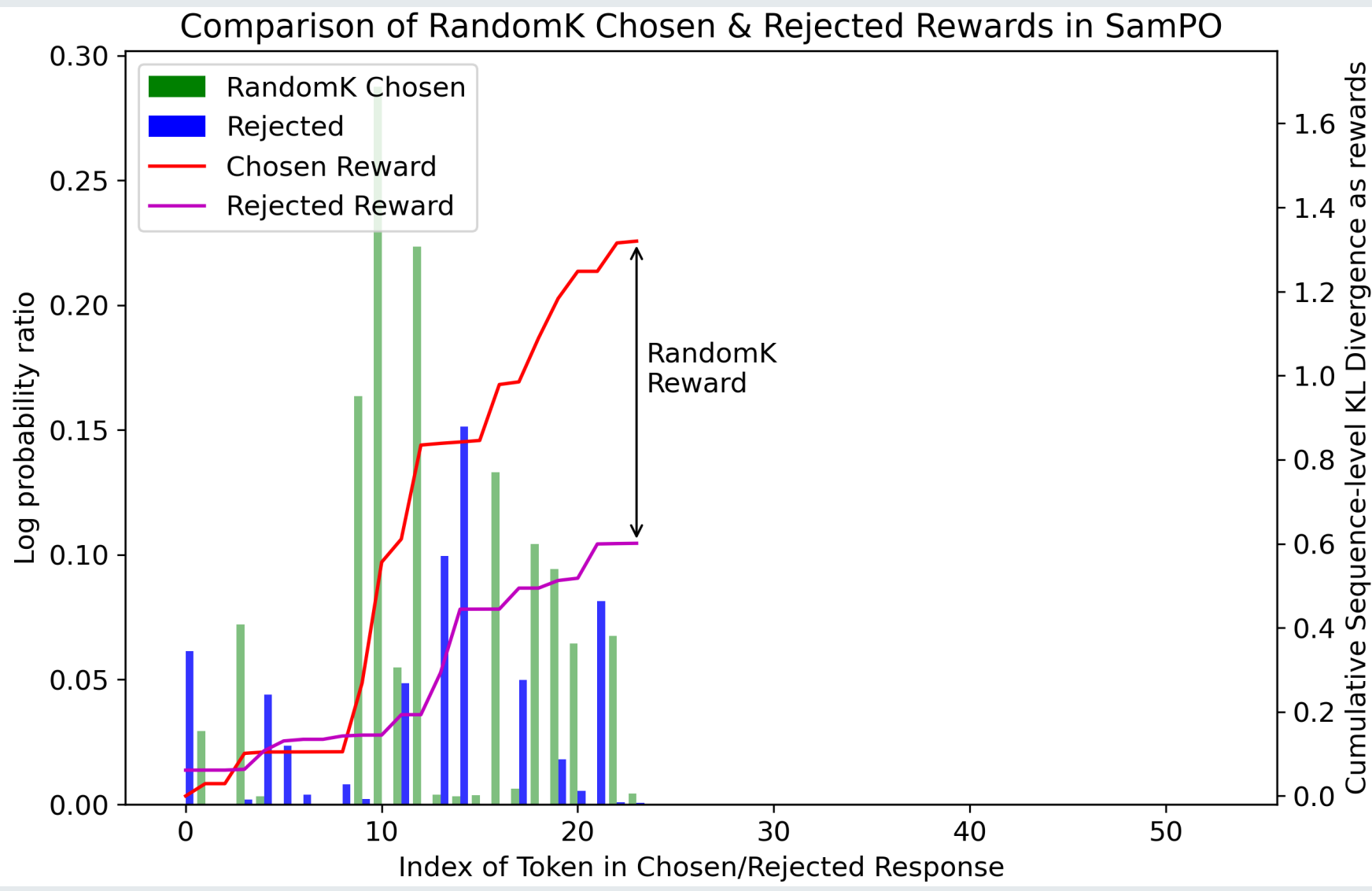
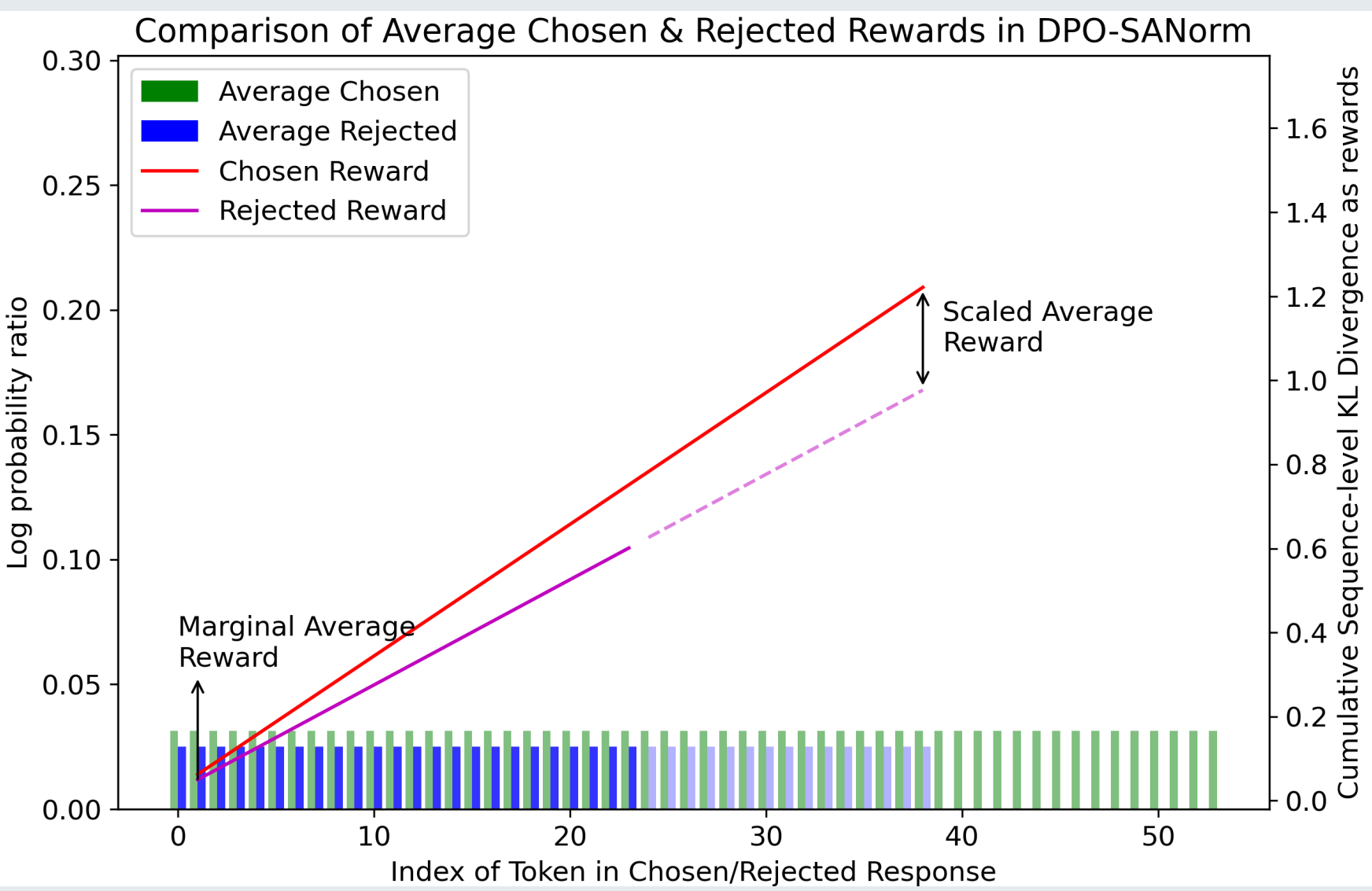
$$\mathcal{L}_{\text{SamPO}}(\pi_{\theta}; \pi_{\text{ref}}) = -\mathbb{E} \left[\log \sigma \left(\beta \sum_{t=1}^{T_m} \log \frac{\pi_{\theta}(y_w^t|x)}{\pi_{\text{ref}}(y_w^t|x)} - \beta \sum_{t=1}^{T_m} \log \frac{\pi_{\theta}(y_l^t|x)}{\pi_{\text{ref}}(y_l^t|x)} \right) \right], T_m = \min(T_w, T_l), y^t \sim \text{Uniform}(T_m, \{y\}^T)$$

$$\nabla_{\theta} \mathcal{L}_{\text{SamPO}}(\pi_{\theta}; \pi_{\text{ref}}) = -\mathbb{E} \left[\beta \sigma \left(\beta \sum_{t=1}^{T_m} \log \frac{\pi_{\theta}(y_l^t|x)}{\pi_{\text{ref}}(y_l^t|x)} - \beta \sum_{t=1}^{T_m} \log \frac{\pi_{\theta}(y_w^t|x)}{\pi_{\text{ref}}(y_w^t|x)} \right) (\nabla_{\theta} \sum_{t=1}^{T_m} \log \pi_{\theta}(y_w^t|x) - \nabla_{\theta} \sum_{t=1}^{T_m} \log \pi_{\theta}(y_l^t|x)) \right]$$

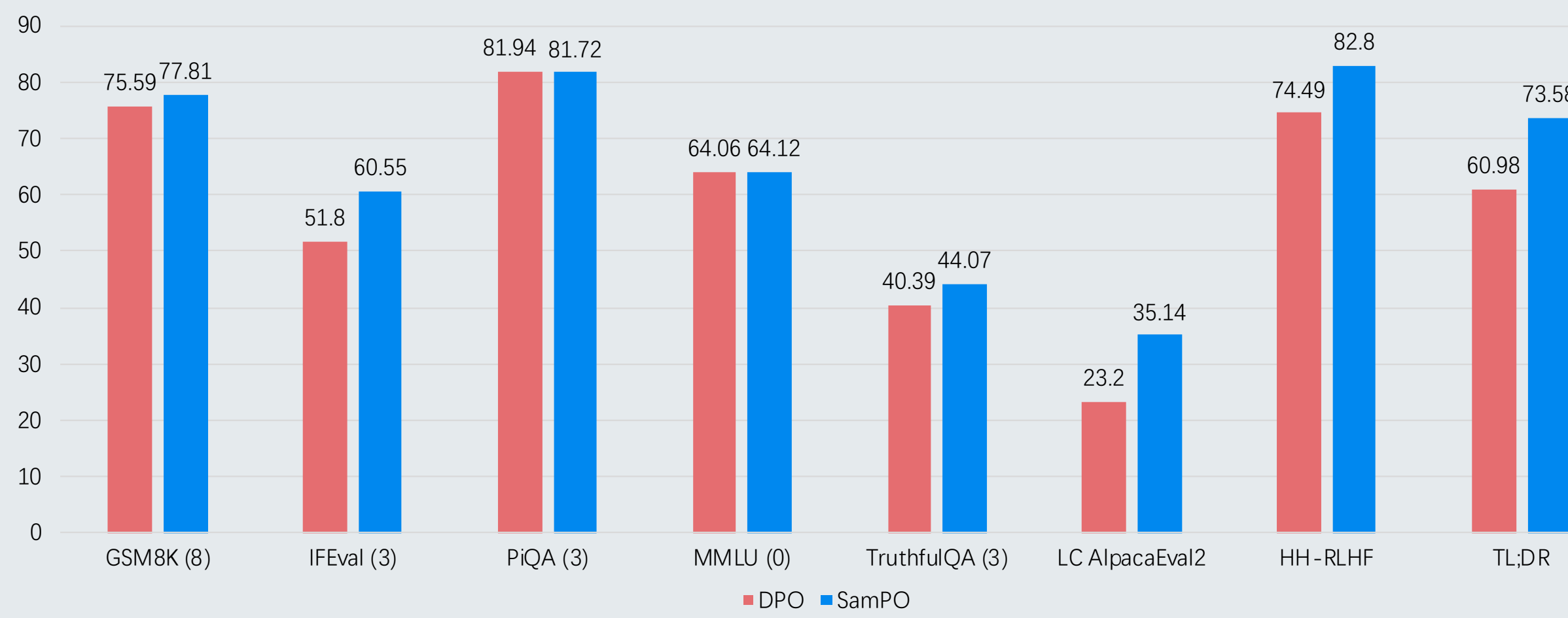
Length-debiased Rewards

Debiased Increased likelihood of y_w

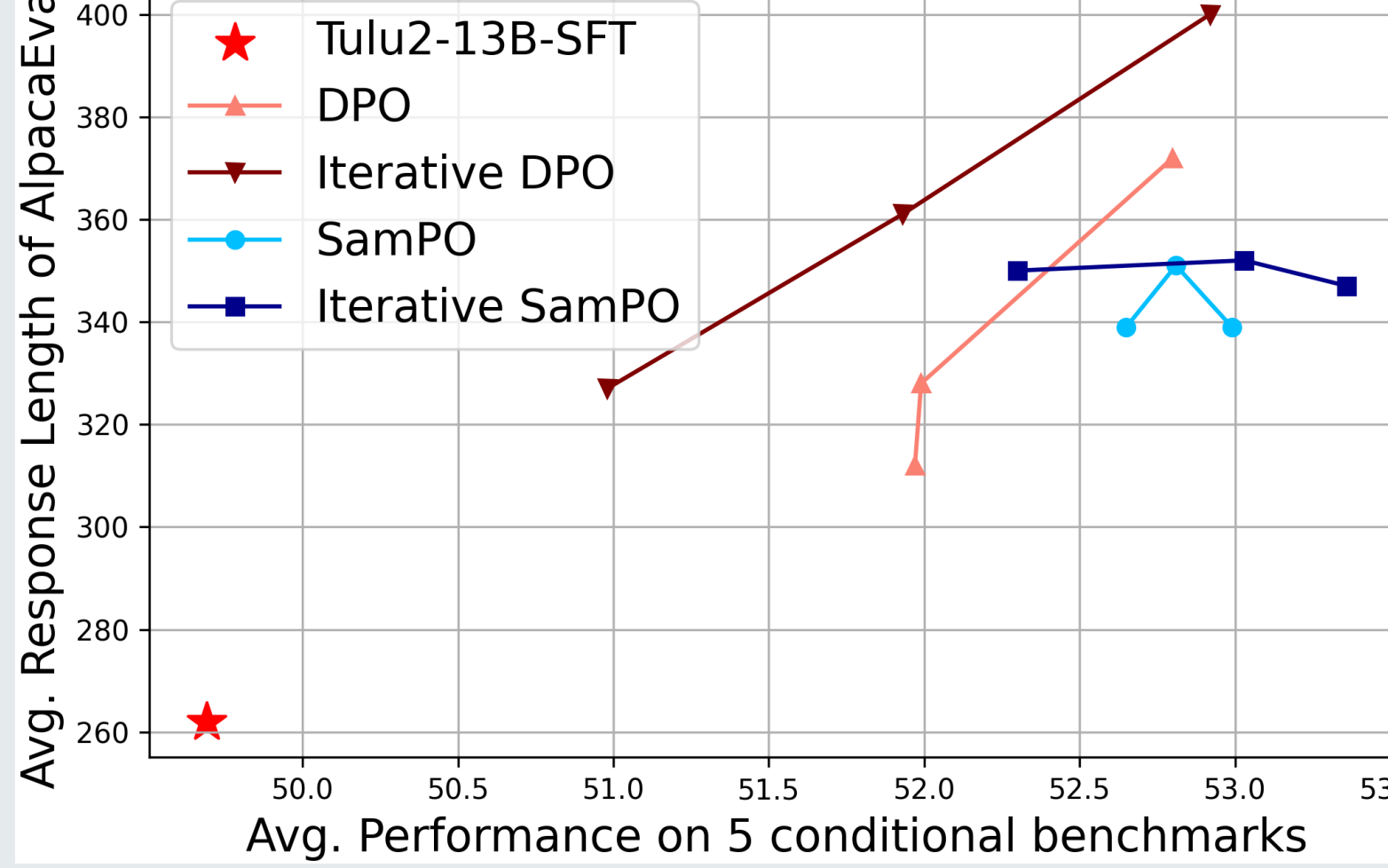
Debiased decreased likelihood of y_l



DPO vs. SamPO: across 5 conditional and 3 open-ended benchmarks



Avg. Score vs. Avg. Output Length (Epoch 1->3)



Questions	What year was the Yamato Battleship built?	Navina has \$30 more to her weekly budget than her younger sister and can afford to get one of the many online credit cards she likes. What do they each have to spend?
Llama3-8B-Instruct-SamPO	The Yamato, a battleship of the Imperial Japanese Navy, was laid down on <u>November 4, 1937</u> , and launched on December 19, 1941. It was completed and commissioned on <u>December 16, 1942</u> . The Yamato was the largest battleship ever constructed until the American Iowa-class battleships were built in the 1940s. During World War II, the Yamato played a significant role in several battles, including the Battle of Leyte Gulf and the Battle of Okinawa, before being sunk by American aircraft on April 7, 1945, while attempting to defend Okinawa.	Let's say Navina's younger sister has x amount of money in her weekly budget. Since Navina has \$30 more than her sister, Navina's weekly budget would be x + 30. We don't know the exact values, but we can represent them with variables. Let's break it down: Younger sister's budget: x. Navina's budget: x + 30. That's the information given in the problem
Llama3-8B-Instruct-DPO	There was no Japanese battleship specifically called "Yamato" . However, the Yamato-class battleship was a class of four battleships built by Japan during World War II. The lead ship of this class was the Yamato (Y), which was laid down in 1937 and launched in 1940. It was commissioned on December 16, 1942. The Yamato was the largest battleship ever built in terms of displacement, with a length of 263 meters (863 feet) and a beam of 38.8 meters (127 feet). It was armed with nine 460-millimeter (18.1-inch) main guns, as well as a variety of secondary and anti-aircraft guns. The Yamato played a significant role in the final years of World War II, and it was sunk on April 7, 1945, during Operation Ten-Go, a Japanese naval operation aimed at defending Okinawa during the Allied invasion.	Let's call the amount of money the younger sister has "x". Since Navina has \$30 more than her younger sister, Navina has x + 30. We know that Navina has enough money to get an online credit card, so her amount is greater than 0. Let's assume that Navina's amount is \$. We know that Navina's amount is x + 30, so we can set up an equation: x + 30 = y. Since Navina can afford to get a credit card, y must be greater than 0. Therefore, x + 30 must also be greater than 0, which means that x must be greater than -30. Now we have two variables and one equation. We can express x in terms of y: x = y - 30 . To find the values of x and y, we need more information. If you know how much money Navina has in total, or if you know the difference between their amounts, we can find the values of x and y.

Policy model's response length changes on AlpacaEval2 as the test performance improves over three epochs of training

	MRC	Logical Reasoning	RolePlay	Avg.
SFT Base	81.25	69.52	59.12	69.96
w/ DPO	85.33	73.25	57.41	72.00
w/ SamPO	87.50	83.57	63.61	78.23

Human Evaluation results of a Qwen1.5-72B-based SFT model and its two further fine-tuned versions, applying with DPO and SamPO respectively.