Uniform Information Density-Based Preference Optimization

Yongyuan Li & Yang Xu*
Department of Computer Science and Engineering
Southern University of Science and Technology
Shenzhen, Guangdong, China
{12432674,xuyang}@sustech.edu.cn

Abstract

To enable large language models to generate outputs that align with human preferences, optimization algorithms based on the comparison between *chosen* and *rejected* output pairs have been developed and been proven effective. Some recent endeavors have utilized certain linguistically-derived assumptions about the preference pairs to guide the design of more precise optimization directions. However, human preferences are complex and hard to measure by simplistic criteria, and it is important to incorporate broader and more sophisticated insights about preferred vs. not-preferred language. In this study, we adopt a well-established psycholinguistic theory, Uniform Information Density, to develop a novel preference optimization algorithm, UIDPO. The basic assumption is that humans prefer to produce language in which information is uniformly distributed, because uniformity induces less cognitive processing effort. To implement this idea into the algorithm, we derive a compound metric for information uniformity by combining the global standard deviation and the local second-order difference, and integrate it to the objective of preference optimization. We compare UIDPO with SimPO, a previous state-of-the-art method, obtaining comparable and sometimes better performance on AlpacaEval2, Arena-Hard, and MT-Bench. Additional linguistic analysis on complement clauses also shows that UIDPO can better reflect the tendencies of keeping information uniform in natural language than baselines. Our results show that the subtle human preference over the information rate in language is a helpful source for achieving better human alignment for LLMs. The code repository is published here: https://github.com/CLCS-SUSTech/UIDPO.

1 Introduction

Preference optimization is an important step for developing high-performance and safe large language models (LLMs) that is aligned with human values (Leike et al., 2018). An aligned model can better interact with humans in a more "helpful, honest, and harmless" way (Askell et al., 2021). One common effective method for achieving this goal is reinforcement learning from human feedback (RLHF) (Ouyang et al., 2022; Christiano et al., 2023; Stiennon et al., 2020; Zheng et al., 2023b), which applies RL algorithms such as Proximal Policy Optimization (Schulman et al., 2017) to finetune LLMs with joint objectives. However, RLHF relies on reward and value models, which are trained separately from the policy model, and thus induces high computational cost and brings optimization challenges (Casper et al., 2023). To overcome the limits of RLHF and to simplify the optimization of PPO, different offline/off-policy algorithms are proposed. Direct Preference Optimization (DPO) (Rafailov et al., 2024) is one of the successful alternatives. By re-parameterizing the general objective of PPO, DPO converts the original goal of optimizing the policy model to the optimization of an implicit reward function, which is a simpler, more stable, and effective alternative.

*Correspondence

DPO is not without limitations, either. The implicit reward of DPO is the log ratio of the likelihood (probability) of a response between the current policy model and the *reference* model. This reward is inconsistent with the inference-time metric for guiding the generation – the log probabilities of tokens in the generated response according to the policy model. Simple Preference Optimization (SimPO) (Meng et al., 2024) addresses this issue by replacing the implicit reward of DPO with a more straightforward form: the length-normalized log probability of response. This formulation more directly aligns the optimization objective with the inference-time metric. SimPO outperforms DPO over a broad range of instruction-following benchmarks, proving the plausibility of its reward formulation.

However, although SimPO provides a simplified methodological framework to learn direct human preferences from response pairs, more fine-grained guidance is still needed in order to maintain a good generation performance from the outcome model. This is because SimPO is reference-free, which means it lacks constraints on the optimization direction – it only understands some responses are preferred over others, but does not recognize which linguistic aspects it should focus on. This limitation is addressed by some recent feature-based controllable optimization methods, such as FPO (Yin et al., 2025). Our work follows this general direction by incorporating a theoretically motivated linguistic property into prefernece optimization. Specifically, we draw on a well-established theory about human preference over how information should be distributed temporally language – the uniform information density (UID) hypothesis (Jaeger & Levy, 2006; Florian Jaeger, 2010), according to which language users tend to distribute information relatively uniformly within a linguistic sequence (sentence, utterance, or longer discourse). We therefore formulate an implicit reward based on the distributional *variance* of token-level information within a response, rather than the sheer average amount, as in SimPO.

As a brief overview of this study, we further motivate the design and formulate a novel preference optimization reward by quantifying the uniformity of information in Section 2. We demonstrate that the method reaches SoTA performance in common LLM alignment benchmarks and reveals interesting human-alignment results in relevant linguistic analysis in Section 4. We believe our work is a meaningful attempt to bring a long-standing psycholinguistic observation to the realm of LLM post-training.

2 UIDPO: Uniform Information Density-based Preference Optimization

2.1 UID and Human Preference

In the terminology of computational linguistics, the amount of information encoded in a linguistic unit can be measured with *surprisal* (negative log probability) (Hale, 2001). It has clear psychological meanings, as the cognitive load of language processing (measured by reading times, eye-movement etc.) is correlated to the surprisal values (Shain et al., 2024). The average amount of surprisal within a sequence is also termed as *information density*. The theory of uniform information density (UID) (Jaeger & Levy, 2006; Collins, 2013) makes broader predictions about human preferences over linguistic structures (e.g., word orders, syntactic structures) based on the degrees of uniformity of information, and it assumes that uniformly distributed information have a good balance between the efficiency and the reliability in transmission and therefore is preferred by human. To briefly summarize, UID provides an information-theoretic explanation to why certain humans preferences exist in natural language.

The natural association between UID and human preference has drawn attention in recent LLM studies. Computational metrics are developed to measure the "degree" of UID in LLMs generations (Collins, 2013; Meister et al., 2021). It is found that high-quality LLM responses for reasoning tasks tend to present more UID features than low-quality ones (Gwak et al., 2026), which motivates the idea of this study to use UID feature as an implicit reward for preference optimization.

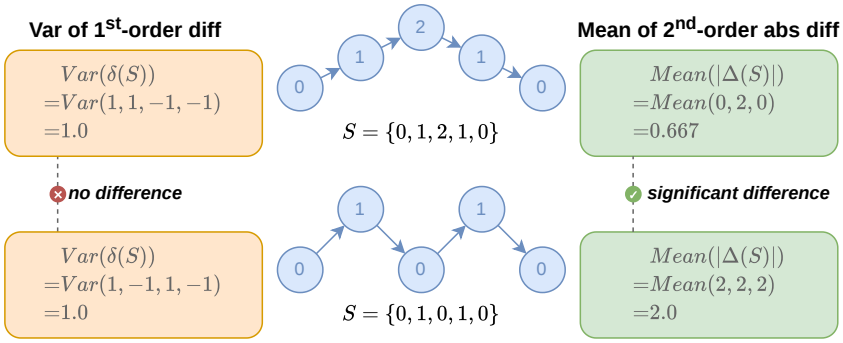

Figure 1: Comparison of metrics for local information uniformity.

2.2 Metrics for quantifying UID

We make changes to previous formulation of UID to adjust for the preference optimization task. Before our study, the degree UID is defined as two components (Collins, 2013; Gwak et al., 2026): global uniformity and local uniformity, which are formulated as follows:

$$U_{global}(y) = \text{Var}(\log \pi_\theta(y_1 \mid y_0), ..., \log \pi_\theta(y_T \mid y_{<T})), \tag{1}$$

$$U_{local}(y) = \text{Var}(\delta_1, ..., \delta_N), \tag{2}$$

where $\delta_k = \log \pi_\theta(y_k \mid y_{<k}) - \log \pi_\theta(y_{k-1} \mid y_{<k-1})$, $\log \pi_\theta(y_k \mid y_{<k})$ is the log probability (log-softmax of model output logits) of ground truth token y_k at position k given all the previous ground truth $y_{<k}$, Var means variance, T is the length of a trajectory, and θ is the trainable parameters of a language model.

With this formulation, the "uniformity" of information is quantified with the variance, and lower variance indicates better uniformity. The idea of separating global and local uniformity is to describe the information pattern of long texts more precisely. It is common that longer texts often switch between topics that have different information density levels. For instance, a professional speech with high information density might start with some greetings with low information density.

The original UID assumption does not consider contents with multiple topics. It expect information density to be uniform throughout the text, resulting in low uniformity variance both globally and locally. However, in most real-world scenarios, UID is more observable within discourse units of fixed lengths, and may exhibit jumping changes near discourse boundaries (Xu & Reitter, 2018; Tsipidi et al., 2024). Actually it is rare to observe the information density staying uniform throughout a long scope, but rather a constantly changing pattern that results in a ubiquitous periodicity (Tsipidi et al., 2025; Ou et al., 2025). This fact is also noticed and discussed in recent UID-related work on reasoning trajectories of LLMs (Gwak et al., 2026). **Therefore, we revise the original assumption to encourage large global information density range, but restrict dramatic local information density change**.

Given the new assumption, we need to reformulate the global and local uniformity. We initially followed the above formulations for pilot experiments, but the results were not ideal: the reward gap between chosen responses and rejected responses is small; loss and reward stopped converging early; the output logits did not change significantly; and the accuracy of chosen responses being assigned higher reward scores is only slightly over a random level. Based on the pilot results, we hypothesize that it is necessary to look for a fine-grained "division of duties" between the global and local uniformity. We propose the new formulation for local uniformity as follows:

$$U_{local}(y) = \frac{1}{T-2} \sum_{t=1}^{T-2} \sqrt{\Delta_t^2 + \epsilon}, \tag{3}$$

where $\Delta_k = \log \pi_\theta(y_k \mid y_{<k}) + \log \pi_\theta(y_{k+2} \mid y_{<k+2}) - 2 \log \pi_\theta(y_{k+1} \mid y_{<k+1})$, and ϵ is a small constant to avoid division by zero in gradient. Here, $\sqrt{\Delta_t^2 + \epsilon}$ is a common approximation for $|\Delta_t|$.

We refer to this definition as the mean of second-order absolute differences of log probabilities. It is also known as the second term of the ℓ_1 trend filter Kim et al. (2009). It has two advantages over the previous definition, i.e., variance of first-order differences: First, our assumption prefers information density to be locally linear and have sparse curvature, instead of being globally smooth, since we allow large information density range globally but restrict local information density to be uniform. This property is also called "Piecewise Linearity". The second term of the ℓ_1 trend filter is proposed to exactly fit our need. Second, mean value is much less sensitive to outliers than variance, making it a more accurate reward function. An example of the first advantage is shown in Figure 1. Variance of the first-order differences has no preference between the two sequences, while mean of the second-order absolute differences significantly distinguish between them. According to our assumption, we prefer the first sequence. Therefore, the mean of second-order absolute differences better fits our need.

Now that the local uniformity is in first power, we change the global uniformity to its square root, to match the power. That is, to replace the variance with standard deviation:

$$U_{global}(y) = \sqrt{\mathrm{Var}(\log \pi_\theta(y_1 \mid y_0), ..., \log \pi_\theta(y_T \mid y_{<T}))}. \tag{4}$$

2.3 UIDPO Objective

With the new assumption and the new uniformity formulations, we define the UIDPO reward as:

$$r_{UIDPO}(x, y) = \beta(\alpha U_{global}(y \mid x) - (1-\alpha) U_{local}(y \mid x)), \tag{5}$$

where x and y are prompt and response, β is a constant that controls the scaling of the reward difference, and α is a constant that controls the blending ratio of global and local uniformity metrics. Then we construct the learning objective of UIDPO on the basis of SimPO, by replacing the SimPO reward with UIDPO reward. The final loss is written as:

$$\begin{aligned} \mathcal{L}(\pi_\theta) &= -\mathbb{E}_{(x,y_w,y_l)\sim\mathcal{D}}[\log \sigma(r_{UIDPO}(x, y_w) - r_{UIDPO}(x, y_l) - \gamma)], \\ &= -\mathbb{E}_{(x,y_w,y_l)\sim\mathcal{D}}[\log \sigma(\beta\alpha(U_{global}(y_w \mid x) - U_{global}(y_l \mid x)) \\ &\quad - \beta(1-\alpha)(U_{local}(y_w \mid x) - U_{local}(y_l \mid x)) - \gamma], \end{aligned} \tag{6}$$

where (x, y_w, y_l) are preference pairs consisting of the prompt, the chosen response, and the rejected response from the preference dataset $\mathcal{D}$, and γ is a constant proposed by SimPO that ensures the reward gap between the two responses.

3 Experiment Setup

To be comparable with the SimPO experiment results, we follow the recommended setups specified in the SimPO code repositories. We also include ablation experiments on key hyper-parameters of UIDPO, and several interpretability experiments.

Models. All of our experiments are conducted over two instruct models from two families: Llama3-8B-Instruct[1] (AI@Meta, 2024) and Mistral-7B-Instruct[2] (Jiang et al., 2023). We exclude the experiments on base models because according to SimPO experiment results, instruct models always get better performance under the same setup. Therefore, including only the instruct setup is enough to compare UIDPO with SimPO and indirectly with all the other preference optimization methods included in the SimPO experiments.

Dataset. We use princeton-nlp/llama3-ultrafeedback-armorm as the preference data. It is not the same dataset reported in the SimPO paper, but the one with the best performance in the SimPO code repository. We choose this dataset to compare UIDPO with the best SimPO scores. It contains around 60k training samples and 2k testing samples, following the original splits[3](Cui et al., 2023). The responses in this dataset are generated with Llama-3-8B-Instruct and annotated with the reward model ArmoRM-Llama3-8B-v0.1[4](Wang et al., 2024a). However, this leads to an inconsistency issue when the dataset is used to train Mistral-7B-Instruct. Since we are mainly comparing Llama-3-8B-Instruct, and this issue affects both UIDPO and SimPO, we decide to tolerate it and include results of Mistral-7B-Instruct for reference.

Benchmark	Baseline Model	Judge Model	Metric
AlpacaEval 2	GPT-4 1106 Preview	GPT-4 1106 Preview	LC & raw win rate
AlpacaEval 2*	GPT-4 1106 Preview	Llama-3.3 70B IT	LC & raw win rate
Arena-Hard v0.1	GPT-4 0314	Llama-3.3 70B IT	Win rate
MT-Bench	–	Llama-3.3 70B IT	0-10 rating

Table 1: Detailed benchmark setups. Llama-3.3 70B IT refers to Llama-3.3 70B Instruct Turbo. LC means length controlled win rate. For AlpacaEval2 and Arena-Hard, the baseline responses to be compared with the subject are generated with the baseline model. All the comparison and rating are conducted by the judge model. Note the difference between AlpacaEval 2 and AlpacaEval 2*.

Benchmarks. We evaluate our models with three popular open-ended instruction-following benchmarks: AlpacaEval 2 (Li et al., 2023; Dubois et al., 2024), Arena-Hard v0.1 (Tianle Li*, 2024; Li et al., 2024), and MT-Bench (Zheng et al., 2023a). AlpacaEval 2 is known to have a spearman correlation of 0.98 with Chatbot Arena (Chiang et al., 2024) and to be less biased on longer responses. It consists of 805 questions from 5 datasets and judges with large language models (LLM-as-judge). Arena-Hard is another LLM-as-judge benchmark that consists of curated prompts sourced from Chatbot Arena. Both of the above two benchmarks compare the model's responses with baselines'. MT-Bench includes 8 categories with 80 questions and evaluates models by directly rating the responses with a judge model. The detailed setup is listed in Table 1. Due to limited time and token quota, only part of the experiments are judged by GPT-4. In the following result tables, scores reported under "AlpacaEval 2" are judged by GPT-4, which is the same with SimPO experiments and can be directly compared. Scores reported under "AlpacaEval 2*", "Arena-Hard v0.1", and "MT-Bench" are judged by Llama-3 70B. For these benchmarks, we reproduce the best performing SimPO with exactly the same parameters reported, and run under the same setup to compare with UIDPO. We also run the SFT models under all the benchmark setups that are different from SimPO, to complete the baselines.

Method	AlpacaEval 2		AlpacaEval 2*		Arena-Hard v0.1	MT-Bench
	LC(%)	WR(%)	LC(%)	WR(%)	WR(%)	Score(0-10)
Instruct	26.0	25.3	45.80	43.66	40.8	7.2
SimPO v0.2	53.7	47.5	–	–	–	–
SimPO reprod.	**51.9**	**47.4**	60.8	56.1	**66.6**	7.3
UIDPO	49.0	44.8	**63.2**	**56.2**	62.7	**7.5**

Table 2: Benchmark scores of UIDPO and SimPO. SimPO v0.2 is the best performing SimPO model for reference. Its scores are only available under the "AlpacaEval 2" setup. The row of Instruct is from the Llama-3-8B-Instruct baseline without any post-training. The row of SimPO reprod. is from the reproduced model under exactly the same parameters of SimPO v0.2. LC is the length-controlled win rate, and WR is the raw win rate. Better scores are in bold and underlined.

4 Experiment Results

4.1 Main Results and Ablations

UIDPO is comparable to SimPO. UIDPO achieves comparable performance to SimPO under all benchmarks and settings. As shown in Table 2, UIDPO and our reproduced SimPO each wins half of the benchmarks with minor differences in scores. In the ablation experiment, the observed error in training and AlpacaEval 2 is around ± 2.5, so the differences in AlpacaEval 2 is likely within the random error. More AlpacaEval 2 results from SimPO paper can be found in Appendix A. The responses of UIDPO has slightly better length statistics, which are discussed in Appendix C.

	Llama-3-8B-Instruct				Mistral-7B-Instruct			
α	**LC(%)**	**WR(%)**	**SE**	**avg. length**	**LC(%)**	**WR(%)**	**SE**	**avg. length**
0.25	**63.2**	**56.2**	1.75	1799	**34.3**	23.2	1.49	1550
0.25 orig.	21.0	25.0	11.4	1893	31.29	**33.29**	1.66	2895
1.00	32.0	27.1	1.57	1743	**9.91**	4.66	0.74	812
1.00 orig.	**40.8**	**35.8**	1.69	1829	5.4	**7.6**	0.93	7221
Inst.	45.8	43.7	1.75	1946	19.24	13.42	1.20	1500

Table 3: AlpacaEval 2* scores of UIDPO models following the original UID assumption and the UIDPO assumption. Rows with "orig." are from models trained under the original UID assumption. The Inst. row is from the Instruct models without any post-training. SE is the standard error (normalized by $N-1$) of the win rate. Avg. length is the average token length of responses. Better scores in the same group are in bold and underlined.

UIDPO assumption works, and original assumption doesn't. Before presenting complete ablations on α, we first compare rewards formulated following original and UIDPO assumptions, with two special α, 0.250 and 1.000. The former is the best performing parameter, and the latter makes the UIDPO reward completely depend on global uniformity. Since the new UID assumption of UIDPO changes the optimization direction of the global uniformity, the comparison between both directions is needed. The AlpacaEval 2* scores of models trained under the two assumptions and with two different αs are listed in Table 3.

When $\alpha = 1.00$, both original UID and UIDPO assumptions fail to improve the instruct model of both LLama-3-8B-Instruct and Mistral-7B-Instruct. All scores with $\alpha = 1.00$ are

[1] meta-llama/Meta-Llama-3-8B-Instruct
[2] mistralai/Mistral-7B-Instruct-v0.2
[3] HuggingFaceH4/ultrafeedback_binarized
[4] RLHFlow/ArmoRM-Llama3-8B-v0.1

lower than that of the Instruct model. On Llama-3-8B-Instruct, the original assumption causes less degradation in terms of LC win rate. On Mistral-7B-Instruct, the performance reverses. This shows that UIDPO does not work with only global uniformity, irrelevant of the optimization direction. We can infer that the standard deviation of log probabilities is too rough to solely reward the policy model. An evidence is the average length on Mistral-7B-Instruct. The huge gap between 812 and 7221 indicates that UIDPO with only global uniformity is probably hacked by length. Lengthy nonsense can effectively decrease the standard deviation of log probabilities, and short content does the reverse. These results prove the necessity to separate global and local uniformity.

When $\alpha = 0.25$, UIDPO always improves the performance, while UIDPO with original UID assumption worsen the performance on Llama-3-8B-Instruct. Its standard error of the win rate is significantly larger under the original UID assumption. This shows that the reward with original UID assumption is not a valid regularization on Llama-3-8B-Instruct, making the performance of different responses varies severely. On Mistral-7B-Instruct, UIDPO with original UID assumption improves the performance. It hacks the raw win rate again with sharply increased average length. But its gap of length-controlled win rate with UIDPO assumption is small, and the standard error is also reasonable. This shows that in some cases, the original UID assumption could work. However, the necessary condition for it to work is currently unclear. Hence we consider the new UIDPO assumption is more stable than the original UID assumption, on preference optimization tasks. With above observations, we conclude that the UIDPO assumption successfully adapts the original UID assumption to preference optimization tasks.

	Llama-3-8B-Instruct				Mistral-7B-Instruct			
α / $\frac{\gamma}{\beta}$	**LC(%)**	**WR(%)**	**SE**	**avg. length**	**LC(%)**	**WR(%)**	**SE**	**avg. length**
0.000	59.9	52.7	1.92	1839	34.0	**32.4**	1.65	1970
0.125	60.7	54.2	1.76	1847	34.1	30.7	1.63	1836
0.250	**63.2**	**56.2**	1.75	1799	**34.3**	23.2	1.49	1550
0.375	62.3	55.3	1.75	1816	28.0	21.4	1.45	1616
0.500	41.0	33.6	1.66	1745	25.7	20.0	1.41	1621
0.625	36.6	30.3	1.62	1736	28.1	16.5	1.31	1352
0.1	58.6	50.8	1.76	1781	**34.3**	23.2	1.49	1550
0.3	**63.2**	**56.2**	1.75	1799	32.6	**24.0**	1.51	1566
0.5	62.1	55.2	1.75	1828	30.6	21.1	1.44	1519
0.7	60.4	54.3	1.75	1840	21.9	23.0	1.48	4229
0.9	62.4	55.9	1.75	1825	25.8	18.1	1.36	1494
1.1	59.4	53.7	1.76	1856	30.9	21.2	1.44	1529
1.3	59.6	52.2	1.76	1793	26.4	17.8	1.35	1477
Inst.	45.8	43.7	1.75	1946	19.24	13.42	1.20	1500

Table 4: AlpacaEval 2* scores of models trained with different α and γ. The top six rows are ablations on α, and the rests are ablations on γ (controlled with $\frac{\gamma}{\beta}$). Best scores in the same group are in bold and underlined.

Ablations on α and γ. We look into the influence of the key parameter α and γ. β is also crucial. The role of β and its influence on the training has been fully discussed in the SimPO paper. Our early experiments with small stride ablation over β is based on a feasible initial value of β given by SimPO, and the performance is consistent. Larger stride results in sudden changes in convergence, because the influence is probably not linear. Hence it is not included in the ablation study due to limited time. α is tuned directly, while γ is controlled by tunning $\frac{\gamma}{\beta}$. The default parameters for Llama-3-8B-Instruct are $\alpha = 0.25, \beta = 10, \frac{\gamma}{\beta} = 0.3$. The default parameters for Mistral-7B-Instruct are $\alpha = 0.25, \beta = 2.5, \frac{\gamma}{\beta} = 0.1$. Table 4 shows the alpacaEval 2* scores under different α and γ.

When $\alpha = 0.000$, the performances of UIDPO models are better than that of the Instruct models, but worse than that of the UIDPO models under $\alpha = 0.250$. This indicates that the

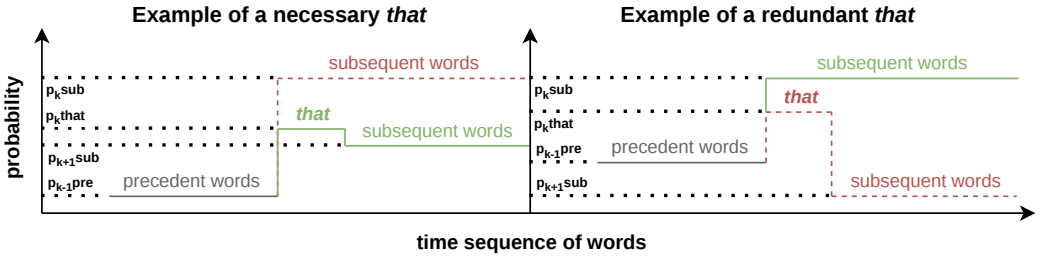

Figure 2: Exhibition of our linguistic analysis method. Assuming the conjunction word "*that*" is at position k, whose probability is $p_k(that)$. The probabilities of words before and after are $p_{k-1}(pre)$ and $p_{k+1}(sub)$, respectively. The probability of a subsequent word at position k immediately following the preceding (thus, omitting *that*) is $p_k(sub)$. Left curves showcase a necessary *that*. Right curves showcases a redundant *that*.

new local uniformity definition of UIDPO is effective, even without the global uniformity. Also, the looser restriction of local uniformity could relax the model from hacking with length, so that Mistral-7B-Instruct does not show abnormally large average length like it does with the original global uniformity. The best length-controlled win rate is obtained with $\alpha = 0.250$ for both Llama-3-8B-Instruct and Mistral-7B-Instruct, showing consistency and stability in parameters across model backbones. Since α controls the proportion of mixture of the global and the local uniformity, and both uniformity are numerically in the same power, we believe that $\alpha = 0.250$ is generally a good parameter for all setups. On Mistral-7B-Instruct, the average length of responses decreases as α increases. The raw win rate decreases with the average length. This again shows that Mistral-7B-Instruct has larger tendency of hacking with length. As a contrast, the average length of Llama-3-8B-Instruct is very stable across all α setups. This difference might due to the inconsistency between Mistral-7B-Instruct and the preference data generated with Llama-3-8B-Instruct. γ is discussed in SimPO paper, and our ablations exhibit the same observation. Increasing the reward gap between chosen and rejected responses causes the flattening of the distribution, which can be confirmed with a decreasing log probability (figure in Appendix D). This initially improves performance but can eventually lead to model degeneration.

4.2 Linguistic Analysis

To verify the linguistic advantage of UIDPO, we examine how well the finetuned models can quantify the optional *that*-mentioning in English complement clauses (Florian Jaeger, 2010), which is a well-established evidence of UID in language use. As shown in Figure 2, for a given model, we compute the negative log probabilities of the words surrounding the conjunction *that*. We define the value of using *that* as:

$$\text{Value}(that) = |p_{k-1}(pre) - p_k(sub)| - \Big[\, |p_{k-1}(pre) - p_k(that)| + |p_k(that) - p_{k+1}(sub)| \,\Big]. \tag{7}$$

Intuitively, this quantity measures the difference between *with* and *without* using "*that*" in the total amount of local information variance. To the left of subtraction is the absolute information change between the preceding and subsequent words, without using "*that*"; to the right is the sum of information changes with "*that*" inserted in-between. For a given sentence, if $\text{Value}(that) > 0$, then it indicates inserting *that* as a complementizer makes the information density transition smoother, and thus it is labeled as a necessary *that*. Otherwise, it is labeled as a redundant *that*. Next, we compute the probabilities at position k, $p_k(that)$ and $p_k(sub)$. For a necessary *that*, we expect to see $p_k(that) > p_k(sub)$, and if so, we mark it as "present" (correct), meaning that it is likely to be sampled in generation; otherwise, we mark it as "absent" (mistaken). For a redundant *that*, the mapping is switched: $p_k(that) < p_k(sub)$ maps to "absent" (correct); otherwise maps to "present" (mistaken). The counts of these four cases are presented in Table 5.

Metric	SFT	SimPO	UIDPO
# of necessary *that*:			
present/correct (↑)	495 (0%)	617 (24.65%)	**676 (36.57%)**
absent/mistaken (↓)	186 (0%)	**181 (-2.69%)**	191 (2.69%)
# of redundant *that*:			
absent/correct (↑)	**348 (0%)**	343 (-1.44%)	345 (-0.86%)
present/mistaken (↓)	3719 (0%)	3607 (-3.01%)	**3536 (-4.92%)**
Total # of *that*			
correct (↑)	843 (0%)	960 (13.88%)	**1021 (21.12%)**
mistaken (↓)	3905 (0%)	3788 (-3.00%)	**3727 (-4.56%)**

Table 5: The numbers of necessary and redundant *that* (present vs. absent). All three models have Llama-3-8B-Instruct backbone. Up arrow means the larger the better, and down arrow means the opposite. The percentages in the brackets are the rates of change relative to the SFT baseline. Best scores are in bold.

We collect the sentences containing *that* complement clauses from the Switchboard corpus (Godfrey et al., 1992), following the same `tgrep2` (Rohde, 2005) pattern-based extraction procedure used in Florian Jaeger (2010)'s analysis. 4817 sentences are extracted in total, from which we count the numbers of necessary vs. redundant *that* being present or absent, using probabilities computed from the SFT, SimPO, and UIDPO (Llama-3-8B-Instruct backbone) models, respectively. From Table 5 we can see that UIDPO greatly improves the number of correct *that*, and also reduces the number of mistaken *that* by a smaller margin – **UIDPO outperforms SFT and SimPO on both metrics**.

Using the sentence S_1 : *My boss confirmed that we were absolutely crazy* as an example, and let c denote the context *My boss confirmed*, we examine the probabilities of *that* and the alternative continuation *we*. When a UIDPO-finetuned model (Llama-3-8B-Instruct, $\alpha = 0.250$) is used to estimate probabilities, we have $p^{\text{UIDPO}}(\textit{that} \mid c) = 0.4996$ and $p^{\text{UIDPO}}(\textit{we} \mid c) = 0.0005$; when a SimPO model (Llama-3-8B-Instruct) is used, we have $p^{\text{SimPO}}(\textit{that} \mid c) = 0.4326$, $p^{\text{SimPO}}(\textit{we} \mid c) = 0.0008$. The log-odds $\frac{p(\textit{that}|c)}{p(I|c)}$ is greater for UIDPO, $6.8125 > 6.2500$. Since this is a case for necessary *that*, it shows UIDPO indeed has a higher chance to produce $p(\textit{that})$ than the alternative. However, situation is more complicated for another example where *that* is redundant, S_2 : *My boss thinks (that) I am absolutely crazy*. UIDPO indeed assigns higher probability to the alternative than SimPO, $p^{\text{UIDPO}}(I \mid c) = 0.6303 > p^{\text{SimPO}}(I \mid c) = 0.6152$, but the redundant *that* is also assigned higher probability, $p^{\text{UIDPO}}(\textit{that} \mid c) = 0.0908 > p^{\text{SimPO}}(\textit{that} \mid c) = 0.0782$. The log-odds $\frac{p(\textit{that}|c)}{p(I|c)}$ is not in favor of UIDPO, $-1.9375 > -2.0625$, because it indicates UIDPO still has relative higher odds to produce *that* than SimPO (though the difference is small). With these two cherry-picked examples, we would like to intuitively demonstrate the potential effect and limit of UIDPO to produce more "human-like" *that* complementizer.

Finally, we use Cochran's Q + McNemar test (Repeated Measures Chi-square Test) to conduct a group-wise comparison (SFT vs. SIMPO, SFT vs. UIDPO, and SIMPO vs. UIDPO) for Table 5, and the resulting p-values are all below 0.001, indicating a statitically significant effect. Therefore, we conclude that UIDPO-finetuned model has a higher chance to produce a necessary *that* and to avoid producing a redundant *that*, which aligns better with the human preference in producing uniform information density. Overall speaking, the effect is more salient in the former than in the latter, suggesting that it is easier to instruct a model to emphasize *what should already there* not to de-emphasize *what could be absent*. Detailed procedures and some additional analyses are provided in Appendix F.

5 Related Work

UID Theory The Uniform Information Density hypothesis originates from the intersection of information theory and psycholinguistics, building on Shannon's formalization of information as surprisal (Shannon, 1948; Hale, 2001; Levy, 2008). Jaeger & Levy (2006) formulates the UID hypothesis, which assumes that speakers prefer utterances that distribute information uniformly across a signal . Florian Jaeger (2010); Frank & Jaeger (2008) conducts tests to systematically discusses the mechanism of redundancy and reduction, and how they are managed by a rational speaker to uniform the information density. Piantadosi et al. (2011) examines that more predictable words are optimized to have shorter length and more uniform information.

Recent work has revisited and refined UID from both empirical and theoretical perspectives. Meister et al. (2021) systematically show that linguistic acceptability and processing difficulty correlate with deviations from uniform surprisal. This paper opens up evaluation and application of UID in the context of large language models. Follow-up research shows that UID-inspired metrics can characterize the quality of reasoning processes (Gwak et al., 2026). It reports a counter-intuitive pattern of LLM reasoning trajectories: high-quality reasoning exhibits local uniformity and global non-uniformity. This observation directly inspires the redefinition of uniformity in UIDPO.

Preference Optimization Reinforcement Learning from Human Feedback (RLHF) is the standard paradigm for aligning large language models with human preferences. (Christiano et al., 2023; Ouyang et al., 2022). Despite its effectiveness, RLHF is computationally expensive and vulnerable to instability and reward mis-specification. Recent work proposes direct preference optimization methods that remove explicit reinforcement learning. Direct Preference Optimization (DPO) (Rafailov et al., 2024) reformulates the RLHF objective into a supervised objective by expressing the reward as a function of the log-probability difference between the policy and a reference model. This avoids training a separate reward model and stabilizes optimization . Subsequent work aims to further simplify and improve this framework. Simple Preference Optimization (SimPO) (Meng et al., 2024) removes the reference model entirely and defines the implicit reward from the model likelihood, making the objective more consistent with the generation. Building on this idea, later methods explore different variants of preference optimization objectives (Wu et al., 2025; Yin et al., 2025).

6 Conclusion

In this work, we draw inspirations from a well-established psycholinguistic theory, Uniform Information Density, to develop a novel preference optimization algorithm, UIDPO. We compare UIDPO with SimPO, one of the State-of-the-Art preference optimization methods, and achieve comparable and sometimes better performance on major instruction following benchmarks. The key of our algorithm is to formulate the original hypothesis of UID theory as the compound of global variance and a second-order difference-based local metrics of information uniformity, which better fit the objective of preference optimization. Extensive experiments prove the necessity and superior performance of our UIDPO formulations over naïve choices. Additional linguistic analysis on complement clauses further reveals the advantages of UIDPO in obtaining language models that can distribute information in more human-aligned manners. Our work is a successful attempt to incorporate the underlying mechanisms of information processing into the post-training of LLMs that potentially better align with human language use.

Acknowledgments

We sincerely thank Linyi Yang for providing insightful suggestions on model training. We thank all reviewers for giving constructive feedbacks that make this paper much improved. This study is funded by Shenzhen Science and Technology Program (No. JCYJ20240813094612017) and Guangdong Province ZJRC Program (No. 2024QN11X145).

References

AI@Meta. Llama 3 model card. 2024. URL https://github.com/meta-llama/llama3/blob/main/MODEL_CARD.md.

Amanda Askell, Yuntao Bai, Anna Chen, Dawn Drain, Deep Ganguli, Tom Henighan, Andy Jones, Nicholas Joseph, Ben Mann, Nova DasSarma, Nelson Elhage, Zac Hatfield-Dodds, Danny Hernandez, Jackson Kernion, Kamal Ndousse, Catherine Olsson, Dario Amodei, Tom Brown, Jack Clark, Sam McCandlish, Chris Olah, and Jared Kaplan. A general language assistant as a laboratory for alignment, 2021. URL https://arxiv.org/abs/2112.00861.

Mohammad Gheshlaghi Azar, Mark Rowland, Bilal Piot, Daniel Guo, Daniele Calandriello, Michal Valko, and Rémi Munos. A general theoretical paradigm to understand learning from human preferences, 2023. URL https://arxiv.org/abs/2310.12036.

Stephen Casper, Xander Davies, Claudia Shi, Thomas Krendl Gilbert, Jérémy Scheurer, Javier Rando, Rachel Freedman, Tomasz Korbak, David Lindner, Pedro Freire, Tony Wang, Samuel Marks, Charbel-Raphaël Segerie, Micah Carroll, Andi Peng, Phillip Christoffersen, Mehul Damani, Stewart Slocum, Usman Anwar, Anand Siththaranjan, Max Nadeau, Eric J. Michaud, Jacob Pfau, Dmitrii Krasheninnikov, Xin Chen, Lauro Langosco, Peter Hase, Erdem Bıyık, Anca Dragan, David Krueger, Dorsa Sadigh, and Dylan Hadfield-Menell. Open problems and fundamental limitations of reinforcement learning from human feedback, 2023. URL https://arxiv.org/abs/2307.15217.

Wei-Lin Chiang, Lianmin Zheng, Ying Sheng, Anastasios Nikolas Angelopoulos, Tianle Li, Dacheng Li, Hao Zhang, Banghua Zhu, Michael Jordan, Joseph E. Gonzalez, and Ion Stoica. Chatbot arena: An open platform for evaluating llms by human preference, 2024. URL https://arxiv.org/abs/2403.04132.

Paul Christiano, Jan Leike, Tom B. Brown, Miljan Martic, Shane Legg, and Dario Amodei. Deep reinforcement learning from human preferences, 2023. URL https://arxiv.org/abs/1706.03741.

Michael Collins. Information density and dependency length as complementary cognitive models. *Journal of psycholinguistic research*, 43, 09 2013. doi: 10.1007/s10936-013-9273-3.

Ganqu Cui, Lifan Yuan, Ning Ding, Guanming Yao, Wei Zhu, Yuan Ni, Guotong Xie, Zhiyuan Liu, and Maosong Sun. UltraFeedback: Boosting language models with high-quality feedback. *arXiv preprint arXiv:2310.01377*, 2023.

Yann Dubois, Balázs Galambosi, Percy Liang, and Tatsunori B Hashimoto. Length-controlled alpacaeval: A simple way to debias automatic evaluators. *arXiv preprint arXiv:2404.04475*, 2024.

Kawin Ethayarajh, Winnie Xu, Niklas Muennighoff, Dan Jurafsky, and Douwe Kiela. Kto: Model alignment as prospect theoretic optimization, 2024. URL https://arxiv.org/abs/2402.01306.

T. Florian Jaeger. Redundancy and reduction: Speakers manage syntactic information density. *Cognitive Psychology*, 61(1):23–62, 2010. ISSN 0010-0285. doi: https://doi.org/10.1016/j.cogpsych.2010.02.002. URL https://www.sciencedirect.com/science/article/pii/S0010028510000083.

A. Frank and T. Florian Jaeger. Speaking rationally: Uniform information density as an optimal strategy for language production. *Proceedings of the 30th Annual Meeting of the Cognitive Science Society*, pp. 939–944, 01 2008.

John J Godfrey, Edward C Holliman, and Jane McDaniel. Switchboard: Telephone speech corpus for research and development. In *[Proceedings] ICASSP-92: 1992 IEEE International Conference on Acoustics, Speech, and Signal Processing*, volume 1, pp. 517–520. IEEE, 1992.

Minju Gwak, Guijin Son, and Jaehyung Kim. Revisiting the uniform information density hypothesis in llm reasoning traces, 2026. URL https://arxiv.org/abs/2510.06953.

John Hale. A probabilistic Earley parser as a psycholinguistic model. In *Second Meeting of the North American Chapter of the Association for Computational Linguistics*, 2001. URL https://aclanthology.org/N01-1021/.

Jiwoo Hong, Noah Lee, and James Thorne. ORPO: Monolithic preference optimization without reference model. In Yaser Al-Onaizan, Mohit Bansal, and Yun-Nung Chen (eds.), *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pp. 11170–11189, Miami, Florida, USA, November 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.emnlp-main.626. URL https://aclanthology.org/2024.emnlp-main.626/.

T. Jaeger and Roger Levy. Speakers optimize information density through syntactic reduction. In B. Schölkopf, J. Platt, and T. Hoffman (eds.), *Advances in Neural Information Processing Systems*, volume 19. MIT Press, 2006. URL https://proceedings.neurips.cc/paper_files/paper/2006/file/c6a01432c8138d46ba39957a8250e027-Paper.pdf.

Albert Q. Jiang, Alexandre Sablayrolles, Arthur Mensch, Chris Bamford, Devendra Singh Chaplot, Diego de las Casas, Florian Bressand, Gianna Lengyel, Guillaume Lample, Lucile Saulnier, Lélio Renard Lavaud, Marie-Anne Lachaux, Pierre Stock, Teven Le Scao, Thibaut Lavril, Thomas Wang, Timothée Lacroix, and William El Sayed. Mistral 7b, 2023. URL https://arxiv.org/abs/2310.06825.

Seung-Jean Kim, Kwangmoo Koh, Stephen Boyd, and Dimitry Gorinevsky. ℓ_1 trend filtering. *SIAM Review*, 51(2):339–360, 2009. doi: 10.1137/070690274. URL https://doi.org/10.1137/070690274.

Jan Leike, David Krueger, Tom Everitt, Miljan Martic, Vishal Maini, and Shane Legg. Scalable agent alignment via reward modeling: a research direction, 2018. URL https://arxiv.org/abs/1811.07871.

Roger Levy. Expectation-based syntactic comprehension. *Cognition*, 106(3):1126–1177, 2008. ISSN 0010-0277. doi: https://doi.org/10.1016/j.cognition.2007.05.006. URL https://www.sciencedirect.com/science/article/pii/S0010027707001436.

Tianle Li, Wei-Lin Chiang, Evan Frick, Lisa Dunlap, Tianhao Wu, Banghua Zhu, Joseph E Gonzalez, and Ion Stoica. From crowdsourced data to high-quality benchmarks: Arena-hard and benchbuilder pipeline. *arXiv preprint arXiv:2406.11939*, 2024.

Xuechen Li, Tianyi Zhang, Yann Dubois, Rohan Taori, Ishaan Gulrajani, Carlos Guestrin, Percy Liang, and Tatsunori B. Hashimoto. Alpacaeval: An automatic evaluator of instruction-following models. https://github.com/tatsu-lab/alpaca_eval, 5 2023.

Clara Meister, Tiago Pimentel, Patrick Haller, Lena Jäger, Ryan Cotterell, and Roger Levy. Revisiting the uniform information density hypothesis, 2021. URL https://arxiv.org/abs/2109.11635.

Yu Meng, Mengzhou Xia, and Danqi Chen. Simpo: Simple preference optimization with a reference-free reward, 2024. URL https://arxiv.org/abs/2405.14734.

Yulin Ou, Yu Wang, Yang Xu, and Hendrik Buschmeier. Identifying the periodicity of information in natural language. *arXiv preprint arXiv:2510.27241*, 2025.

Long Ouyang, Jeff Wu, Xu Jiang, Diogo Almeida, Carroll L. Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, John Schulman, Jacob Hilton, Fraser Kelton, Luke Miller, Maddie Simens, Amanda Askell, Peter Welinder, Paul Christiano, Jan Leike, and Ryan Lowe. Training language models to follow instructions with human feedback, 2022. URL https://arxiv.org/abs/2203.02155.

Ryan Park, Rafael Rafailov, Stefano Ermon, and Chelsea Finn. Disentangling length from quality in direct preference optimization. In Lun-Wei Ku, Andre Martins, and Vivek Srikumar (eds.), *Findings of the Association for Computational Linguistics: ACL 2024*, pp. 4998–5017, Bangkok, Thailand, August 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.findings-acl.297. URL https://aclanthology.org/2024.findings-acl.297/.

Steven T. Piantadosi, Harry Tily, and Edward Gibson. Word lengths are optimized for efficient communication. *Proceedings of the National Academy of Sciences*, 108(9):3526–3529, 2011. doi: 10.1073/pnas.1012551108. URL https://www.pnas.org/doi/abs/10.1073/pnas.1012551108.

Rafael Rafailov, Archit Sharma, Eric Mitchell, Stefano Ermon, Christopher D. Manning, and Chelsea Finn. Direct preference optimization: Your language model is secretly a reward model, 2024. URL https://arxiv.org/abs/2305.18290.

Douglas LT Rohde. Tgrep2 user manual. *Unpublished manuscript*, 2005.

John Schulman, Filip Wolski, Prafulla Dhariwal, Alec Radford, and Oleg Klimov. Proximal policy optimization algorithms, 2017. URL https://arxiv.org/abs/1707.06347.

Cory Shain, Clara Meister, Tiago Pimentel, Ryan Cotterell, and Roger Levy. Large-scale evidence for logarithmic effects of word predictability on reading time. *Proceedings of the National Academy of Sciences*, 121(10):e2307876121, 2024. doi: 10.1073/pnas.2307876121. URL https://www.pnas.org/doi/abs/10.1073/pnas.2307876121.

C. E. Shannon. A mathematical theory of communication. *The Bell System Technical Journal*, 27(3):379–423, 1948. doi: 10.1002/j.1538-7305.1948.tb01338.x.

Nisan Stiennon, Long Ouyang, Jeffrey Wu, Daniel Ziegler, Ryan Lowe, Chelsea Voss, Alec Radford, Dario Amodei, and Paul F Christiano. Learning to summarize with human feedback. In H. Larochelle, M. Ranzato, R. Hadsell, M.F. Balcan, and H. Lin (eds.), *Advances in Neural Information Processing Systems*, volume 33, pp. 3008–3021. Curran Associates, Inc., 2020. URL https://proceedings.neurips.cc/paper_files/paper/2020/file/1f89885d556929e98d3ef9b86448f951-Paper.pdf.

Evan Frick Lisa Dunlap Banghua Zhu Joseph E. Gonzalez Ion Stoica Tianle Li*, Wei-Lin Chiang*. From live data to high-quality benchmarks: The arena-hard pipeline, April 2024. URL https://lmsys.org/blog/2024-04-19-arena-hard/.

Eleftheria Tsipidi, Franz Nowak, Ryan Cotterell, Ethan Wilcox, Mario Giulianelli, and Alex Warstadt. Surprise! Uniform Information Density isn't the whole story: Predicting surprisal contours in long-form discourse. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pp. 18820–18836, Miami, Florida, USA, 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.emnlp-main.1047.

Eleftheria Tsipidi, Samuel Kiegeland, Franz Nowak, Tianyang Xu, Ethan Wilcox, Alex Warstadt, Ryan Cotterell, and Mario Giulianelli. The harmonic structure of information contours. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 31636–31659, Vienna, Austria, 2025. Association for Computational Linguistics. URL https://aclanthology.org/2025.acl-long.1527/.

Haoxiang Wang, Wei Xiong, Tengyang Xie, Han Zhao, and Tong Zhang. Interpretable preferences via multi-objective reward modeling and mixture-of-experts. *arXiv preprint arXiv:2406.12845*, 2024a.

Yubo Wang, Xueguang Ma, Ge Zhang, Yuansheng Ni, Abhranil Chandra, Shiguang Guo, Weiming Ren, Aaran Arulraj, Xuan He, Ziyan Jiang, et al. Mmlu-pro: A more robust and challenging multi-task language understanding benchmark. *arXiv preprint arXiv:2406.01574*, 2024b.

Junkang Wu, Xue Wang, Zhengyi Yang, Jiancan Wu, Jinyang Gao, Bolin Ding, Xiang Wang, and Xiangnan He. Alphadpo: Adaptive reward margin for direct preference optimization, 2025. URL https://arxiv.org/abs/2410.10148.

Haoran Xu, Amr Sharaf, Yunmo Chen, Weiting Tan, Lingfeng Shen, Benjamin Van Durme, Kenton Murray, and Young Jin Kim. Contrastive preference optimization: Pushing the boundaries of llm performance in machine translation, 2024. URL https://arxiv.org/abs/2401.08417.

Yang Xu and David Reitter. Information density converges in dialogue: Towards an information-theoretic model. *Cognition*, 170:147–163, 2018. doi: h10.1016/j.cognition.2017.09.018.

Qingyu Yin, Chak Tou Leong, Minjun Zhu, Hanqi Yan, Qiang Zhang, Yulan He, Wenjie Li, Jun Wang, Yue Zhang, and Linyi Yang. Constrain alignment with sparse autoencoders, 2025. URL https://arxiv.org/abs/2411.07618.

Hongyi Yuan, Zheng Yuan, Chuanqi Tan, Wei Wang, Songfang Huang, and Fei Huang. Rrhf: Rank responses to align language models with human feedback. In A. Oh, T. Naumann, A. Globerson, K. Saenko, M. Hardt, and S. Levine (eds.), *Advances in Neural Information Processing Systems*, volume 36, pp. 10935–10950. Curran Associates, Inc., 2023. URL https://proceedings.neurips.cc/paper_files/paper/2023/file/23e6f78bdec844a9f7b6c957de2aae91-Paper-Conference.pdf.

Yao Zhao, Rishabh Joshi, Tianqi Liu, Misha Khalman, Mohammad Saleh, and Peter J. Liu. Slic-hf: Sequence likelihood calibration with human feedback, 2023. URL https://arxiv.org/abs/2305.10425.

Lianmin Zheng, Wei-Lin Chiang, Ying Sheng, Siyuan Zhuang, Zhanghao Wu, Yonghao Zhuang, Zi Lin, Zhuohan Li, Dacheng Li, Eric. P Xing, Hao Zhang, Joseph E. Gonzalez, and Ion Stoica. Judging llm-as-a-judge with mt-bench and chatbot arena, 2023a.

Rui Zheng, Shihan Dou, Songyang Gao, Yuan Hua, Wei Shen, Binghai Wang, Yan Liu, Senjie Jin, Qin Liu, Yuhao Zhou, Limao Xiong, Lu Chen, Zhiheng Xi, Nuo Xu, Wenbin Lai, Minghao Zhu, Cheng Chang, Zhangyue Yin, Rongxiang Weng, Wensen Cheng, Haoran Huang, Tianxiang Sun, Hang Yan, Tao Gui, Qi Zhang, Xipeng Qiu, and Xuanjing Huang. Secrets of rlhf in large language models part i: Ppo, 2023b. URL https://arxiv.org/abs/2307.04964.

A More AlpacaEval 2 Results

Method	LC(%)	WR(%)	Length
RRHF (Yuan et al., 2023)	37.9	31.6	1700
SLiC-HF (Zhao et al., 2023)	33.9	32.5	1938
DPO (Rafailov et al., 2024)	48.2	47.5	2000
IPO (Azar et al., 2023)	46.8	42.4	1830
CPO (Xu et al., 2024)	34.1	36.4	2086
KTO (Ethayarajh et al., 2024)	34.1	32.1	1878
ORPO (Hong et al., 2024)	38.1	33.8	1803
R-DPO (Park et al., 2024)	48.0	45.8	1933
UIDPO	49.0	44.8	1825
SimPO reproduced	51.9	47.4	1861

Table 6: AlpacaEval 2 results of different post-training methods on Llama-3-8B-Instruct, utilizing SimPO v0.2 preference labels. All the scores are reported in SimPO paper. LC is the length-controlled win rate, and WR is the raw win rate. Length is the average token length of the responses.

Table 6 lists more AlpacaEval 2 results retrieved from SimPO paper. The reproduced SimPO and the UIDPO are under the same training and testing setup with all other methods. That is, using princeton-nlp/llama3-ultrafeedback-armorm as the preference data and GPT-4 1106 Preview as the judge for AlpacaEval 2.

B MMLU results

We test UIDPO and SimPO on MMLU-Pro (Wang et al., 2024b), and compare their performances with that of Llama-3-8B-Instruct without any post-training on the Leader-board.

answer extraction	Llama-3-8B-Instruct	SimPO run 1	SimPO run 2	UIDPO
Level 1	0.4098	0.3935	0.3910	**0.4122**
Level 2		0.3983	0.3965	**0.4122**

Table 7: MMLU results. Level 1 & 2 are different answer extraction mechanisms. The score of Llama-3-8B-Instruct is retrieved from the Leader-board, where only the best score is reported.

Since SimPO has not report its performance on MMLU-Pro, we can only compare with our reproduced SimPO. Thus we decide to include this result in the appendix instead of the main result table, to avoid confusion. According to Table 7, UIDPO slightly outperforms Llama-3-8B-Instruct and SimPO under both answer extraction mechanisms. Since our reproduced SimPO under-performs Llama-3-8B-Instruct, we run the benchmark for the second time, to confirm the result scores. This result shows that UIDPO can provide slight improvement on language understanding.

C Benchmark Statistics

Method	AlpacaEval 2		AlpacaEval 2*		Arena-Hard v0.1
	SE	avg. length	SE	avg. length	CI(%)
SimPO reprod.	1.54	1861	1.75	1861	-2.0/+1.8
UIDPO	**1.50**	**1825**	1.75	**1799**	-1.9/+1.7

Table 8: Key statistics reported by main experiment benchmarks. SE is the standard error (normalized by $N-1$) of the win rate. Avg. length is the average token length of responses. CI is the non-overlapping confidence interval with the baseline.

UIDPO has slightly better statistics than SimPO. Table 8 lists some key statistics reported by each benchmark. A smaller standard error implies more stable performance. Shorter average length means that UIDPO is less likely to exploit length bias. It is also a preferable feature in application, in terms of both user experience and cost. The Confidence Interval (CI) reported by Arena-Hard measures how similar a model's responses are to the baseline (GPT-4 0314 in this case). It is not directly related to performance.

D Average Log Probability of Chosen Responses in γ Ablations

Figure 3 displays three examples of the average log probability curves during training. We select three curves of evenly spaced γs for clearer exhibition, and other curves also conform the same pattern. As γ becomes larger, the average log probability drops more sharply. This confirms the discovery in SimPO paper, that a larger γ flattens the probability distribution.

E Reward Margins and UID Margins

To exhibit the approximate numerical range and the trend of the UIDPO reward and UID terms, we plot the following two figures. Figure 4 displays the widening gap between rewards of chosen responses and rejected responses. Figure 5 displays the dropping UID, where the global term drops but slower. This is probably due to the choice of α. More weight on the local term (negative contribution) forces the global term to drop, while the positive contribution of the global term resists. this causes the widening gap between the global UID and the local UID.

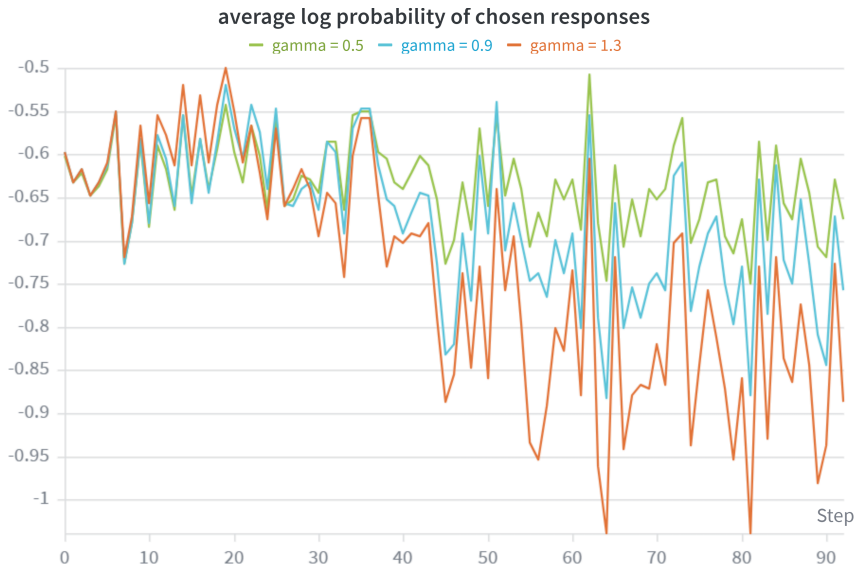

Figure 3: Average log probability of chosen responses during training, when $\gamma = 0.5, 0.9, 1.3$.

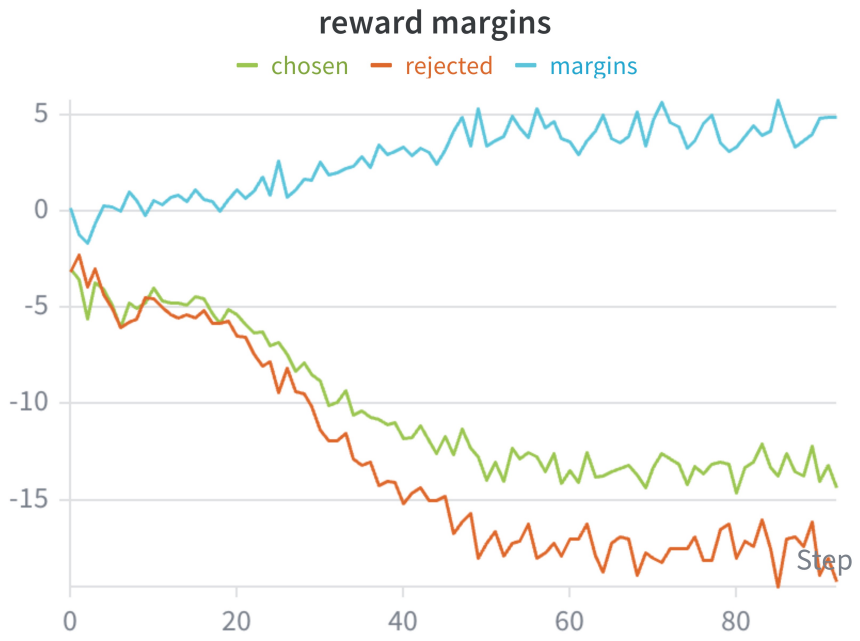

Figure 4: The reward margins between chosen responses and rejected responses during training, under the best performing setup.

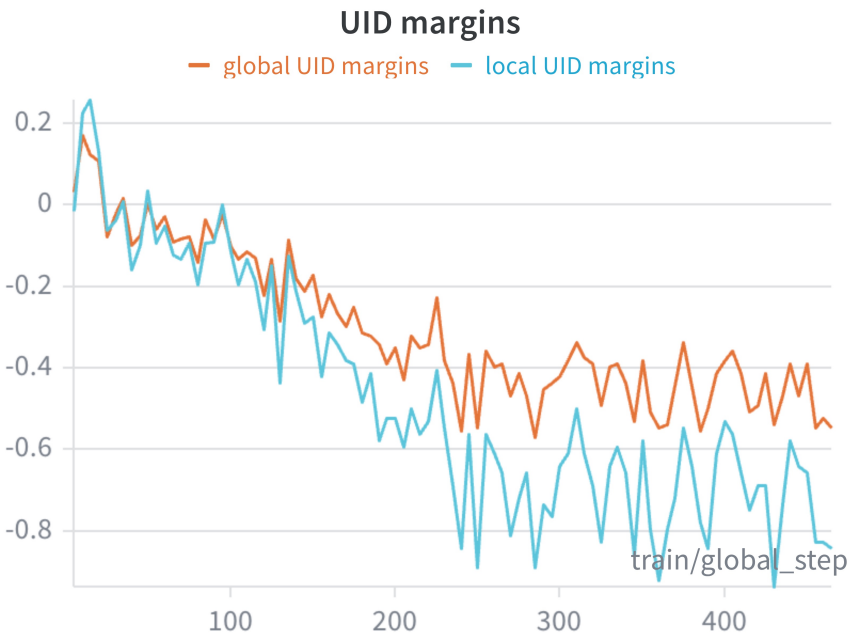

Figure 5: The global and local UID margins between chosen responses and rejected responses during training, under the best performing setup.

F Linguistic Analysis Details

This section provides additional details about the complement-clause analysis in Section 4. All three models in this analysis use the Llama-3-8B-Instruct backbone.

Corpus and clause extraction. We use the fluent portion of the parsed Switchboard corpus (Godfrey et al., 1992). Candidate clauses are extracted using an NLTK-based approximation of the `tgrep2` query used by Florian Jaeger (2010). The procedure selects trees containing an `SBAR` associated with a verbal projection and requires the clause to begin with either an overt complementizer *that* or a Treebank null-complementizer annotation. It also removes clauses with a subject gap, extraposed clauses, clauses containing a wh-expression, and common parenthetical uses such as *I mean* and *you know*.

The structural constraints are adapted from the following `tgrep2` query:

```
/^sbar/ $,, /^vb/
[<(in<"that") |< "-none-"]
<(/^s/ < (/-sbj/ !< "-none-"))
!<(/^n/ < ! /^'0,1[a-zA-Z]+.*/)
!</^wh/
```

Our NLTK implementation approximates these constraints with the following procedure:

```
for each parsed tree:
    require at least one SBAR and one VP
    require overt "that", null "0", or a -NONE- node
    reject trees containing an empty subject
    reject extraposed clauses identified by an indexed NP
    reject WH constituents and clauses headed by "whether"
    reject short parenthetical "I mean" or "you know" uses
    retain the remaining tree
```

The filtering procedure retains 4,867 of the 177,804 parsed trees. Within these trees, we identify 4,817 complement-clause instances: 962 with an overt *that* and 3,855 with a null complementizer. These instances correspond to 1,768 distinct prefixes. The descriptive statistics below treat clause occurrences, rather than unique prefixes, as separate observations.

Construction of model inputs. For an overt complementizer, we retain the observed *that*. For a null complementizer, we insert *that* at the annotated clause boundary. We then retain the prefix preceding this position and the first token w of the embedded clause. This construction allows *that* and w to be evaluated as alternative continuations of the same prefix c.

Probability-based measures. For each model, we compute the negative log likelihoods $\mathrm{NLL}_\theta(\textit{that} \mid c) = -\log p_\theta(\textit{that} \mid c)$ and $\mathrm{NLL}_\theta(w \mid c) = -\log p_\theta(w \mid c)$. The latter is the probability of beginning the embedded clause without an overt complementizer; it is **not** the probability of w conditioned on *that*. We also count the number of vocabulary items assigned a higher probability than *that* under the same prefix. A value of zero therefore indicates that *that* is the model's most probable continuation. We report the mean of this rank-based measure below.

Descriptive results. UIDPO assigns a higher NLL to *that* than SFT on 85.3% of the analyzed instances. Across all instances, the mean ratio between the UIDPO and SFT NLLs is 1.070. Conversely, UIDPO assigns a lower NLL than SimPO on 96.5% of the instances, with a mean UIDPO-to-SimPO NLL ratio of 0.945.

The mean difference

$$\mathrm{NLL}_\theta(\textit{that} \mid c) - \mathrm{NLL}_\theta(w \mid c) \tag{8}$$

is -2.39 for SFT, -2.27 for UIDPO, and -2.13 for SimPO. Thus, all three models assign a higher probability, on average, to *that* than to beginning the embedded clause directly.

For the rank-based measure defined above, the average numbers of vocabulary items assigned a higher probability than *that* are 46.18 for SFT, 59.91 for UIDPO, and 63.28 for SimPO. As a qualitative complement, we also count the tokens that most frequently rank above *that* across the analyzed contexts. Although their exact order varies slightly across models, the same ten items occur most frequently for SFT, SimPO, and UIDPO: *I*, *it*, *you*, *the*, comma, period, *to*, *we*, *in*, and *there*.

Taken together, these aggregate statistics place UIDPO between SFT and SimPO in its overall propensity to produce *that*. They are descriptive and, on their own, do not establish whether a model distinguishes contexts that favor realization from those that favor omission.